



## Introduction: ways of machine seeing

Mitra Azar<sup>1</sup> · Geoff Cox<sup>2</sup> · Leonardo Impett<sup>3</sup>

Received: 23 November 2020 / Accepted: 26 November 2020 / Published online: 20 February 2021  
© The Author(s), under exclusive licence to Springer-Verlag London Ltd. part of Springer Nature 2021

How do machines, and, in particular, computational technologies, change the way we see the world? This special issue brings together researchers from a wide range of disciplines to explore the entanglement of machines and their ways of seeing from new critical perspectives. As the title makes clear, we take our point of departure in John Berger’s 1972 BBC documentary series *Ways of Seeing*, a four-part television series of 30-min films created by Berger and producer Mike Dibb, which had an enormous impact on both popular and academic perspectives on visual culture. Berger’s scripts were adapted into a book of the same name, published by Penguin also in 1972. The book consists of seven numbered essays: four using words and images; and three essays using only images. Seeing is evidently a political act, exemplified in the third episode-chapter, where images of women in early modern European painting (Pol de Limbourg, Cranach the Elder, Jan Gossaert, Tintoretto) and commercial magazines are juxtaposed to demonstrate the ways in which women are rendered as objects of the male gaze. More broadly, Berger emphasised that “the relation between what we see and what we know is never settled”. In this special issue, we explore how these ideas can be understood in the light of technical developments in machine vision and algorithmic learning, and how the relations between what we see and know are further unsettled.

What we see above (Fig. 1) is clearly not a book as such but a technically reproduced image of a book (Cox 2016). This testifies to the ways in which what, and how, we see and know is further unsettled through complex assemblages of elements, the details of which are largely kept from view. Access to the means of production here emphasises the Marxist approach of Berger (and, in turn, the Marxism of Benjamin, who Berger credits with furnishing the key concepts in *Ways of Seeing*), and how exposing what stays invisible allows a political understanding of social relations and thereby the possibility of their transformation. The TV programmes employed Brechtian techniques: revealing the technical apparatus of the studio, to encourage viewers not to simply watch (or read) in a straightforward way but rather be forced into an analysis of their alienation. Although rather less sophisticated in form, we hope for something similar with this journal that readers will reflect on what they see and read in ways that lead to a “return from alienation” or recognition of distancing-effects—breaking the “fourth wall” of machine vision, so to speak—to expose the unevenness of social relations in new ways.

Alienated forms of social interaction have become the “new normal” as we write. The current pandemic seems to heighten uncertainties about what is rendered visible and invisible to human perception. When the visual field is increasingly nonhuman, how is the world made knowable to us when much of its operations lay outside our visual register and consequently outside the scope of human action? Adrian Mackenzie and Anna Munster refer to an “operationalization” of visibility, in which images operate within a field of “distributed invisibility” in which relations between images count more than indexicality or iconicity (we might add, “aura” (Benjamin 2008)) of a single image (Mackenzie and Munster 2019: 16). Seeing, or what they call “platform seeing”, becomes distributed through data practices and machinic assemblages that “emphasize the importance of the formatting of image ensembles as datasets across contemporary data practices; the incorporation of platforms into hardware in devices; forms of parallel computation; and the computational architectures of contemporary artificial

---

✉ Geoff Cox  
geoffcox@lsbu.ac.uk

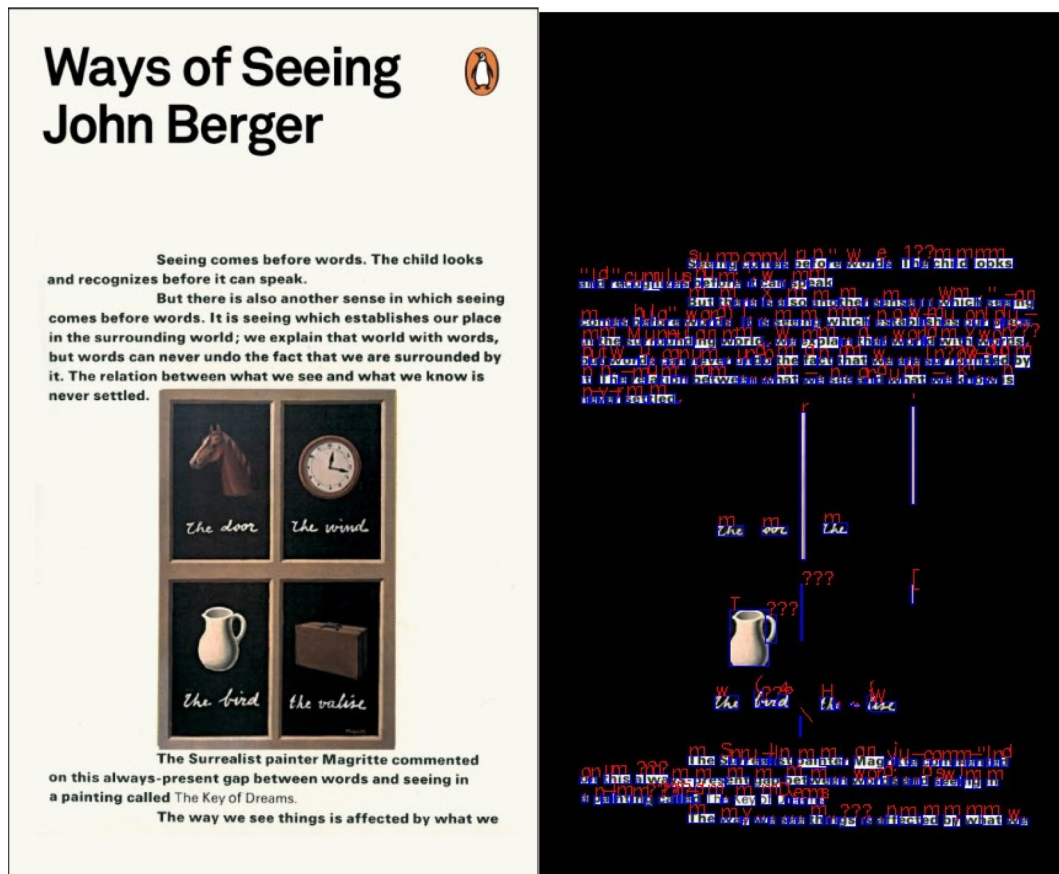
Mitra Azar  
azar@cc.au.dk

Leonardo Impett  
leonardo.l.impett@durham.ac.uk

<sup>1</sup> Department of Digital Design and Information Studies, Aarhus University, Aarhus, Denmark

<sup>2</sup> Centre for the Study of the Networked Image, London South Bank University, London, UK

<sup>3</sup> Department of Computer Science, Durham University, Durham, UK



**Fig. 1** The Cover of *Ways of Seeing* by John Berger (1972), and as seen through an optical character recognition programme. Images from Penguin Books and Scandinavian Institute for Computational Vandalism

intelligence. These assemblages constitute the (nonhuman) activities of perception as mode of cutting into/selecting out of the entire flux of image-ensemble world” (2019: 3). The importance of this understanding is that a new mode of perception is operationalized, what they call “invisual perception” (2019: 4): a new way of (machine) seeing which is an assemblage of its various parts; including imaging devices (such as cameras), the data they produce (which might take the form of an image), and the wider practices and infrastructures through which they are operationalized (in terms of its application).

So how to begin to understand the relation between seeing and knowing within this new operational context? Artist Trevor Paglan bluntly explains: “We’ve long known that images can kill. What’s new is that nowadays, they have their fingers on the trigger” (2014). This intensification of visuality takes “necropolitical” form (defined by Achille Mbembé (2003) as the question of who lives and who dies). Central to this is the “right to look” as Nicholas Mirzoeff has previously put it, the need to reclaim autonomy from authority, and generate new forms of “countervisuality” that turn the unreality created by visuality’s fake authority into real alternatives

(2011: 476, 485). For this journal, the special relationship between the formation of “racial surveillance capitalism” and the artificial vision of colonial power is further elaborated by Mirzoeff, for instance, as a form of fake authority over asylum seekers and refugees (see “Artificial Vision, White Space and Racial Surveillance Capitalism”, herein). The “realism” of this authority is rendered open to question, and some of the contradictions inherent to machine seeing are exposed. Indeed, “Reality is not only everything which is, but everything which is becoming. It’s a process. It proceeds in contradictions. If it is not perceived in its contradictory nature it is not perceived at all” (Brecht, cited in Mirzoeff 2011: 477).

We might well ask how machine seeing further exposes contradictions, or perhaps there is another kind of negation at work? To paraphrase, Luciana Parisi in “Negative Optics in Vision Machines” (herein), can machine vision step beyond the “oculocentric metaphysics of the Western gaze and the reproduction of racial capital”? She refers to this “inhuman” mode of machine vision as “negative optics” (like countervisuality) in order to offer an internal critique of ocular metaphysics, but also to defy the equation of value

between 0 and 1 s that sustains the universal law of (computational) capital. By starting from the negativity of the image, the racialized and gendered conditions of “artificial intelligence capitalism” demonstrate that “the equation of value maintains the condition of the zero of blackness”. What is challengingly proposed is that randomness in computation is part of the expansion of “heretic epistemologies” that start from dark optics and address what Denise Ferreira Da Silva calls “blackness”, namely “matter without form”, or “matter beyond the equation of value” (2017). Parisi suggests that, instead of being just invisible, blackness, as matter without form, brings forward the nullification of the ocularcentric field of vision.

What constitutes knowledge or value can be seen to be arranged in ways that further recall Berger’s reflections on the medium of television through which his ideas were made public: “But remember that I am controlling and using for my own purposes the means of reproduction needed for these programmes [...] with this programme as with all programmes, you receive images and meanings which are arranged. I hope you will consider what I arrange but please remain skeptical of it.” (1972). We reiterate this statement here to further stress reflection on the means of production, including journals like this, that purport to impart useful knowledge. What is learnt should not be separated from the means by which it is transmitted or circulated. More to the point, the production of meaning lies at the core of our discussion, as do concerns about what is being learnt, and to what extent this has been compromised by the mode of production, or inflected by reductive ideas of how the world operates. Under these conditions, the relations between human and machine learning become quite blurry. The overall idea of “learning” implies new forms of control over what and how something becomes known and how decisions are made, as for instance, in the ways in which images are classified and categorized by humans and machines (e.g. concluding that this image of a person most likely represents a specific gender, race, likely terrorist, and so on). Knowledge is often set at the lowest common denominator in such cases, backed up by the enormous infrastructural power of the companies that profit from generalisations, as is the case for platform-based media empires—such as Amazon and Google—who are concerned that users simply supply (visual) data.

That machines can be said to “see” or “learn” is shorthand for calculative practices that only approximate (or generalise) likely outcomes by using various probabilistic algorithms and models that all have built upon inherent human–machine prejudices such as those related to gender and race (see Myers West et al. 2019). What are the political consequences of automated systems that use labels to classify humans, including by race, gender, emotions, ability, sexuality, and personality? Kate Crawford and Trevor

Paglan’s “Excavating AI” (2019, republished in a new version for this journal) examines how and what computers recognize in an image, and indeed what is misrecognized? Computer vision systems make judgements, and decisions, and as such exercise power to shape the world in their own images, which, in turn, is built upon flattening generalities and embedded social bias.

A cartography of the limits of AI is provided by Matteo Pasquinelli and Vladan Joler in “The Nooscope Manifested” (herein), applying the analogy of optical media to “diagram” machine learning apparatuses. Ultimately, they wish for more collective intelligence about machine intelligence, more public education instead of “learning machines” and their “colonial regime of knowledge extractivism”. The interplay between truth and fiction is again part of this, and “deepfakes” for example (a wordplay on “deep learning”) make a good case study for the ways in which synthetic instances can pass for real data (or human reason) as perpetuated by corporate regimes of knowledge extractivism and epistemic colonialism (to paraphrase the evocative description of Pasquinelli and Joler). Furthermore, this interplay of truth/fake is what Abelardo Gil-Fournier and Jussi Parikka expand upon in “Ground Truth to Fake Geographies” (herein) to chart the development of what they call “ground truth”, with reference to the shift from physical, geographical ground, to the “ground of the image”. They discuss contemporary practices that mobilize geographical earth observation datasets for experimental purposes, including “fake geography” as well as artistic practices, to show how ground truth is “operationalised” (citing Harun Farocki 2004).

Seeing, then, is no longer centred, singular or indexical truth or reality, as it was mistakenly thought to be, indicative of a wider need to manifest authority and power through visibility, but takes on new distributed and contradictory forms. Fabian Offert and Peter Bell in “Perceptual Bias and Technical Metapictures” (herein), argue for a transdisciplinary approach—across computer science and visual studies—to understand the inherent biases of machine vision systems. What they call “perceptual bias” accounts for the differences between the assumed “ways of seeing” of a machine vision system, “our reasonable expectations regarding its way of representing the visual world, and its actual perceptual topology”. This shifts the discussion from critical attention to dataset bias and on fixing datasets by introducing more diverse sets of images. The point is upheld by Nicolas Malev in “On the Dataset’s Ruins” (herein) who acknowledges the dataset as significant cultural form, but also wants to shift attention from fixing the database (as in the case of the inherent biases of ImageNet) to highlight the invisible labour that labels and classifies the images, as well as to the operations of the apparatus itself. To understand the specific character of the “scale” of computer vision as he puts

it, comparison is made to André Malraux's "museum without walls" in which the photographic apparatus allows for unlimited access to the world's image resource. The social implications of this attention to scale—from the molecular to urban—are further elaborated in Benjamin Bratton's commentary on the The New Normal think-tank at Strelka Institute in Moscow ("AI Urbanism: A Design Framework for Governance, Program, and Platform Cognition", herein), thereby understanding AI as "a property that can be designed into objects of different scales" and thus more alert to questions of "what kind of *planetarity* can and must be composed".

From our open call for papers, contributors further address the ways in which existing references from visual culture—such as Berger's *Ways of Seeing*—useful as they remain, require additional work, not least in dialogue with other disciplines to further emphasise the importance of social and political aspects in technical fields such as AI (Agre 1997). (And we might reiterate the name of this journal to stress our point.) Reflecting what he calls "cultural analytics", Lev Manovich's "Computer Vision, Human Senses, and Language of Art" (herein) argues for the importance of using computer vision methods in humanities research, and how this can offer analytical insights into cultural objects where existing analytical tools fall short. In another article, "On Machine Vision and Photographic Imagination" (herein), Daniel Chávez Heras and Tobias Blanke discuss the experimental television programme *Made by Machine: When AI met the Archive* (2018), created with materials from the BBC television archive using different computational techniques. They establish a conceptual and material link between photographic practice and deep learning computer vision, and the implied "optical perspective of the computer vision system itself". Carloalberto Treccani's "The Brain, the Artificial Neural Network and the Snake" (herein) is an account of the evolutions of vision systems, from "intelligent" animals to the functioning of Artificial Neural Networks. Both human and machine learning can be seen to demonstrate, through trial and error (recurrent neural nets), how, for better or worse, artificial systems inform us about the workings of biological systems, and how and "why we see what we see". Claudio Celis Bueno and María Jesús Schultz Abarca's article "Memo Akten's Learning to See" (herein) employs the philosophy of Bernard Stiegler to insist that "human vision is always already technical", and the result of constant training processes. Akten's art installation *Learning to See* (2017) becomes an example of machinic imagination and "machinic unconscious" (referring to Walter Benjamin's notion of the "optical unconscious", and how technical reproduction provided access to new forms of vision). The optical logic is transformed, in ways that augment new layered realities. This layering is what Manuel van der Veen describes, in his "Crossroads of

Seeing" (herein), in which two ways of seeing are produced simultaneously, as our field of vision is superimposed with additional information and images. Augmented reality is compared to traditional procedures, such as *trompe-l'œil*, to suggest that new kinds of reflection become possible; not only to look at the intersection itself, but also to see where the ways "divide" (we might add contradict).

The inherent fallibility of AI systems is explored in Gabriel Pereira and Bruno Moreschi's "Artificial Intelligence and Institutional Critique 2.0" (herein), drawing upon their intervention within the art collection of the Van Abbemuseum in Eindhoven, Netherlands. Using widely available image-recognition software to "read" images, the inherent values of both are exposed. It is the "untrained eye" of computer vision that offers critique of the art system and possible new ways of approaching visual culture on the one hand, as well as understanding the commercial imperatives of computational ways of seeing on the other. The point is reiterated that new methods of analysis are required that work with computational practices and existing theoretical readings, in which the human and the machine are critical partners. In Iain Emsley's "Causality, Poetics, and Grammatology" (herein), a "critical assemblage" of philosophy and computational thinking allows for an analysis of the *Next Rembrandt* project (2016, a 3D printed painting made from the data of Rembrandt's total body of work using deep learning algorithms and facial recognition techniques). In Rebecca Uliasz's "Seeing like an Algorithm" (herein), the relationship between the image and the human subject is explored in detail. Techniques of machine vision are described as "techniques of algorithmic governance", and the way the human subject is made visible through computation. She refers to the relationship between the "operative image" (Farocki 2004) and the formation of "emergent subjects". Perle Møhl's "Seeing Threats, Sensing Flesh: Human-machine ensembles at work" (herein) further develops the discussion of "human-machine ensembles", how they work together mutually to "see" specific things and "unsee" others. Her key point is that seeing in both cases is "not automated but unskilled and mutually co-constituted". Examples are provided that demonstrate this coming together of "material, political, organisational, economic and fleshy entities in order to configure what can be seen and sensed, and what cannot".

We hope with this coming together of articles there is sufficient attention to critical-technical practices that illuminate the complexity of human-machine relations, and their transformations, and not least to serve to emphasise the uncertainties and contingencies that characterise these contemporary ways of seeing.

The relation between what we see and what we know has always been unsettled. It has been argued that seeing and ways of seeing have been intertwined since bipedism

bifurcated the evolution of the first anthropoids, which found themselves with hands for crafting (rather than only using them for walking) and a mouth for speaking (rather than only for eating) (Leroi-Gourhan 1993). The changing of orientation happens simultaneous to the development of an erect posture and the consequent changed function of the mouth, from an organ to simply gather and ingest food, to an organ increasingly capable of articulating sounds. Building on the philosophy of André Leroi-Gourhan, it is possible to imagine that this changing of perspective stimulated a vocal response from the liberated mouth—and the sequence of sounds were a first attempt to articulate the sense of awe (Plato 2004) or dread (Nietzsche 2006) associated with the new orientation of the body. At the same time, the liberated hands could start intervening with the surrounding landscape, scratching stones and, in a way, attempting inscription of those early articulations of sound emerging from the liberated mouth. Otherwise said, the technical ability acquired by the hands proceeds simultaneous to the possibility of articulating sounds to signify things, thanks to the new orientation of the erect body, and the freeing of the head and the mouth.

Although Berger seems to be strongly influenced by phenomenology—especially the phenomenology of perception of Merleau-Ponty (1964), and defines human seeing as primordial and in anticipation of language—the relation between seeing and ways of seeing is unsettled because seeing (as much as speaking) is always already technical, and, as such, written (Derrida 1976). Despite these phenomenological traces found in Berger’s approach, his work is also about the hidden intricacies between image and language, explored through a reverse engineering of the grammar of images (no doubt influenced by the influence of semiotics and structuralism at that time). Berger maintains the primordially of the image over the language, as is made clear in the first sentence of the book—“Seeing comes before words. The child looks and recognizes before it can speak”—or in the assertion that although we “explain that world with words [...], words can never undo the fact that we are surrounded by it” (Berger 1972: 7). Furthermore, “the reciprocal nature of vision [of simultaneously seeing and being seen] is more fundamental than that of spoken words” (Berger 1972: 8). Yet, to Berger, it is the very grammatical in-expressibility of the image that inevitably generates the power of signification through sounds and language, which in our earlier brief sketch of evolutionary palaeontology (inspired by Leroi-Gourhan) began with the changed orientation of the body, producing new ways of seeing together with inscription practices.

Moreover, Berger turns the phenomenological ideal of a primordial image into a trigger for a historical materialist analysis of images which departs from the idealism of a virgin gaze and white canvas, and instead explores the originarity of the image in relation to the proliferation of

strategies for registering and manipulating it. It’s not difficult to imagine these early anthropoids emitting sounds while simultaneously experimenting with modifying rocks into what became tools. In this sense, Berger’s phenomenological and semiotic references, find their current synthesis in a post-phenomenological approach which rethinks the originarity of the (human) image as always already captured by the originarity of technicity (Stiegler 1998), which allows its perception in the first place, turning the image into a way of seeing.

In this sense, “all images are [hu]man-made” (Berger 1972: 8), and all images imply knowing in their making as much as in their realisation—this knowing, and its temporality, being inscribed, or spatialised, in the proto-technical object, if we follow Stiegler. Thus, not only “the way we see things is affected by what we know or what we believe” (Berger 1972: 8), but also the images we make (or how we make them) are “affected by what we know or what we believe”. It seems there is always a grammar beyond an image, and this grammar proliferates because of its impossibility to grasp the fullness of an image. The relation between image and language is necessarily unsettled because, although they are co-originary, the ability of the latter to signify the former is never settled, and changes as a consequence of what we know or believe.

Since the beginning, the unsettling relation between (human-made) image (ways of seeing) and language (knowing) has turned into the unsettled relation between *techne* (the *techne* to make an image) and *episteme* (the socio-cultural milieu that allows for certain *technics*, and images, to emerge). In fact, the relation between *techne* and *episteme*—or, in Berger’s terms, seeing and knowing—is so deeply unsettled that philosophy, according to Derrida (1981) and Stiegler (2013), has been unable to think it properly since at least the time of Plato’s *Phaedrus* (370 BC). At this time, the unsettling relation between ways of seeing as always technical and knowing, and as always supported by technicity, is exemplified by the practice of “sophism”, understood as a break of the linkage between *techne* and *episteme*. In brief, sophists applied *techne* without a proper *episteme*, turning it into a poison, and not a cure, given the pharmacological nature of technology, already highlighted by Plato, and further emphasised by Derrida and Stiegler. Berger’s work also dives into this de-linkage, not at the level of philosophical inquiry, but at the level of the concrete production of artefactual images and the implicit process of knowing they support (or not). In doing so, Berger provides tactics for navigating their hidden epistemologies, within a Marxist framework in the most literal sense, through which to re-think the relation between the technical exteriorisations which produce images and the concept of alienation as the distancing-effect between the means of production and the knowledge of conditions. Alienation here is understood as

a departure from the relation between the proliferation of technical exteriorisations and the implicit (gnoseological and) epistemological framework they bring forth.

These inscriptions, or technical artefacts, are written, and represent images. Technical exteriorisations (images, in Berger's sense) attempt the inscription of the non-specialised sensory-motor schema of the hands and of the erect body which, as a consequence of its non-specialisation, produces the corticalisation of the brain (Leroi-Gourhan 1993). Non-specialization through externalisation opens up the therapeutic or curative side of the technical object, while its poisonous side emerges when externalisation starts to produce, instead, specialization. Factory workers are specialised because they only know how to operate the cog they see in front of them—without knowledge of the larger mechanism of which it is part—as dramatically shown in Farocki's (1968) film *Inextinguishable Fire*, depicting the production of Napalm by Dow Chemicals by units of workers producing single parts necessary to assemble the weapon. Workers remain unaware of the ultimate goal of their work, given the strict division of labour enforced by the factory. The distancing between seeing and knowing increases with computational technology, which often reduces cognitive workers into keyboard operators, pushing buttons completely unaware of the algorithmic consequence of their actions. The invisibility of the grammar is triggered by computational infrastructures that turn ideology (and knowledge) into a cloud of diffuse, networked services constantly extracting and processing data. This invisibility is a consequence of the specialization of knowing and it is de-linking from the ways of seeing that enable it in the first place. This alienation, or “generalized proletarianization” (Stiegler 2017) involves producers and consumers alike, and their common loss of *savoir-faire* (how-to-do), *savoir-vivre* (how-to-live) and *savoir-théorizer* (theoretical knowledge). Rethinking the linkage between knowing and seeing in relation to ways of machine seeing means understanding alienation as the theft of knowledge operated by technics animated by sick epistemological frameworks, and vice-versa. Furthermore, it means to understand that the technical object has been deprived of its ability to produce new long circuits of individuation and trans-individuation (Stiegler 1998), and instead, as we will see, produces short circuits of dividuation.

These tendencies proliferate and become more complex as the relation between visibility and invisibility keeps shifting and reaches a point where the demand to visualise drives a bulimic “drive to visibility” which aims at making everything visible (van Winkel 2005: 1). At the same time, the drive to visibility functions by hiding the processes through which visibility emerges. In the current technological milieu, big data is funneled from raw data assemblages into datasets that furnish the materials for the constitution of users'

algorithmic doubles (which arise from the extraction of their geolocation, frequency of communication, choice of topics, and so on). In this way, the algorithmic double becomes a data matrix for the molecular-tailored production of “missing visuals” (van Winkel 2005). In this architecture, missing visuals are those that appear at the level of the interface on the basis of the user's algorithmic double, and function as bait to keep the user clicking and producing new data to enrich the algorithmic double which, in turn, will produce new (missing) visuals. In short, the proliferation of big data's invisibilities function as means for the production of new data and, as a consequence, new missing visuals (Azar 2020). These invisibilities capture the subject in a circuit of algorithmic dividuations which produce the subject's algorithmic double, designed to overlook the production of (missing) images, which is where the circuit of individuation (and trans-individuation), possibly enabled by the technical object and its ways of seeing, is interrupted. The projection of the algorithmic double on the user's interface generates mirroring effects and digital echo-chambers where previous tastes and beliefs are fed back to confirm their positioning (and orientation) in the world. In Berger's terms, and put simply, new ways of algorithmic seeing allow the constant production of new visuals while simultaneously increasing the gap between seeing and knowing.

If this relation between seeing and knowing was once fundamental to acting in the world, the current distribution of agency across complex networks of non-human agents allows simultaneously more visibility—and, as a consequence, more knowledge about processes that before were not visible—and less knowledge about the very processes behind the way in which these new visualities are rendered visible. Although we see more, we are not given the instruments to understand the ways in which we see more—this being mainly a problem of property, for example in relation to our reliance on proprietary platforms and infrastructures, but also a problem related to the complexity of algorithmic networks, which in their operations often escape current epistemological frameworks.

Paraphrasing Stiegler (2015, 2019), if we see (and know) more, it is because of the computational ability of algorithmic networks to take over (and exponentially increase beyond human capacity for reason; what Kant defines as the analytical faculty of reason, or understanding). If we see and know less, simultaneously, it is because algorithmic networks colonise also the faculty of synthesis, “short-circuiting the deliberative functions of the mind” (Stiegler 2019: 26), to the point of moving beyond the given epistemological framework which embed them, or, otherwise said, to the point of destroying the possibility of theory (and of a human-accessible epistemology) as such (Stiegler 2015). In this context, how to recover a sense of agency when it has

been distributed to wider systems and assemblages, or more to the point, how to adjust to this reality (Tsing 2015)?

Evidently, images themselves have the ability to act in the world, and upon us, nowadays via the artefactuality of computational technology, characterised by its simultaneous invisibility and proximity. According to Stiegler, adjusting to reality consists of adopting a given (artefactual) reality rather than adapting to it (2010), with adoption standing for the ability to recover a sense of agency through long circuits of individuation that increases knowledge as *savoir-faire* (2018) and non-specialisation via the adoption of the technical object and its emancipatory potential. In a way, this was once the project of critical visual culture, to which Berger's essay contributes: to render visible the underlying conditions that allow us to see reality as it really is, instead of bringing to visibility new visuals by hiding the processes that make them visible, as happens with AI-driven big data analysis for example. Moreover, this is about the relation between what is visible and the names that we give to what is seen, as well as what is invisible—a politics of representation and nonrepresentation. But things are not so easy to comprehend as they were once thought to be, as images now proliferate and circulate in such vast quantities and are mostly made by machine for machines (Paglen 2016), and the knowledge related to this circulation can be only partially traced given the ability of machines to write according to a grammar not fully graspable by humans, further occluded by the proprietary nature of this form of writing, and knowing.

That machines play their part in this multilayered assemblage of images, turning images into operations, was famously articulated by Farocki: “images no longer represent an object but are part of an operation” (2004). In his video trilogy *Eye/Machine* (2001–3), and in his writing “Phantom Images”—key references for various articles in this journal—this is made evident, as the image-making machine no longer simply takes the position of a person but one of “intelligence”, combining what Farocki calls “the ill-considered notion of intelligence with an equally ill-considered subjectivity” (2004: 13) Images then are “operative” inasmuch as they do more than simply display themselves and offer themselves for human interpretation, but begin to interpret themselves. Machines “see”, but not simply like the eye of modernism (as in the case of the “kino eye” of Vertov's 1929 *Man With a Movie Camera*). Image-machines “act”, but no longer in the metaphorical or animistic “image-act” that art historian Horst Bredekamp saw in the war-photograph or the medieval altarpiece. In their social interaction with humans, images have always created, and not just depicted, reality. Now, as in the case of images created by machines for machines (synthetic datasets, QR codes, calibration

routines, and so on), image-machines create reality in their autonomous interaction with each other.

The removal of the human subject from the assemblage of seeing, or its role as a disposable element from which to extract value and knowledge, is symbolised by what Farocki calls the “suicidal camera” (2004). Suicidal cameras generate suicidal images: images produced by cameras located on remote missile systems and offered to the human operator until the moment the missile hits its target, at which point the camera disintegrates with the explosion and the transmission ends. If on the one hand, the human element fulfills a monitoring function, while the (ill) intelligence is delegated to the machine, on the other it is the pray of the image, and appears and disappears with it. Farocki's suicide images work well to explain the relation between the iconic indexicality of the algorithmic image understood as the possibility of its human-oriented use-value (otherwise said, as the form in which it is accessible by humans), and the speed at which it is forced to circulate and to be exchanged (according to its exchange value), which, in fact, compromises its human-oriented (iconic) indexicality.

It is within this type of operability—understood by expanding on Berger's Marxist framework—that the drive to visibility bargains truthfulness for novelty, and the relation between seeing and knowing becomes a problem of truth, or, to say it with Foucault, a problem of “games of truth” (Lorenzini 2015). In our computational context, iconic indexicality falls short not really because algorithmically-produced images look fake, rather the opposite—they can almost seamlessly pretend to be factual, indexing reality while being fully algorithmically-generated. For example, Generative Adversarial Networks (or GANs)—a framework in which two neural nets are dialectically opposed—have managed not only to allow real time facial re-enactment (as in the case of the DeepFakes) but also to process autonomously huge databases of real human faces and to generate new hyper-realistic faces that do not replicate any of the faces of the dataset. These AI human faces are both faces of missing humans (who do not exist in the actual world) and faces of algorithmically-generated ghosts, as paradoxically shown by the project *DoppelGANger.agency* (Azar 2018), which turns AI-generated faces into street posters for missing persons (Fig. 2).

These DoppelGANgers are a form of missing visuals emerging from the real human faces the GANs were trained on (Azar 2020), and as such can help to re-open the question of iconic indexicality in relation to algorithmic images that resemble their referents, even though their referents do not exist. In other words, the iconic indexicality of the algorithmic image serves to support the algorithmic indexicality which is at the core of the operational drive and consequent circulation of the image. This algorithmic indexicality allows for the clustering of images with similar parameters,

**Fig. 2** Examples from Mitra Azar, Doppelganger.agency (2019), doppelganger.agency



and to profile users exposed to the iconic indexicality of the image on the basis of those clusters, so as to predict the next (missing) image. In fact, the seeing and knowing on the side of the algorithmic indexicality is almost alien to the seeing and knowing happening on the side of the iconic indexicality, yet, at the same time one would not function without the other.

The relation between ways of seeing and knowing is always unsettled then because although the former is a grammatical object which, on the basis of its grammatology (of which the technical inscription is a trace), manifests itself as a knowable object, its very grammatology doesn't exclude—but rather depends on—the incomputable opening that the relation between ways of seeing and knowing inevitably produce. Ways of seeing and knowing—technical artefacts and human knowledge—are rooted in the non-specialised senso-motor and cortical potential defined by its capability of bifurcating, or finding ways of *savoir-faire* and *savoir-vivre* not fully inscribed by the ways of seeing and the forms of knowing they enable.

Finally, in this last section, we should say more about our editorial process and motivation for our work. We started this project with a series of small events at the University of Cambridge in 2016 (initially developed by Anne Alexander, Alan Blackwell, Geoff Cox and Leonardo Impett), to bring

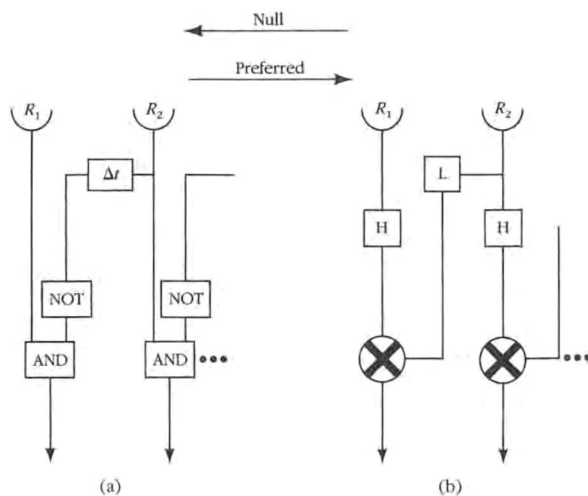
some of the politics of Berger to the discussion of machine vision, working across the Digital Humanities Learning programme and Computer Lab, and reaching out to other collaborators, and in turn to the contributors of this journal. Our starting point was to speculate on the comparison between Berger's *Ways of Seeing* and David Marr's *Vision*, a foundational text in computational neuroscience—and the axiomatic philosophy of vision behind machine vision for the past four decades. If Berger's concern was understanding how humans seeing with machines changed the ways in which they could represent the world, Marr was interested in the theoretical work necessary to make machines which could see. With a degree in mathematics, his early work on neuroscience was largely done at Cambridge, before moving to MIT; where in 1980, he was tenured at the Department of Psychology. He died later that year. His book, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, was published posthumously two years later, becoming the foundational text for the new field of computational neuroscience.

The explanation of the human visual system contained in Marr's *Vision* has remained the go-to theory of human vision for generations of computer scientists, for teaching and research: the most prestigious award in the computer vision community (awarded by the International Conference



on Computer Vision) is called the *Marr Prize*. What must be stressed, however, is that computer scientists didn't just adopt Marr's paradigm because it was the most up-to-date, successful, or accurate physiological theory of the human visual system. Computer vision is not, in general, meant to be a simulation of a biological visual system; that is not the epistemic role of a theory of human vision for the computer vision community. Rather than giving a biological precedent that machine vision researchers could copy, Marr's theory does almost the reverse: it understands human visual processes *in computational terms*. The parallels Marr draws between human and machine vision are far wider, and more structural, than that of the individual "neuron". Rather than anthropomorphic computing, Marr's *Vision* is technomorphic physiology.

Marr was interested in creating an information-theoretical (and even computer-scientific) model of biological seeing: before *Vision*, his early work was to propose computational models of the cerebellum, neocortex and archicortex (hippocampus) (Marr 1969, 1970, 1971). These were computational models not in the sense of computer simulations, but rather in that they used the metaphors of theoretical computer science—pattern recognition, memory systems—to explain the role and mechanisms of the human visual system. Electronics had long furnished a conceptual and notational system with which to describe human neuroscience (as in the Fig. 3 below, from Marr 1982); Marr's innovation was to extend this metaphor to information theory itself (Fig. 3). This is the basis of Marr's central axiom: *vision as an information processing system*.



**Figure 3–30.** (a) Barlow and Levick's (1965) model for directional selectivity connects two detectors to an AND–NOT gate, one via a delay. Thus the network does not respond to stimuli moving with roughly the right speed in the null direction. (b) Hassenstein and Reichardt's (1956) model operates on the same principle except that the delay is replaced by a temporal low-pass filter (L). H = high-pass filter.

**Fig. 3** Different models (in Marr, *Vision*, 1962: 163)

In the decades that followed Marr's *Vision*, a number of results supported his general theory of vision as information processing: that the human trichromatic system (Parkkinen and Jaaskelainen 1987), and the activation functions of primary visual cortex neurons (Field 1987), could be derived statistically from data about the natural world. In the so-called *Tri-Level Hypothesis*, David Marr and Tommaso Poggio (1976) outlined how these information processing systems—both biological and machine vision—ought to be understood at three separate levels:

- Computational level: what (computational) problem does the system solve, and why?
- Algorithmic/representational level: what statistical representations and operations does the system employ to solve this problem?
- Implementation/physical level: how are these statistical operations implemented, either as neural structures or computer code?

Marr's proposal, therefore, is to understand machine and human vision as a separable *stack*: we might share the same computational problem (say, face recognition) across multiple algorithms (eigenfaces, wavelet transforms), or the same algorithmic approach (e.g. learning sparse feature embeddings), through various physical implementations (neurons, GPUs, CPUs). Though fond of computational metaphors, Marr might not have recognised the concept of the *software stack*, which only really took off with the introduction of the OSI model for telecommunications in 1984 and the 2000s proliferation of stack models for web technology. Software systems are increasingly designed, conceived, operated, and programmed around the metaphor of the stack: not just from a technical perspective ("full-stack developer"), but also in their political-economic dimension ("full-stack business analyst").

Bratton's intuition (2016) about the "stack"—that the structural, transversal metaphors of software engineering and techno-capitalism are equally useful critical tools—is relevant to our project. We can expand Marr and Poggio's tri-level model of human vision:

- (1) Social level (where are such systems deployed, by whom, for what purpose)
- (2) Computational level (which problems are being solved: e.g. "object detection")
- (3) Data level (who labels, which images are chosen, who takes the photographs)
- (4) Algorithmic/representational level (e.g. Siamese convolutional neural network with Adam gradient descent optimization)
- (5) Implementation/physical level (abor. Tensorflow on cuDNN/CUDA on Nvidia GPU)

(6) Philosophical/axiomatic level (e.g. *vision as inverse graphics*)

This vertical cartography of machine vision—though far coarser than Joler and Pasquinelli’s Nooscope—nonetheless exposes black-spots in our collective critical dissection of the machine vision stack. We have brought forward an excellent critical understanding of the datasets of machine vision and of their social applications, but have almost nothing to say about the ideological content of specific algorithms, technical axioms, compiler languages, or massively parallel silicon implementations.

One of these technical axioms comes from the work of Marr himself: the computational paradigm of [robotic] *Vision as Inverse* [computer] *Graphics*. Marr outlined what he saw as the three stages of vision in an information-processing pipeline in the construction of:

- (1) Firstly, a 2D primal sketches: including edge detection, silhouettes, etc.;
- (2) Subsequently, 2.5D images, including textures, foreground and background;
- (3) Finally, a full 3D model of the environment.

Here, the machinery of machine vision is chained to that of the early video-games industry (a link which extended, since widespread use of CNNs in 2012, to the hardware level of the stack—how Nvidia overtook Intel); through the implicit separation of mesh and texture, object and background, image and sprite. This is what makes it possible to create *synthetic datasets* in machine vision; to use CGI to synthesise models of faces, cars, streets, or warzones, which machine vision algorithms then learn from.

Where Berger sets out a dialectical relation between seeing and knowing, and the role of knowing in seeing, Marr’s suggestion is that the visual system is fine-tuned to efficiently compress (i.e. recognise) visual stimuli that it has evolved to encounter (faces, shadows, motion); and in doing so has a kind of implicit, embodied knowledge about the natural world and its images. For Marr, then, and for computer vision scientists after him, seeing is not so much a product of *knowing* (nor, we might add, of *belief*, and thus ideology) as a product of *data*. A consequence of this perspective—of the unmediated nature of visual representation—is an inability to deal with the ambiguities and inconsistencies of visual perception, whose technical implications have been highlighted by Aaron Sloman, former chair of AI at Birmingham University: “that common idea is mistaken: visual systems do not represent information about 3-D structure in a 3-D model... but in a collection of information fragments, all giving partial information. A model cannot be inconsistent. A collection of partial descriptions can” (2011). Sloman’s

critique of Marr here echoes Berger: “the relation between what we see and what we know is never settled.”

As Berger wrote of photography, the technical aspects of machine vision “are not, as is often assumed, a mechanical record”, rather they are saturated with ideology. What is to be done? Taking political issue with the corporate machinery of machine vision is, for at least two reasons, an intrinsically slippery task. First, because this corporate machinery is difficult to delineate; it’s as far as it’s possible to get from the perfect-substitute-producing factory of textbook economics (Srnicsek 2016). It includes conventional private companies which sell machine vision for profit—but also corporate-funded AI labs which publish research openly in the scientific community; privately-funded open-source software libraries, on which much of AI depends; and infrastructure, from cloud computing to hard silicon. Google does all of these; selling machine-vision-as-a-service (Google Vision AI), publishing open research (Google Brain), producing fundamental shared open-source libraries (Google TensorFlow), and selling the cloud services (Google Cloud) and even the chips (Google’s Tensor Processing Units). Microsoft, Amazon and others have similar profiles.

But there is a second reason for which the intersection of political critique and corporate research in machine vision is so confusingly nuanced. The academia-versus-industry debates of the past five decades had clear demarcation lines; for instance, in the tobacco industry’s long refutation of the link between smoking and cancer, or the oil industry funding research which attempted to obscure the links between man-made emissions and climate change. In the case of machine learning and machine vision, industry (including corporate AI labs) publicly agrees, in many cases, with its critics. Microsoft, for instance, claims “inclusiveness” and “fairness” as two of the six guiding principles for Responsible AI. This is not simply a marketing proposition—a paper on fairness in AI from Microsoft Research won Best Paper at CHI 2020 (Madaio et al 2020), and similar internationally-relevant research is produced by other major corporate tech players. Through efforts like increasing demographic diversity in the workplace, technology companies have endlessly publicised those occasional win-win situations in which a degree of corporate fairness increases long-term profitability. But it’s not yet clear to what degree this mandate for socially—and ethically responsible machine learning extends to activities which might seriously endanger a company’s bottom line: well-publicised ethical AI policies notwithstanding, Amazon, Google, Microsoft, Oracle and IBM all still bid for a \$10bn cloud computing contract with the US Department of Defense in 2018–19. When Google dropped its bid in October 2018, it claimed the first reason behind this decision was that “we couldn’t be assured that it would align with our AI Principles”; in practice, it was responding to seven months of significant organised pressure

from Google workers, including several resignations and a widely-signed employee petition.

Where policymakers and journalists tend to talk about analysis of abstract *data* (almost an implication of Excel spreadsheets), a very large proportion of today's AI controversies centre around *vision*. The debate over Google's involvement in the DoD contract started when it was revealed that Google supplied machine vision technologies for the automated video analysis of drone footage. Amazon (who, at time of writing, was attempting a legal challenge to the contract being awarded to Microsoft) continues to publicly host an online demo of Amazon Rekognition, their machine vision platform, for the defence industry; highlighting Amazon's ability to "analyze images and recognize faces, objects, and scenes", and finding the "likelihood that faces in two images are of the same person, in near real-time". Without a hint of irony, Amazon is at the same time the principal sponsor of a 3-year \$20 million funding program of the U.S. National Science foundation on Fairness in AI.

Such paradoxical situations, in many cases, are not borne of inconsistency or hypocrisy, but *irresolvability*—a general symptom, for Matthew Fuller and Olga Goriunova (2019), of the post-Cold War society. In a structurally unequal society, it is exceedingly difficult to make a "fair" algorithm; and it is effectively impossible to make an algorithm which is both fair *and* effective. Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan (2016) have shown, in the most general case it is *impossible* to classify data in a way that meets several well-established definitions of fairness (lack of active discrimination; balanced false positive rates; balanced false negative rates). All three definitions of fairness also produce classifications which are, in general, different from the highest possible accuracy (and therefore, in most situations, different from the highest potential generated profit). There are two special cases where it is possible to satisfy all fairness conditions: *perfect predictions* (total information—nobody is ever misclassified because of their background), or *equal base rates* (i.e. a society/ground-truth where different groups have statistically identical behaviours). In a society which is unfair, a classification-machine will *always* be unfair (in at least one sense). What is so important about Kleinberg et al.'s mathematical proof is that it's fundamentally about the final *classification results*, regardless of how any individual decisions are taken. The impossibility of fair classifiers in an unfair world, therefore, is equally true if those classifiers are human. In the specific case of machine vision, the risk of unintended discrimination through high-dimensional correlation has greatly increased with the advent of deep convolutional neural networks (Impett 2018), compared to their simpler (and less powerful) predecessors which relied on hand-crafted geometric features.

Nonetheless, the research behind such a profound mathematical result—with enormous political implications for human decision-making—would not have existed were it not for current debates around machine-learning-based classification systems. Our hope, in this journal and beyond, is that critical work on machine vision will lead to similarly profound political insights. "To ask whether machines can see or not is the wrong question [...] rather we should discuss how machines have changed the nature of seeing and hence our knowledge of the world" (Cox 2016). In this sense, the project of algorithmic literacy behind *Ways of Machine Seeing* mirrors Berger's didactic project of visual culture.

Beyond being a powerful or harmful new technology in its own right, machine vision gives us a new, precise set of metaphors with which to think about vision differently—in computational, biological, aesthetic, and consequently *political* terms—to enhance our ability to see and thereby act in the world.

**Acknowledgements** *Ways of Machine Seeing* first emerged as a workshop organised by the Cambridge Digital Humanities Network, convened by Anne Alexander, Alan Blackwell, Geoff Cox and Leonardo Impett, and held at Darwin College, University of Cambridge, on the 11 July 2016, and which has involved many more people and institutions since. We wish to thank our reviewers—without whom the journal would not have been possible—for their work, their detail and their rigour: Alan Blackwell, Annet Dekker, Andrew Dewdney, Jonathan Impett, Paul Melton, Gabriel Menotti, Darío Negueruela del Castillo, Winnie Soon, and Pablo Rodrigo Velasco González.

Most of all, we must thank the authors contained in this journal, for the many hours of uncompensated labour, and for the quality of their articles: María Jesús Schultz Abarca, Peter Bell, Tobias Blanke, Benjamin Bratton, Claudio Celis Bueno, Kate Crawford, Iain Emsley, Abelardo Gil-Fournier, Daniel Chávez Heras, Vladan Joler, Nicolas Malevé, Lev Manovich, Nicholas Mirzoeff, Perle Møhl, Bruno Moreschi, Fabian Offert, Trevor Paglan, Jussi Parikka, Luciana Parisi, Matteo Pasquinelli, Gabriel Pereira, Carloalberto Treccani, Rebecca Uliasz, and Manuel van der Veen.

## References

- Agre PE (1997) Toward a critical technical practice: lessons learned in trying to reform AI. In: Bowker G, Gasser L, Star L, Turner B (eds) Bridging the Great Divide. Social Science, Technical Systems, and Cooperative Work, Erlbaum
- Azar M (2019) Pov-data-doubles, the dividual, and the drive to visibility. In: Natasha L (ed) Big data—a new medium? Routledge, London, pp 177–190
- Azar M (2018) DoppelGANger.agency. Available from: <http://doppelganger.agency/>.
- AWS (2017) Amazon Rekognition Demo for Defense. Blog, August 7. <https://aws.amazon.com/blogs/publicsector/amazon-rekognition-demo-for-defense/>
- Benjamin W (2008) The work of art in the age of mechanical reproduction. Belknap Press of Harvard University Press, Cambridge
- Berger J (1972) Ways of Seeing. BBC and Penguin.

- Crawford K, and Paglen T (2019) Excavating AI: The Politics of Images in Machine Learning Training Sets. [www.excavating.ai](http://www.excavating.ai)
- Cox G (2016) Ways of Machine Seeing. *Unthinking Photography*. [unthinking.photography/articles/ways-of-machine-seeing](http://unthinking.photography/articles/ways-of-machine-seeing)
- Farocki H (2004) Phantom Images. *Public* 29:12–22
- Farocki H (1968) *Inextinguishable Fire*. Film.
- Derrida J (1976) *Of Grammatology*. John Hopkins University Press, Baltimore
- Derrida J (1981) *Disseminations*. Athlone Press, New York
- Ferreira da Silva D (2017) “1 (life) ÷ 0 (blackness) = ∞ – ∞ or ∞ / ∞: On Matter Beyond the Equation of Value.” *e-flux* 79, February. [www.e-flux.com/journal/79/94686/1-life-0-blackness-or-on-matter-beyond-the-equation-of-value](http://www.e-flux.com/journal/79/94686/1-life-0-blackness-or-on-matter-beyond-the-equation-of-value)
- Field D (1987) Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc Am* 4(12):2379–2394
- Fuller M, Goriunova O (2019) *Bleak Joys: Aesthetics of Ecology and Impossibility*. University of Minnesota Press, Minneapolis
- Leroi-Gourhan A (1993) *Gesture and Speech*. MIT Press, Cambridge
- Lorenzini D (2015) “What is a regime of Truth?”. *Le Foucauldien* 1/1. Open Access Journal for Research along Foucauldian Lines. 2015.
- Impett L (2018) Artificial intelligence and deep learning: Technical and political challenges. *Theory & Struggle* 119:82–92
- Kleinberg J, Mullainathan S, and Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores, arXiv preprint
- Mackenzie A, Munster A (2019) Platform seeing: image ensembles and their invisibilities. *Theory Cult Soc* 26(5):3–22
- Madaio M, Stark L, Wortman Vaughan J, and Wallach H (2020) Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI, 2020 CHI Conference on Human Factors in Computing Systems.
- Made by Machine: When AI met the Archive (2018) Dir. Hannah Fry. BBC Four, September 5th, <https://www.bbc.co.uk/programmes/b0bhwk3p>.
- Marr D (1982) *Vision: a computational investigation into the human representation and processing of visual information*. MIT Press, Cambridge
- Marr D, Poggio T (1976) From understanding computation to understanding neural circuitry. MIT Technical Report, Cambridge
- Mbembé A (2003) *Necropolitics*. *Public Cult* 15(1):11–40
- Merleau-Ponty M (1964) *Phenomenology of perception*. Routledge, New York
- Mirzoeff N (2011) The right to look. *Crit Inq* 37(3):473–496
- Myers West S, Whittaker M, and Crawford K (2019) *Discriminating Systems: Gender, Race and Power in AI*, AI Now Institute, New York University, [ainowinstitute.org/discriminatingystems.html](http://ainowinstitute.org/discriminatingystems.html)
- The Next Rembrandt (2016) Wunderman-Thompson/Microsoft, [www.nextrembrandt.com](http://www.nextrembrandt.com)
- Nietzsche F (2006) *On the genealogy of morality*. Cambridge University Press, Cambridge
- Paglen T (2014) “Operational Images.” *e-flux journal* #59, [www.e-flux.com/journal/59/61130/operational-images](http://www.e-flux.com/journal/59/61130/operational-images)
- Paglen T (2016) Invisible images (your images are looking at you). *The New Inquiry*. <https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/>.
- Parkkinen J, Jaaskelainen T (1987) Color representation using statistical pattern recognition. *Appl Opt* 26(19):4240–4245
- Plato (2004) *Meno*. Focus Publishing, Newburyport, MA
- Sloman A (2011) “What’s vision for, and how does it work? From Marr (and earlier) to Gibson and Beyond.” Birmingham Vision Club, [www.cs.bham.ac.uk/research/projects/cogaff/talks/sloman-beyond-gibson.pdf](http://www.cs.bham.ac.uk/research/projects/cogaff/talks/sloman-beyond-gibson.pdf)
- Stiegler B (1998) *Technics and Time I-II-III*. Stanford University Press, Stanford
- Stiegler B (2010) *Taking care of youth and the generations*. Stanford University Press, Stanford
- Stiegler B (2013) *What Makes Life Worth Living: On Pharmacology*. Polity Press, Cambridge
- Stiegler B (2015) *States of shock: stupidity and knowledge in the 21st century*. Polity Press, Cambridge
- Stiegler B (2017) *Automatic Society, vol I*. Polity Press, Cambridge
- Stiegler B (2019) For a neganthropology of automatic society. In: Pringle T, Koch G (eds) *Machine*. Meson press, Lüneburg
- Srnicek N (2016) *Platform Capitalism*. Polity, London
- Tsing AL (2015) *The Mushroom at the End of the World: On the Possibility of Life in Capitalist Ruins*. Princeton University Press, Princeton, NJ
- Van Winkel C (2005) *The Regime of Visibility*. NAI, Rotterdam

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# On the data set's ruins

Nicolas Malevé<sup>1</sup>

Received: 5 August 2020 / Accepted: 14 October 2020 / Published online: 11 November 2020  
© The Author(s) 2020

## Abstract

Computer vision aims to produce an understanding of digital image's content and the generation or transformation of images through software. Today, a significant amount of computer vision algorithms rely on techniques of machine learning which require large amounts of data assembled in collections, or named data sets. To build these data sets a large population of precarious workers label and classify photographs around the clock at high speed. For computers to learn how to see, a scale articulates macro and micro dimensions: the millions of images culled from the internet with the few milliseconds given to the workers to perform a task for which they are paid a few cents. This paper engages in details with the production of this scale and the labour it relies on: its elaboration. This elaboration does not only require hands and retinas, it also crucially mobilises the photographic apparatus. To understand the specific character of the scale created by computer vision scientists, the paper compares it with a previous enterprise of scaling, Malraux's *Le Musée Imaginaire*, where photography was used as a device to undo the boundaries of the museum's collection and open it to an unlimited access to the world's visual production. Drawing on Douglas Crimp's argument that the "musée imaginaire", a hyperbole of the museum, relied simultaneously on the active role of the photographic apparatus for its existence and on its negation, the paper identifies a similar problem in computer vision's understanding of photography. The double dismissal of the role played by the workers and the agency of the photographic apparatus in the elaboration of computer vision foreground the inherent fragility of the edifice of machine vision and a necessary rethinking of its scale.

**Keywords** Computer vision · Data set · Photography · Imageability · Scale · Micro-labour · Le musée imaginaire · ImageNet

## 1 Introduction

In *On the Museum's Ruins*, Douglas Crimp (1980), reflects on the proximity between the museum and the mausoleum. The museum is the place artworks, disconnected from the practices and lived conditions that breathe life into them, have come to die. Crimp's text dramatises a moment of change in the museum's history. The museum, once considered the shelter of a selection of significant artworks, evolves into a site, where genres that were deemed secondary are excavated from the storerooms and given exposure. The inclusion of salon painting in the installation of nineteenth century painting at the Metropolitan Museum signals, for Crimp, the demise of modernism. And it triggers, among

his contemporary critics, a wave of criticism against the lack of rigour, the laissez-faire attitude of a postmodern era. Refusing to embrace such criticism and its nostalgia, Crimp moves on to articulate a distinct critique of the new museal institution. The pretension to knowledge from the part of the museum is to obtain a coherence from objects and works extracted from increasingly heterogeneous sources with art history being assigned the role of homogenising the disparate museum's contents. To extract knowledge through a dialogue between regularised fragments is the condition at which the museum resists its assimilation to the mausoleum. Crimp sees this tendency to a larger inclusion expressed by André Malraux's *Le Musée Imaginaire* in which oeuvres from all civilisations can be confronted and compared. As Crimp remarks, Malraux's project emphasises the fact that art history cannot produce a universal plane of comparison without another organising device: photography. Photography, here, acts as a leveller. Every art object that can be photographed can enter Malraux's museum. Through the

✉ Nicolas Malevé  
maleven@lsbu.ac.uk

<sup>1</sup> Centre for the Study of the Networked Image, London South Bank University, 103 Borough Road, London SE1 0AA, UK

photographic medium, a part of a plate can be compared to a detail of a sculpture, an enlarged brush stroke to a mural painting. Photography is a condition for the inclusion of objects and for the museum's knowledge production. But to function this way, photography must be made a transparent vehicle. As soon as photography itself is considered as an autonomous agent, heterogeneity comes back and the museum's pretension to knowledge is in jeopardy. The ruins, for Crimp, are not only those of the museum's contents which roots are severed from the soil that gave them birth. The ruins are those of the museum itself, an institution of knowledge production based on the active yet transparent role of a levelling device: the photographic apparatus and the scale it introduces. Repressing the apparatus through which it obtains its coherence makes it blind to the artificiality and the limits of the knowledge it produces.

The current text is not an essay on art history. In the following pages, I will address the problems of a discipline of computer science, computer vision (CV), in which engineers train machines to make sense of images and, for that purpose, assemble huge collections of photographs. In this endeavour, I won't use Crimp's critical dialogue with Malraux as a strict analytical framework, but rather as an insistence and call to increase sensitivity to the relation between the photographic device, scale and a pretension to knowledge in the vast sets of images assembled in computer vision. It is with the notion of selective use of photography, and its repression, related to the homogenisation of a potentially limitless source of visual imagery, that I interrogate the formation of computer vision. To understand the relevance of such a conceptual device for CV, it will be necessary to take a step by step approach to photographic practices at the core of the discipline.

## 2 Computer vision, learning to see by example

Nowadays, contemporary computer vision is increasingly understood as a sub-field of artificial intelligence and machine learning. This is the disciplinary affiliation given in introductory courses and manuals (Brownlee 2019; Fullscale 2019; Wikipedia Computer vision no date). To construe CV first and foremost as a field of artificial intelligence (AI) means several hierarchies are taken for granted. For instance, it means that a discipline, AI, applies the same techniques, with a few variations, to different kinds of data. In computer vision, visual input is often unproblematically conflated with photographs or media abiding to the representational codes

of the photographic image.<sup>1</sup> Photographs understood as data are presented as passive samples awaiting the mining of algorithms to be made meaningfully part of computational systems. In such a view, agency is located on the side of the algorithm, and data, as the name suggests, is simply given.

The work of Fei-Fei Li, leading computer vision scientist, director of Stanford AI, and initiator of the ImageNet data set, shows how a different valence can be given to photographs. As she (GoogleTechTalks 2011) states, "another way to understand CV is through the evolution of its data sets". Taken seriously, data sets challenge an algorithm-centric view of CV. The algorithm becomes relative. For Li, data sets are no longer suppliers of data, they are tools to formulate CV's problems: they engage actively in the modelling process. To understand such a claim, we need to look at how modelling functions in contemporary machine learning systems.

Traditional AI and early computer vision relied on explicit modelling. In a framework, where modelling is explicit, developers design themselves a model that matches the complexity of the problem domain. For instance, a cat's face is decomposed into simple shapes like a circle for the face, two triangles for the ears and two circles for the eyes. This approach is only efficient for a limited amount of cases. Its advantage is to produce a legible model: circle + 2 triangles + 2 circles = a cat face. Such a model may function in a strictly controlled environment, but leads to overwhelming problems in real-world scenarios, where the cat may appear from profile, with eyes closed, at rest or jumping. Additionally, one cannot expect the animal to be perfectly centred and illuminated and partial occlusions or unexpected perspectives often change the organisation of the patterns. As neuroscientist David Marr remarked, photographs rarely follow a predictable order. Noticing the almost despairing feeling of early computer vision researchers, he concluded that "practically anything could happen in an image and furthermore that practically everything did" (Marr 1982, p. 16). Every time a photograph deviates from an expected pattern, the algorithm needs to be updated and optimised. Such an approach lacks generalising power as the algorithm must not merely discriminate a pattern adequately, it also needs to make sense of the other patterns interfering with what it attempts to detect (Deng et al. 2010). Furthermore, the developer needs to learn in detail about the object to detect and to analytically decompose it before writing the code. Under this paradigm, to write a cat classifier, a programmer needs to become a feline expert.

<sup>1</sup> See, for instance, one of the most cited papers in the discipline, "A large-scale hierarchical image database – ImageNet" (2009), in which image and photograph are used interchangeably.

In contrast, current techniques of machine learning based on neural networks do not rely on a previous analytical decomposition. The developer assembles a data set reflecting the variations of the domain under study and utilises automated means to calculate an optimal function that treats the features of the data as parameters. In computer vision, this technique, at its most simple level, uses large visual databases in which discrete units such as pixels can be considered as data points. Common techniques of machine learning in computer vision are said to be “supervised”, which means that the data is curated (Beheshti et al. 2016) to provide examples from which the machine learning algorithm extracts regularities: the software “learns by example”. To come back to the case of the cat, in the data-oriented paradigm, the developer does not try to decompose the animal in distinct shapes and explicitly summarise their relations. Instead, she curates a large series of photographs, where the cat is displayed in various positions, and lets the algorithm detect the regularities traversing the various samples. Through this phase of “learning”, the algorithm produces the model of the cat.

The change in algorithmic design does not merely correspond to an evolution of the technical process. It provokes more largely a reconfiguration of the positions of the actors in the field. The analytical decomposition of the problem is now replaced by the production of a data set exhibiting the regularities that define the problem domain. Engineers are said to write programs that “discover” the rule inherent to the data (Simpson Center for the Humanities UW 2017). However, these data sets do not appear by magic, they need to be curated, assembled, maintained and annotated. Concretely, the modelling is outsourced to those who curate and annotate the data set. The annotators and curators are in fact implicitly coding the model that will be discovered algorithmically. The engineers say the model is learned end to end. This means in fact that it doesn’t learn from them anymore. In the current machine learning paradigm, the engineer doesn’t need to be a feline expert to produce a cat detector, but the engineer relies on a population of curators and annotators actively engaged in defining what counts as photographs of cats. The paradigm change in machine learning has externalised the modelling process and, by doing so, has produced a new division of labour.

If we take this seriously, then the data set becomes a site, where computer vision is made, rather than a component of its supply chain. The techniques to compose a data set are not subsidiary to the problem of computational vision, they are integral to it. But as important as data sets are for machine vision, two intertwined aspects of their production are generally overlooked: the micro-labour of annotation and the role given to photography in the process. To examine the labour that goes into CV and its photographic dimension, I will analyze how ImageNet, one of the largest

database of human annotated visual content to date (Deng et al. 2009) has been assembled. The ImageNet project is a collection of images for visual research that offers tens of millions of images manually annotated, sorted and organised according to a taxonomy. It aims to serve the needs of computer vision researchers and developers for training data. Due to its size<sup>2</sup> and its extensive annotations, ImageNet has become de facto the most used knowledge base in the world of computer vision (Fei-Fei 2010). Following the work of the annotators who selected the photographs and reflecting on the role given to photography, I will try to understand their roles in the process and the reasons why they remain treated as contingent rather than crucial to the definition of what it means for machines to see.

### 3 The elaboration of CV

Computer vision data sets depend on a standing reserve of large volumes of manually annotated photographs. Advertising a new update of the Google Photo service, Chuck Rosenberg (2013) wrote that, thanks to the advances of computer vision, users could search their own images *without having to manually label each and every one of them*. This tedious work can be executed automatically from now on. Yet, to automate this task, large volumes of annotations are themselves necessary. While the users of the new Google software may not be obliged to manually tag their photos anymore, thousands of hands on keyboards, retinas glued on screens are producing an unprecedented amount of labelling and descriptions to train the algorithms that automate the user’s tagging. Perhaps, the amount of annotation work involved in the production of data sets is more impressive than the amount of photographs included in ImageNet. After all, 14 million images are only a fraction of the monthly 575 million public uploads of photos on a platform like Flickr. (Smith 2019) The work of manually cross-referencing and labelling the photos is what makes data sets like ImageNet so unique. In fact, there has been rarely in history so many people paid to look at images and report what they see in them (Vijayanarasimhan and Grauman 2009). The automation of vision has not reduced the amount of eyeballs looking at images, of hands typing descriptions, of taggers and annotators. On the contrary, it has increased their number. Yet what has changed is the context in which the activity of seeing is taking place, how retinas are entangled in heavily technical environments and how the speed at which candidate images must be evaluated. Due to the pressure of generating annotations at the scale of the web, the process

<sup>2</sup> The average number per categories is 10.5k, see ImageNet\_2010.pdf.

needs to be massively “parallelized”.<sup>3</sup> A parallel architecture optimises the calculation and makes maximal use of all the computational resources of a machine. On the Amazon Mechanical Turk (AMT) platform, where a large part of the annotation effort is produced, the workers are treated as computational processes. The workers, the “Turkers”, are abstracted away. The advantage of having parallel processes is that the execution time of a given task can be divided among the available workers/processes. Li calculated the amount of human labour required for ImageNet this way: estimating that a person can annotate two images per second, such that 19 human years are necessary for an individual working 24 h a day to verify the tens of millions of images in her data set. AMT allows her to compress the 19 years into 2 years by providing the necessary workforce to handle the workload in parallel (GoogleTechTalks 2011).

Apart from speed, there is another less “technical” reason to adopt such architecture. The parallel architecture isolates the workers and makes it difficult for them to create bounds, to build a collective identity and to unionise. The Turkers have no name, only an anonymous identifier and the platform doesn’t provide any mean of communication between them (Irani and Silberman, 2013). There are political reasons for crowdsourcing platforms to fear workers’ unions: their working conditions are exploitative. AMT offers people with large data sets the possibility to outsource the annotation work using their massive workforce. The tasks on AMT are rewarded with micro-payments. An annotation is estimated between 1 and 4 cents and the estimated hourly revenue for a Turker is approximately two dollars (Hara et al. 2018). The annotator must find an optimal trade-off between the precision and attention required to make a good enough annotation and a speed that allows her to “maximise” her financial gain. The hourly rate depends on the task and access to the most lucrative tasks depends on the age, the origin and the class of the workers as one needs to be “culturally compatible” (Irani 2015, p. 726) with the request. Annotation tasks in particular require to quickly learn the knowledge relevant to very different subjects including religious differences (i.e., be able to discriminate between photographs of Catholics and Old Catholics, archbishops and archpriests, Buddhist and Zen Buddhists from sources as various as the Vatican website, English tabloids, and family albums), military ranks and uniforms (i.e., differentiate between Redcaps and Green Berets, adjutant generals and generals, sergeants and first sergeants, recruiting sergeants and gunnery sergeants, from sources ranging from military websites, news outlets, TV series or wedding pictures), or bacterial species (i.e., identify

the *Bacillus anthracis*, the spirillum, the clostridium perfringens, or the gonococcus from microscopic imagery).

The workers, if they want to make a living, need to work at a pace that barely allows them to see the images. For the annotators, structurally, the “glance” is the norm. Speed is built in the platform economically. Training sets must be produced fast. Lots of workers are mobilised intensely for a short period of time. Through the interface of the AMT, the requesters are managing the cadence of the annotation work. They want to ensure the workers go fast enough to match production deadlines. And, at the same time, they attempt to preclude them from overlooking their task to avoid drops in quality. The interfaces of annotation are designed to control workers’ productivity, to find the optimal trade-off between speed and precision. The time estimation for labelling tasks is especially difficult as it varies according to the annotator’s experience and the nature of the photographs. Andrej Karpathy, now Tesla’s director of AI, reported that when he annotated an ImageNet’s recognition challenge’s data set, the labelling started at a rate of one image per minute and decreased with the accumulated experience (Karpathy 2014). But, even if progress was noticeable, the rate was not constant as some images like those depicting some particular breeds of animals required a longer time and additional research. To cope with the variability of the annotation process, numerous techniques are currently tested to streamline it. To begin, as manual labelling is costly, the priority must be given to the annotations producing the highest information gain. As Vijayanarasimhan and Grauman put it, unlabelled images must be ranked “according to their expected ‘net worth’ to an object recognition system” (Vijayanarasimhan and Grauman 2009). The cost of the labelling effort leads the computer vision researchers to approach visual content in the form of informational currency and attention scarcity. This approach informs the architecture of annotation workflows, wherein a pre-labelling attempt is made by an algorithmic detector and corrected by human annotators (Papadopoulos et al. 2016). Pattern recognition techniques are used to absorb the bulk of the “easy” detection work, spot the potential targets and redirect the “ambiguous” cases to the human annotators (Vijayanarasimhan and Grauman 2009). While these techniques of semi-automation are in their early stages, their existence indicates the anxiety of the data set makers to control the annotator’s attention and prevent her distraction. The requesters invest in the annotator’s attention and treat attention as an asset that needs to be protected. Besides guiding the annotator’s eyes to specific regions of interest and regulating their attention, researchers are exploring how to augment the annotators’ ability to absorb visual content. For instance, Krishna et al. (2016, p. 2) claim to have devised a technique of rapid serial visual

<sup>3</sup> In Informatics jargon, to “parallelize” a process means to design a workflow where two processes can take place without interference.



presentation that allows workers to produce one “hit”<sup>4</sup> in 100 ms by immersing them in an uninterrupted visual flow. As the volume of requests augments, such experiments indicate that the unit of measurement for hits is moving towards the millisecond. Finally, others are concentrating on evaluating workers’ performances over large periods to identify the annotators able to sustain the rhythm over time without decline in submission quality and sifting out the “satisficers” (Hata et al. 2017) who strive to do the minimal amount of work to meet the acceptance threshold. In response, the workers themselves try to figure out what is a good ratio between speed and accuracy. On social media platforms like Reddit, AMT workers exchange tips about “good paying tasks” and compare their performances. To be able to estimate the amount of work required for a task is a matter of survival in the crowdsourcing environment: both for the data set maker who is under pressure to deliver in a competitive environment and the annotators who internalise the system’s speed.

Where does this analysis of the annotation architecture leave us? First, it emphasises that data is not given. Data needs to be produced and engineers are deeply involved in setting up the material conditions of production of said data. If the work of curating data sets and their annotation is central, so is the management of the populations of workers involved. More importantly perhaps, we see how the notion of scale transpires into the whole process. With this notion of scale, we find a point in the midst of the technicalities of CV, where the discussion of *Le Musée Imaginaire* starts to resonate. A scale here is not a simple measurement of the accumulation of more material which has for consequence an accumulation of knowledge. Scaling corresponds to a production of difference not just to an increase of the same. If Malraux emphasises that his contemporaries have access to more artworks than their predecessors, he also insists that their relation to art differs in kind as well as in breadth. They have access to more categories of works and to a larger selection of works inside these categories. The expansion of categories and expansion of intra-categorical comparisons are the engines of a different relation to artworks and their representations. In this perspective, *Le Musée Imaginaire* is more than a printed book, it is an operation that brings the museum to the scale of the printed press.

Even if Li’s concerns feel remote from Malraux’s concerns, to read *Le Musée Imaginaire* from this perspective, invites to pay closer attention to the generative character of scale. It shifts our attention from quantity increase (the addition of discrete content) to scale as a qualitative architecture of relations. And it helps take the measure of what is at

stake when Li declares ImageNet is meant to bring computer vision to the scale of the web. Between the French Minister of Culture and the director of the Stanford AI lab, there is a resonance in the investment in scale as vector of knowledge transformation. There are also striking differences. For instance, in Malraux’s essay, one could hardly find any reference to the labour involved in the production of the photographs or any economical consideration. Malraux’s museum is abstracted away from these concerns, as the art historian benefited from the budgets and apparatus of the Ministry of Culture. Li, on the other hand, brings the logistics and the economics of scale to the front. When she states that a crucial step to advance her discipline is to resolve the scale of the problem, she makes clear that this involves getting one’s hands dirty with issues of management and control.

This dissonance calls for further elaboration on the nature and dynamics of the knowledge that is produced. A scale, manifested in logistics and industrial infrastructure, is deeply generative. As media scholar Matthew Fuller writes, a scale provides “a certain perspectival optics by which dimensions of relationality and other scales may be ‘read’ [...]” (Fuller 2005). From a newly sensible scale, it becomes possible to “read new dimensions of potentiality” (Fuller 2005). Li’s work is an articulation of scales. It is the millions of images with their labels, the articulation of vision with the decomposition of work in micro-tasks and micro-payments. The problem of scale is an articulation of both the macro and the micro. 14 million images need a model of vision, where the milliseconds are the unit of measure for perception. When thinking about the problem of scale of computer vision, we need to keep in mind the opposite poles of the problem’s dimension: on one hand, the infinitesimally small units of perception (the eye saccades), the miniaturisation of the work process in discrete hits that can be performed at full speed; and on the other hand, the massive amounts of photographs available on digital platforms and the vast population of precarious workers ready to annotate on demand.

This articulation of scale brings about a new distribution of labour crucially affecting the task of modelling. A task that was previously in the hands of the engineers is distributed among the annotators. The engineers delegate a series of crucial decisions, while keeping control over the way the problem is formulated and its interpretation. Through the interfaces and contracts of AMT, the engineers design the frame of “transepistemic relations” (Knorr-Cetina and Malakay 1983).<sup>5</sup> For Karin Knorr-Cetina, an example of a transepistemic relation can be found in the interaction between a funding agency and the researchers it supports. Often the

<sup>4</sup> In Amazon Mechanical Turk’s jargon, a “hit” stands for a “human intelligence task”.

<sup>5</sup> “Transepistemic”, an adjective coined by sociologist of science Karin Knorr-Cetina refers to relations crucial to the inquiry that exceed the boundaries of the scientific community.

funding agency participates to the framing of the research problem by negotiating the methods to be used or the interpretation of measurements (Knorr-Cetina and Malkay 1983, p. 132). Research happens in and ex situ, yet the contribution of the agency remains unacknowledged in the reports and research outputs. To produce “pure” knowledge, the scientists do not refer to the funding agency as a collaborator as it would open the door to a criticism of external influence, and a suspicion of introducing non-scientific criteria. With this concept, Knorr-Cetina emphasises the active role of external agents in the production of science and their relative invisibility. The concept stresses the importance of the circulation of knowledge and its distributedness. Although in an inverse balance of power, the engineers and the annotators together contribute to the production of the knowledge relevant to computer vision. AMT can be characterised as a transepistemic device that enables the collaboration while masking one party’s contribution. It elaborates computer vision as it provides the modelling for the training process at the appropriate scale, and it elaborates it (this time with the emphasis on *laborare*, the latin root of the word) as it articulates computer vision’s division of labour.

#### 4 The data set as photographic alignment

What is the nature of the workers contribution in the transepistemic relation? Workers are said to be *cleaning* the data set, a term that invokes hygiene and manual rather than intellectual labour. To clean is to scrub away the germs and parasites intoxicating a set of samples. It also indicates clearly, where is the annotator’s side in the division of labour. By contrast, the role of the worker would be framed much differently if she was considered as contributing to the curation rather than the cleaning of the data set. The nature of the decisions the annotator has to take is of central importance for what is included in, or excluded from, the data set. Yet, as we begin to understand better, the labour so crucial to the process remains undervalued financially and unacknowledged (the workers are anonymised and interchangeable). Furthermore, the decision-making power of the workers, as we will see, must be simultaneously enabled and delimited.

With the annotator’s work, we are at the core of the process of regularisation of computer vision’s objects. Regularisation means to process the objects in a manner that makes them suitable for use in scientific work. In this case, it means to fashion the objects in a manner that they exhibit the regularities that will be picked up by the algorithms at a later stage. As the historians of science Daston and Gallison (2017, pp. 19–22) remark, “No science can do without such standardized working objects, for unrefined natural objects are too quirkily particular to cooperate in generalizations and comparisons.” Regularities are never simply given in

the data, they always need a helping hand. The workers are given a set of photographs pulled from the Internet through automatic queries of visual search engines. The results from these queries are noisy, ambiguous, and at times obviously unrelated.<sup>6</sup> A considerable amount of work of disambiguation has to be performed. The workers are tasked to filter the search results and extract the relevant photographs. They are expected to find by consensus which photographs are “imaging” a given concept. Imageability, a concept imported from psycholinguistics, describes the ease or difficulty with which a concept can be characterised visually (Paivio 1986; Yang et al. 2019). ImageNet is composed of a wide array of heterogeneous entities including stars, cities, animals, vehicles, micro-particles and diseases. But as different as they are, computer scientists think they may all be *imaged* in a way that make them available for comparison and differentiation. Imprints of light captured by a digital camera, computer simulations, screenshots or photoshoped images are all reduced to pixels. Therefore, molecules, pencils, acrobats, coliphages, olives and tetraspores belong to a same *imaged* register, where regularities and differences can be observed. What do they have in common? They can be photographed, or more precisely represented according to the codes of photographic realism. Significantly, the interface warns the ImageNet’s annotators: “PHOTOS ONLY, NO PAINTINGS, DRAWINGS, etc.” A pivotal role is given to photography conceived as a leveller, an instrument that automatically converts light into pixels according to predictable rules and at the same time that *images* a concept. Photography is mobilised as an instrument to homogenise the visual world, to transform the visual into data, where data of different origins can be compared and classified.

In the first part of this text I have discussed how the notion of scale was central to Li’s project. I have differentiated scale from a simple quantitative indicator. Resolving the scale mobilises micro-labour, eye saccades, and billions of data items. Now we see that the billions of elements selected to enter the data set are not random visual files, they are photographs, and as photographs, they are treated as suppliers of representations, dedicated to the task of “imaging” concepts. Furthermore, photography is given a role of leveller which offers a plane over which individual representations can be compared. In *Le Musée Imaginaire*, Malraux forges a word that encapsulates photography’s role in regulating the representations entering the collection. Everything in the museum has a place if it is “photographable” (Malraux 1965, p. 123). By photographable, Malraux doesn’t mean only something that

<sup>6</sup> Referring to the Torralba et al. (2008), Yang et al. evaluate to 10% the number of photos matching a query, <https://people.csail.mit.edu/torralba/publications/80millionImages.pdf>.

can be reproduced as a photograph. In Malraux's account, photography is active and is perfectly suited to address the question of style. For him, a photograph doesn't transparently describe an artwork. Light, frame, composition, angle, all contribute to reveal style within the visual object. Yet even if these correspond to carefully selected choices, they are revealing something that exists within the object. The photograph makes the objects speak. This revealing, even if it is obtained through a technical procedure, is by no means purely mechanical. In the book, Malraux compares different photographic reproductions of a same artwork and insists some capture the artwork better than others. Therefore, Malraux's photograph is not just a passive reproduction. However, as a revealing, its active contribution is intimately bound to its effacement, its disappearance. Any surplus to what it reveals can be dismissed as irrelevant, and when two photographs are juxtaposed (or "confronted" in Malraux's terms), the dialogue we are invited to witness happens between two objects, not between two photos. To be photographable means to be opened up to comparison by the means of photography.

To bring the discussion back to computer vision, the problem faced by the data set creators can be formulated this way: how to make 20,000 categories imageable and photographable? Diversity is an important criteria to constitute a data set. The diversity in question concerns the objects represented in the photographs. The objects must exhibit variance in positions, viewpoints, appearance. They must be placed in front of different backgrounds and must appear with various degrees of occlusion (Deng et al. 2009, pp. 4–5). The search engine is the provider of such diversity as it concentrates in one page, photographs of distinct sources. Concretely, it means the engineers delegate to the search engine the task to extract the photos from the context in which they operate online. To perform this extraction, the search engine needs to undo a large series of relations that held these photos in place. For example, as most ImageNet's photographs are originally posted on Flickr, it requires the undoing of the relations the platform has established among the many entities that partake in the photograph. On the platform, the photograph is a composite. It is made of a series of JPEG files of different formats. It includes tags and comments. It circulates in communities and albums. Metadata information are attached to it, and so on. All these relations form an "alignment" that makes the Flickr photograph an entity that can be shared, liked, viewed rated. In the Flickr ecosystem, the photograph has currency. When the photograph gains popularity, it opens the door to new groups, the photographer receives a badge, and his visibility increases. The term alignment stresses the fact that a photograph doesn't exist alone. It is enacted through a

series of relations taking place in an apparatus (in this case, the Flickr platform) through which it gains properties, affordances, and currency.

To be included in the data set, a digital file is excised from its Flickr alignment. The comments do not travel with the file, and neither do the albums, metadata information, or the author's name. The Flickr photograph, once tagged by the author or the community, is now categorised according to the WordNet thesaurus. This operation is far from innocuous. The photograph is enacted differently. It is re-aligned, and this re-alignment requires a considerable amount of work. Once selected by the search engine, the files are acquired by the data set makers and they are shown in grids of hundreds of thumbnails to the Turkers. Moving from the original context, where they were published, they enter into another framework of attention. With the AMT interface, a photograph once seen in an individual page or inside a blog article is displayed among many other candidates. The screen is filled with thumbnails. Proximities change radically, the photograph sits next to new neighbours. As already mentioned, AMT privileges the glance rather than a sustained observation as the exploitative labour conditions imply a rapid pace. Decisions must be taken fast. Moving from Flickr to AMT is not just a change in location and relations, it is a change of rhythm, speed, a change of metrics. It is also the conversion of an economy of sight. The Flickr photograph, a product of free labour, becomes the AMT thumbnail, an object of minimal attention for the Turker struggling to make ends meet. The translation from free labour to micro-payment goes along with the translation from an environment that celebrates the full screen view to a device that optimises its workflow with the thumbnail. In AMT, considerations of aesthetics do not apply, legibility becomes more important. Imageability is the dominant criteria of selection and the worker's competence is her ability to capture the photograph's content fast.

## 5 Decision-making, micro-labour in the photographic alignment

A series of examples will help understand the imaging process happening through the re-alignment of the search results. I will begin with the synset "Ratatouille". The description given to the workers for the ratatouille concept is "a vegetable stew; usually made with tomatoes, eggplant, zucchini, peppers, onion, and seasonings". It is filed under Misc → food, nutrient → nutriment, nourishment → dish → stew. The photos selected in the synset are for one half extracted from Flickr accounts and the other half from culinary websites, foodie blogs, cooking tutorials, or restaurant pages. The distribution of photographs in the synset seems to correspond to the stated goals of the

**Illustration 1** synset  
n07592768, Ratatouille

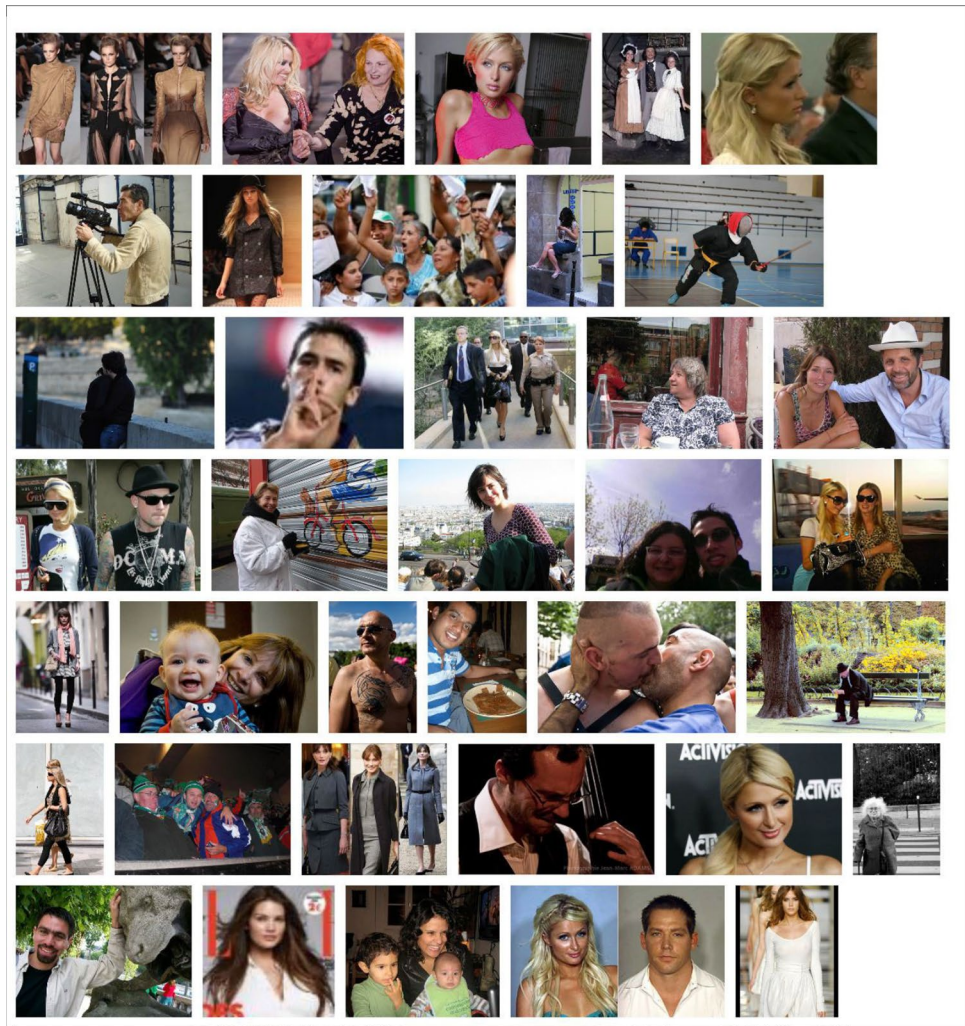


data set makers: variations in shapes, lighting, and points of view. The selection is testament to the variety of forms the dish can take. Layered slices of vegetables in a baking tray, cubes of tomatoes and eggplants in a pot, the meal is displayed alone or in combination with other dishes, on a stove, next to a sink, on a garden table at a barbecue party, in a living room on an immaculate white sofa, in a tupperware container on a flower tablecloth. The technical competences of the photographers vary. The framing and cropping of the photographs, the sharpness, the blurs, the light balance relate to different intentions and belong to contrasting aesthetics. A photograph capturing the delicate reflections of light in the marinade sits next to a quick snapshot in which the details of the meal vanish under reflection of the flashlight. In many ways, the variety of the selections meet the engineering requirements. Yet a closer look at the selected items reveals the presence of entities that do not match the synset's definition. The dish has been confused with various forms of stew (e.g., meat stew) or tomato and mozzarella salads. More surprisingly, pasta meals and grilled shrimps

on skewers have also been confused with ratatouille. The Pixar character of the eponymous movie also appears to be a candidate for the combination of zucchini, tomatoes and eggplants. These items can be seen as the results of mistakes by individual workers who didn't pay enough attention, and that can be partially explained by the nature of the dish inasmuch as it takes multiple forms (e.g., a "rata" in French military slang originally being a rough mix of cheap vegetables) (Illustration 1).

The data set makers anticipated human errors of judgement and the difficulty to decide whether the candidate images contain a particular object. They designed a solution to evaluate inter-annotator agreement. Essentially, a candidate image is included in the data set only if a consensus is reached among annotators. As I have explained, the platform relies on parallelism for technical and organisational reasons. The workers are never aware of the decisions taken by others and cannot interact with them. Consensus on the platform means that several workers are presented the same candidate images and the platform compares their selection.

**Illustration 2** synset  
n09708750, Parisian



Where the selections overlap a consensus is reached. Furthermore, this consensus is modelled according to a scale that varies according to the “semantic difficulty” (Deng et al. 2009) of the data set (its grade of imageability). The ImageNet authors give the example of the difficulty to reach a consensus for a synset like “Burmese cat” in comparison with “cat”. As cat is deemed more imageable than Burmese cat, the number of annotators who need to agree on the label “Burmese cat” for the same photograph needs to be higher than the number of agreeing annotators for “cat”. “Ratatouille” at the bottom of the branch Misc → Food Nutrient, is one of those terms deemed more semantically difficult and requiring, therefore, more eyeballs and a higher level of consensus.

Yet a higher level of consensus doesn’t guarantee a greater conformity to the definitions given to the workers. In the synset n09708750, the concept “Parisian”, is described as “a native or resident of Paris” (Illustration 2). A remarkably high portion of the selected images are photos of Paris Hilton in different forms including candid and

tabloid photos, 3D models, or selfies. Representations of the socialite in bathrobe, bikini, gown or casual wear dominate by far other familiar cultural tropes of Parisians like tourists at the Louvre Pyramid, passers-by wearing berets and striped sailor shirts, men kissing at Gay Pride, people enjoying a coffee at a terrace, or visiting the Eiffel tower. Unlike the previous example, the imbalanced proportion of photographs of Hilton cannot be explained by the errors of a minority of annotators that have introduced a negligible percentage of erroneous candidates.

The cause is often attributed to the workers. Engineers develop techniques to track the workers that are considered as “satisficers” (Hata et al. 2017) or “spammers” (Quach 2019) who accomplish their tasks with too little care. Other causes are evoked, such as the annotators are not culturally compatible (Hata et al. 2017) with the request or they don’t read the definition. For these reasons, many precautions are in place to ensure the definition is read by the annotators. Before seeing the candidate images,

the annotator is presented the definition of the synset. Subsequently she is presented another screen, where she has to choose the right description among several others. For instance for the synset N04070003, “reformer”, the annotators are given the following “gloss”<sup>7</sup>: “An apparatus that reforms the molecular structure of hydrocarbons to produce richer fuel; a catalytic reformer” (imagenet.org no date). However, none of the photos selected in the synset depicts such an apparatus. Instead they depict Pilates reformers and people doing exercises, even if to access the hit, the annotator had to confirm explicitly that reformer was a chemistry-related apparatus and not a gymclass prop. For the annotator, glancing is not only a mode of perception related to the rapid scanning of thumbnails, text too is read in a glimpse (Illustration 3).

This brings us back to the problem of decision-making in the annotation environment. The decision is delegated and regulated through consensus. Supervision happens *post-hoc* when consensus cannot be achieved or through punctual quality checks. Supervision here differs starkly from an idea of a subject dominating a scene from a bird’s eye perspective. To contrast, it may be useful to remind ourselves of a famous photograph representing Malraux in his spacious office at the French Ministry of Culture. The floor is covered with photographs. Malraux standing in front of his desk holds one of the photographs for further consideration. Malraux literally supervises, visually dominates the pictures distributed at his feet. This mode of supervision does not correspond to any of the actors involved in ImageNet. Li and her colleagues control the process through the reports given by the AMT interface. They receive numerical indicators and they can insert them in spreadsheets. But they never supervise ImageNet’s visual content from an overhanging position. There is no floor large enough to contain 14 million photographs, and even if such floor could exist, there would be no position from which an observer could embrace its totality. If such a position is not available to the researchers, the situation is even less comparable for the annotators. If the annotators are given the role to look at the photos on their screens, they are immersed in a flow and unable to negligently pick one photograph and consider it with an air of detached interest. This doesn’t mean, however, that no decisions are taken and that no supervision is in place. What this comparison suggests is that we need to further enquire into the mode of decision-making and that we have to renew our notion of supervision to adapt to what is happening in the annotation environment.

To annotate at speed does not consist of a mechanical response issued from a passive subject. To understand the

annotator’s contribution, it is fundamental to understand the process as one of elaboration which goes beyond rational choice and explicit judgement. The epistemic contribution consists in embodying a scale, figuring out rhythms and levels, understanding and refraining involvement. To attend to the process of elaboration means to avoid concentrating exclusively on the semantic decision. The elaboration is not limited to a pivotal moment, where the annotators assert the meaning of a photograph. It includes the complex methods through which they synchronise within an alignment and embody a scale. Synchronisation, scale embodiment, attunement, and seeing at speed are at the core of the photographic mediation of computer vision. They intervene crucially in the resolution of the photograph in a given alignment. To understand the annotator’s contribution is, therefore, to be attentive to how she creatively relates with the apparatus, how she probes, where the apparatus begins and ends. To attend to the resolution of a candidate image into a data set item requires a rethinking of the nature of the decisions that are taken. It requires to shift from an understanding of a decision mechanism based on a pivotal moment that involves an “isolated who” (Mol 2002) to a more distributed consensus of actors and devices. The question becomes: how do the Turkers achieve consensus without explicit coordination? A crucial part of the answer lies in the methods the annotators mobilise to figure out the rhythm they have to follow.

The cadence of the platform comes from the remuneration. To secure a minimal income, the Turkers have to perform at high speed. They have to calibrate their involvement. They need to figure out how precise they must be to make enough annotations and have their hits accepted. This balancing act requires finding one’s place in the alignment. The Turkers are contributing to the resolution of a scale, to the correlation between semantic hierarchies and speed. They do not debate together to decide whether a candidate image should be considered “tomato mozzarella” and consequently be excluded from the “ratatouille” set. They have to judge whether it is worth slowing down to give attention to this candidate or if, at first glance, it can be assimilated to the concept ratatouille without threatening their remuneration. They have to intuit if the difference they notice is worth changing pace or if they can remain indifferent to this difference. The Turkers develop a common sense of the apparatus’ resolution without having to explicitly negotiate. The architecture and the pace of the AMT define its resolution, its grade of precision in the descriptions, and concomitantly, it creates a zone of indiscernibility for the apparatus.

A consensus is not always a matter of explicit argumentative deliberation. It can be a matter of levelling, of finding a common grade of involvement. To ask which arguments led to the selection of some items rather than others is less important than asking how is a consensual echo propagated through the AMT. To build a consensus, the

<sup>7</sup> A technical term in WordNet’s parlance for the description of a synset.

workers develop a sense of the speed required by the apparatus. This speed as I have said is induced by the remuneration, but it is also induced by many more elements. For instance, AMT stabilises the candidate images as thumbnails detached from their original context. They are search engine results and as such the search engine already conveys a sense of consensus. If many candidate images look similar, it suggests that a consensus exists over them. The Turkers do not validate mechanically the dominant representations they are given by the interface, but the regularities that spring out from the grid of thumbnails function as a cue. It is an accelerator. To choose against the coherence that emerges from the results requires more work and time. It requires looking at the candidates that do not stand out. Another important cue comes from the jobs being accepted or refused by the requester. The reasons why a job is accepted (understand, remunerated) or refused (the worker has no recourse against the requester) are extremely rarely given by the requester. It is, therefore, not always possible to correlate a decision made by the worker and a rejection from the requester. But as they have significant economic consequences for the annotator, she evaluates her work based on her interpretation of the requester's decision. Therefore, the Turkers never rely on any explicit guidance, but they follow an echo, where different forms of feedback and cues resonate with each other. Workers need to learn how to listen to the apparatus more than they have to read the labels and the definitions of the synsets.

The Parisian or the reformer cases make clear that the composition of the synsets cannot be explained by a process in which workers carefully read the gloss and select the candidate images accordingly. Focusing on a pivotal moment of an “isolated who” making a decision takes us only so far. The Turkers hover over a visual configuration emanating from the interface. Rushing through the pages, they see an overwhelming presence of the shining blondness of a familiar icon whose name matches the synset's label. The Turkers do not necessarily believe that Paris Hilton is a resident of, or born in, Paris. Their decision is that there are enough responding echoes in the apparatus to validate the selection without having to spend time further verifying. The object of the consensus is not the fact that Hilton is a Parisian, but that such an approximation will not isolate them. A course of recognition traverses the apparatus. To recognise is to be recognised. To be recognised as a Turker, one doesn't need to recognise things that are factually right but to recognise the grade of approximation that is expected. The Turker's competence is multi-scalar, not just a semantic affair. As a Turker wrote on a forum, it is to have a sense of what is “enough to get away with”. To get away is indeed the right term as a Turker must always have an eye on the previous hit and another on the next. Speed is conducive to consensus.

## 6 Arbitrating photographic alignments

Following the production of consensus in the annotation environment, we gained insights into the nature of the workers' transepistemic contribution. They are accomplishing speed and accuracy rather than emitting explicit judgements. They are constantly engaged in probing the plasticity of the consensus rather than producing discrete statements about facts. Having established that, I need to relate the workers' contribution to the alignment they are part of. I have already stated, following Crimp that the condition for the diversity of the data set is that the objects are represented in various conditions but that the medium which represents them is uniform and transparent. Yet as we have seen the workers are involved in more than classifying concepts represented differently through a seamless medium. The heterogeneity of the candidate images doesn't only reflect a disparity of styles and contexts. It also reflects the heterogeneity of larger apparatuses through which photographs are stabilised and set in motion. To understand the full consequence of this, we need to come back to the notion of alignment and the role played by the search engine in the data set's acquisition pipeline. The search engine hides the relations and context in which a photograph lives online. It excises the photograph from its alignment and, therefore, contributes to the appearance of the photographic object being independent from the apparatus through which it is made visible. This extraction is also in part the validation of the authority of an apparatus over a query term. In many synsets there are large amounts of photographs from the same source (e.g., a Flickr album, a blog or website). The distribution of the photographs in a synset reflect the effectiveness of the strategies of the various platforms competing for the visibility of their contents. A platform like Flickr very much anticipates the requirements of a search engine and offers it as a structured set of objects (file, metadata) optimised for machine readability. By securing and reinforcing the internal connection of its content (Flickr photos are related to each other by various mechanisms like links, inclusions in groups or tags), the platform also anticipates the search engine's ranking criteria (i.e., the more links the higher visibility). The platform is “search engine optimised” and the users of the platforms are recruited in this effort. The search engine and the source platform are not clearly delineated entities between which independent JPEG files are in transit. Rather the platform very much interiorises the search engine in many ways. This explains that various categories are “taken over” by content popular on a sharing platform as the platform is optimised to do so.

In this competition, platforms have an advantage as they are built from the start to capitalise on the circulation of

**Illustration 3** synset  
n04070003, Reformer



their assets. But this advantage is only partial. For certain categories, other apparatuses beat the platforms at their game. The monopoly then comes from other sources. The synset n09633969 “wrongdoer, offender”, for example, is essentially constituted from mugshots published by US states or county websites. These administrations operate their own apparatuses of capture, identification and dissemination that dramatically contrasts with those of amateur photography. These apparatuses inscribe their own regularities in terms of pose, colours and point of views. They also inscribe other regularities in terms of race and gender by excluding images of white collar (or female) criminals and focusing on people of colour. As a consequence, a synset like wrongdoer, constituted of mugshots does not merely validate photos that are imaging a concept, it perpetuates the politics of shaming inherent to state department’s websites.

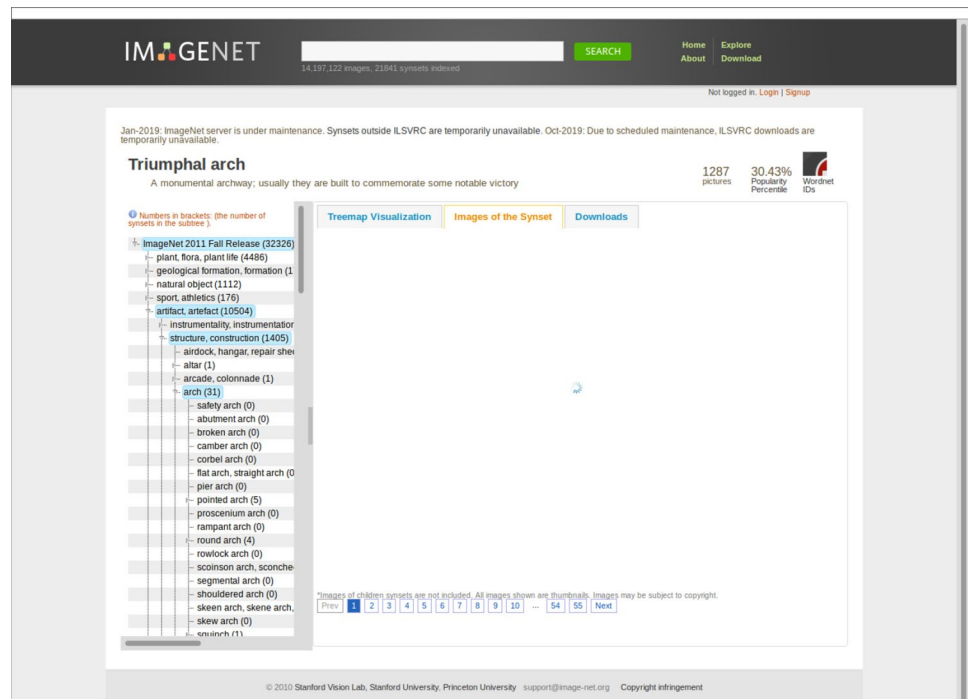
The data set is not a merely a collection of photographs, it is an arbitration between different forms of alignment: Flickr’s amateur snaps, police portraits, submarine

photographs are re-aligned as items of a training set. It doesn’t make sense to talk about this process of selection as if a group of distinct individuals were choosing discrete photographs in a void, but as the data set feeding on different photographic alignments, branching itself into various apparatuses and offering them different shares of the data set’s space, and delegating various levels of decisions to the search engine algorithm and the AMT workers.

Having followed this process of realignment, it is clear that using photography as a tool to homogenise the visual world leads to a paradox. The paradox is that to resolve a photograph as data, all the heterogeneity that pertains to the medium, its apparatus and its circulation needs to be repressed. The photograph needs to be *made* a transparent vehicle. And to make it transparent, computer scientists need to engage in the production of their own apparatuses and produce their own alignments. This long chain of translations and ruptures from search queries to AMT to data set is in itself an alignment and also conditioned to the existence and performance of apparatuses. The paradox, therefore,



**Illustration 4** A screen capture of the Triumphal Arc synset on [imagenet.org](http://imagenet.org)



lies in this: it is an alignment that needs to negate itself, that needs to remove itself to construct photography as a transparent leveller of data. The workers are, therefore, engaged in a double process of elision. Their own contribution is unacknowledged, as they are considered to be merely mechanically responding to hits. Their contribution consists of erasing the traces of the apparatuses involved in the resolution of online photographs into data set items. The *invisibilisation* of the workers goes hand in hand with the resolution of photography as a transparent leveller of data.

## 7 Conclusion

At this point, I have made the case that engaging with the photographic elaboration of computer vision requires the acknowledgement of the importance of the workers epistemic contribution and the photographic alignment traversing CV. To conclude this paper, I will show how this notion of photographic elaboration can help question the terms by which current controversies in CV are being framed and the responses that have ensued (Illustration 4).

Despite its immense academic success, its inclusion in many pieces of software, and its celebration as a benchmark for AI algorithms, ImageNet, today, is undergoing a grave crisis. For those who consulted its website during the last years, ImageNet’s online presence has been gradually disappearing. Its disappearance is due in larger part to mounting controversies both from the public (for instance, see *ImageNet Roulette*, Crawford and Paglen 2019) and from inside

the tech community (i.e., Dulhanty and Wong 2019; Shankar et al. 2017; Recht et al 2019). The criticism mainly targets the data set’s cultural, racial and gender bias. In response, the data set creators and a larger team of researchers are engaged in a process of revision of the data set as an attempt to remedy to the various shortcomings expressed in and out the CV community.

Current reparative efforts are concentrated on the mainstreaming of the training data, the objective being to make ImageNet fairer and updating its “acquisition pipeline”. The researchers’ response to the problem, as expected, essentially concentrates on the question of representation. Problematic categories are annotated to attach properties such as gender, age or skin colour to images. To give fairer representations means, for instance, rebalancing the roles portrayed in the photographs according to gender or racial criteria, and to produce a more suitable distribution of these roles in a synset. Leaving aside the questions of how criteria such as gender or skin colour are assessed, or what a suitable distribution of photographs based on such criteria might mean, it is clear from the outset that the response once again concentrates on the depicted objects to evaluate how well they image a concept. As a response to external pressure, the representationalism of the data set makers increases. The “People” subtree is annotated to evaluate the imageability of its concepts. All the concepts with a low imageability grade are removed. The solutions designed by the research team keep on reducing the number of selected elements in the data set, and potentially problematic categories are discarded. Data demining is in order. At the end of the process, the

researchers estimate that out of the 2832 subcategories of the “people” subtree, only 139 will remain.

But as we have seen, the problem runs deeper than representation. A photograph is not just a constructed representation, it is also very much defined by its circulation, the currencies in which it is accepted, how it comes into series. It acquires meaning and function through use. It requires a reflection on the conditions in which a photograph is temporarily resolved, to the different elements it aligns with. It requires a renewed attention to the relation with the apparatuses and the engineer’s own apparatus of capture, his/her own alignments. Therefore, a response that addresses a criticism of bias only narrowly defines the problem. If the problem is understood solely as a question of right or wrong data (more balanced and fairer), it will leave unquestioned how photography is made to be data in the first place. It remains blind to the performative role of the methods, the labour, the agents, the scales, the apparatuses and the alignments it relies upon.

If we acknowledge that photographs are not representations independent from their alignments, what is the consequence for data set makers? If the process of production of the data set doesn’t take into account the whole chain of aligned elements it rests upon, the nature of its own apparatus, it will keep treating visual data as a mere collection of elements that can be replaced by safer ones or entirely avoid risky categories. But this improvement will then necessarily lead to further impoverishment of the data. As an apparatus in denial of its own performativity will only neutralise further the relational nature of the photograph. Engineers apply the GIGO maxim, “garbage in, garbage out”, and imagine getting rid of the garbage is the path forward. Accordingly, a criticism based on bias will lead to an increased cleaning of the data set. But image *is* garbage. It is entangled in conflicting regimes of representations, divergent apparatuses, disruptive alignments and irreconcilable practices. If those are cleaned up, there is no image left. There will be perhaps less risk for algorithms to be contaminated by obscene or offensive imagery if the data set keeps on shrinking. But this entails a bigger risk, the risk for machine learning to not learn anything worth knowing. A data set of perfect representations is an empty one. The most striking images ever produced by computer vision are not deep, either fakes or dreams, they are flat and white and on their centre, an icon suggests the page is loading without ever displaying any content. They are the millions of empty pages of the ImageNet website. These images of the data set’s ruins are the products of its inability to engage with the labour it relies on and the apparatus it is entangled with.

To introduce this text, I referred to Crimp’s take on the *musée imaginaire*. I stated that the ruins Crimp refers to were not solely the ruins of extracted elements dying in the confines of a mausoleum. They were the ruins of a project

that relied on a particular change of scale. Malraux’s project, Crimp stated, was an hyperbole representing the transformation of the museum. The “real” museum is seen as limited in what it can acquire and keep within its walls. Photography makes it possible to enlarge the scope of what one sees and how one sees. The museum as imagined by Malraux aims to produce knowledge by comparison. Works can be shown side by side, monuments can be scaled down and details can be enlarged. Before, one would compare a painting with a memory, now one can compare two objects through their photographs. The problem Crimp addresses is that photography is enrolled in the process of articulating the relations between a potentially unlimited amount of works on the condition that its role should be limited to one of revealing. Therefore, the process of comparison of otherwise heterogeneous elements is only meaningful, because the process of extraction and homogenisation is made invisible. But as Crimp ironically remarks, when photography enters the museum as an art object, heterogeneity returns as: “Even photography cannot hypostatize style from a photograph” (Crimp 1980, p. 53).

Crimp’s lesson is that we must extend our concern for the objects to the larger project, to the investment in a certain scale obtained through photography and the putative knowledge it purports to generate.

Instead of lamenting the presence of dead objects in a mausoleum, one should question the ruins of an ideal plane of comparison. To translate this in the context of CV, instead of taking off every element that could lead to controversy, the data set makers should instead reckon with their own apparatus of annotation and the labour it conceals. Instead of denying the collaboration, they should address the trans-systemic dimension of the work carried out by the annotators and engage with it. This is hard work, because it questions the discipline, where it hurts most: at the level of its economy. Recognising the work carried out by the Turkers obviously opens the question of the financial revalorisation of their labour. Furthermore, computer scientists should consider afresh their understanding of photography and do away with the entrenched representationalism of the discipline. This is hard word again, because it rejects the notion of an ideal plane on which all images can be made data. Combined together these two observations make clear that the work to be done is much more ambitious than bias-proofing sensitive categories. It confirms Li’s insight: data sets are indeed tools to resolve the scale of computer vision’s problem. However, it suggests that resolving the scale is a project that needs a fresh start.

**Acknowledgements** This article is based on a research conducted thanks to the support of London South Bank University and The Photographers’ Gallery. It has benefited from the attentive comments of Gaia Tedone, Katrina Sluis, and Ruben Van de Ven. It draws on a reflection shared with Laurence Rassel, Geoff Cox, Andrew Dewdney

and my colleagues at the Centre for the Study of the Networked Image, as well as the Institute of Computational Vandalism and the Constant collective.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Beheshti S-M-R, Tabebordbar A, Benatallah B, Nouri R (2016) Data curation APIs. *CoRR abs/1612.03277*
- Brownlee J (2019) A gentle introduction to computer vision. <https://machinelearningmastery.com/what-is-computer-vision/>. Accessed 25 Mar 2020
- Crawford K, Paglen T (2019) Excavating AI: The Politics of Images in Machine Learning Training Sets. <https://www.excavating.ai/>. Accessed 19 Mar 2020
- Crimp D (1980) On the museum's ruins. *October* 13:41–57. <https://doi.org/10.2307/3397701>
- Daston L, Galison P (2007) *Objectivity*. Zone Books, New York
- Deng J, Berg A, Li K, Li F-F (2010) What does classifying more than 10,000 image categories tell us? In: *ECCV*. pp 71–84
- Deng J, Dong W, Socher R, et al (2009) Imagenet: A large-scale hierarchical image database. In: *CVPR*
- Dulhanty C, Wong A (2019) Auditing ImageNet: Towards a Model-driven Framework for Annotating Demographic Attributes of Large-Scale Image Datasets. *CoRR abs/1905.01347*
- Fei-Fei L (2010) ImageNet, crowdsourcing, benchmarking & other cool things. In: *CMU VASC Seminar*
- Fuller M (2005) *Media Ecologies: Materialist Energies in Art and Technoculture*. The MIT Press
- Fullscale (2019) Machine learning in Computer Vision. <https://fullscale.io/machine-learning-computer-vision/>. Accessed 25 Mar 2020
- GoogleTechTalks (2011) Large-scale Image Classification: ImageNet and ObjectBank. <https://www.youtube.com/watch?v=qdDHP29QVdw>. Accessed 28 Feb 2019
- Hara K, Adams A, Milland K, et al (2018) A data-driven analysis of workers' earnings on amazon mechanical turk. In: *proceedings of the 2018 CHI conference on human factors in computing systems*. ACM, New York, NY, USA, pp 449:1–449:14
- Hata K, Krishna R, Fei-Fei L, Bernstein MS (2017) A glimpse far into the future: understanding long-term crowd worker quality. In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. ACM, New York, NY, USA, pp 889–901
- Irani LC (2015) The cultural work of microwork. *New Media Soc* 17:720–739. <https://doi.org/10.1177/1461444813511926>
- Irani LC, Silberman MS (2013) Turkopticon: interrupting worker invisibility in amazon mechanical turk. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, New York, NY, USA, pp 611–620
- Karpathy A (2014) What I learned from competing against a ConvNet on ImageNet. <https://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/>. Accessed 10 Jul 2019
- Knorr-Cetina KD, Malkay M (1983) *Science observed: perspectives on the social study of science*. Sage, London
- Malraux A (1965) *Le muséeimaginaire*. Gallimard, Paris
- Marr D (1982) *Vision. A computational investigation into the human representation and processing of visual information*. MIT Press, Cambridge, MA
- Mol A (2002) *The body multiple: ontology in medical practice*. Duke University Press, Duke
- Paivio A (1986) *Mental representations: a dual coding approach*. Oxford University Press, New York
- Papadopoulos DP, Uijlings JRR, Keller F, Ferrari V (2016) We don't need no bounding-boxes: training object class detectors using only human verification. *CoRR abs/1602.0*
- Quach K (2019) Inside the 1TB ImageNet data set used to train the world's AI: Naked kids, drunken frat parties, porno stars, and more. [https://www.theregister.co.uk/2019/10/23/ai\\_dataset\\_image\\_net\\_consent/](https://www.theregister.co.uk/2019/10/23/ai_dataset_image_net_consent/). Accessed 25 Mar 2020
- Recht B, Roelofs R, Schmidt L, Shankar V (2019) Do ImageNet Classifiers Generalize to ImageNet? <https://arxiv.org/pdf/1902.10811.pdf>. Accessed 10 Jul 2019
- Rosenberg C (2013) Improving Photo Search: A Step Across the Semantic Gap. <https://ai.googleblog.com/2013/06/improving-photo-search-step-across.html>. Accessed 27 Feb 2019
- Shankar S, Halpern Y, Breck E, et al (2017) No classification without representation: assessing geodiversity issues in open data sets for the developing world
- Simpson Center for the Humanities UW (2017) Lorraine Daston on Algorithms Before Computers. <https://www.youtube.com/watch?v=pqoSMWnWTwA>. Accessed 25 Mar 2020
- Smith C (2019) 20 Interesting flickr facts and stats 2019 by the Numbers. <https://expandedramblings.com/index.php/flickr-stats/>. Accessed 10 Jul 2019
- Torralba A, Fergus R, Freeman WT (2008) 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Trans Pattern Anal Mach Intell* 30(11):1958–1970. <https://doi.org/10.1109/TPAMI.2008.128>
- Vijayanarasimhan S, Grauman K (2009) What's it going to cost you? Predicting effort vs. informativeness for multi-label image annotations. In: *2009 IEEE conference on computer vision and pattern recognition*. pp 2262–2269
- Wikipedia Computer Vision. [https://en.wikipedia.org/wiki/Computer\\_vision#Related\\_fields](https://en.wikipedia.org/wiki/Computer_vision#Related_fields). Accessed 25 Mar 2020
- Yang K, Qinami K, Fei-Fei L, et al (2019) Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Perceptual bias and technical metapictures: critical machine vision as a humanities challenge

Fabian Offert<sup>1,2</sup> · Peter Bell<sup>2</sup>

Received: 30 July 2019 / Accepted: 18 August 2020 / Published online: 12 October 2020  
© The Author(s) 2020

## Abstract

In many critical investigations of machine vision, the focus lies almost exclusively on dataset bias and on fixing datasets by introducing more and more diverse sets of images. We propose that machine vision systems are inherently biased not only because they rely on biased datasets but also because their *perceptual topology*, their specific way of representing the visual world, gives rise to a new class of bias that we call *perceptual bias*. Concretely, we define perceptual topology as the set of those inductive biases in machine vision systems that determine its capability to represent the visual world. Perceptual bias, then, describes the difference between the assumed “ways of seeing” of a machine vision system, our reasonable expectations regarding its way of representing the visual world, and its actual perceptual topology. We show how perceptual bias affects the interpretability of machine vision systems in particular, by means of a close reading of a visualization technique called “feature visualization”. We conclude that dataset bias and perceptual bias both need to be considered in the critical analysis of machine vision systems and propose to understand critical machine vision as an important transdisciplinary challenge, situated at the interface of computer science and visual studies/*Bildwissenschaft*.

**Keywords** Machine learning · Computer vision · Bias · Interpretability · Perception

## 1 Introduction

The susceptibility of machine learning systems to bias has recently become a prominent field of study in many disciplines, most visibly at the intersection of computer science (Friedler et al. 2019; Barocas et al. 2019) and science and technology studies (Selbst et al. 2019), and also in disciplines such as African-American studies (Benjamin 2019), media studies (Pasquinelli and Joler 2020) and law (Mittelstadt et al. 2016). As part of this development, machine vision has moved into the spotlight of critique as well,<sup>1</sup> particularly where it is used for socially charged applications like facial recognition (Buolamwini and Gebru 2018; Garvie et al. 2016).

In many critical investigations of machine vision, however, the focus lies almost exclusively on dataset bias (Crawford and Paglen 2019), and on fixing datasets by introducing more, or more diverse sets of images (Merler et al. 2019). In the following, we argue that this focus on dataset bias in critical investigations of machine vision paints an incomplete picture, metaphorically and literally. In the worst case, it increases trust in quick technological fixes that fix (almost) nothing, while systemic failures continue to reproduce.<sup>2</sup>

We propose that machine vision systems are inherently biased not only because they rely on biased datasets (which they do) but also because their *perceptual topology*, their specific way of representing the visual world, gives rise to a new class of bias that we call *perceptual bias*.

Concretely, we define perceptual topology as the set of those inductive biases in machine vision systems that determine its capability to represent the visual world. Perceptual bias, then, describes the difference between the assumed

✉ Fabian Offert  
offert@ucsb.edu

Peter Bell  
peter.bell@fau.de

<sup>1</sup> University of California, Santa Barbara, CA, USA

<sup>2</sup> Friedrich Alexander University Erlangen-Nuremberg, Erlangen, Germany

<sup>1</sup> As shown, for instance, by the increasing prominence of the FATE-CV workshop, organized by Timnit Gebru at CVPR, one of the major international computer vision conferences.

<sup>2</sup> The most recent (at the time of writing) example of this process at work can be seen in the controversy around the PULSE paper, see Kurenkov (2020) and Offert (2020).

“ways of seeing” of a machine vision system, our reasonable expectations regarding its way of representing the visual world, and its actual perceptual topology. Research in computer science has shown that the perceptual topologies of many commonly used machine vision systems are surprisingly non-intuitive, and that their perceptual bias is thus surprisingly large.

We show how perceptual bias affects the interpretability of machine vision systems in particular, by means of a close reading of a visualization technique called “feature visualization” (Erhan et al. 2009). Feature visualization can be used to visualize the image objects that specific parts of a machine vision system are “looking for”. While, on the surface, such visualizations do make machine vision systems more interpretable, we show that the more legible a feature visualization image is, the less it actually represents the perceptual topology of a specific machine vision system. While feature visualizations thus indeed mitigate the opacity of machine vision systems, they also conceal, and thus potentially perpetuate, their inherent perceptual bias.

Feature visualizations, we argue, should thus not be understood so much as direct “traces” or “reproductions” of the perceptual topology of machine vision systems (analog to the technical images of photography) but more as indirect “illustrations”, as “visualizations” in the literal sense of forcibly making-visual (and thus making visible and subsequently making interpretable) the non-visual. They should be understood as *technical metapictures* in the sense of W. J. T. Mitchell (Mitchell 1995), as images about (machine) seeing.

We also show how feature visualization can still become a useful tool in the fight against dataset bias, if perceptual bias is taken into account. We describe a case study where we employ feature visualization to discover dataset bias in several ImageNet classes, tracing its effects all the way to Google Image Search.

We conclude that dataset bias and perceptual bias both need to be considered in the critical analysis of machine vision systems and propose to understand *critical machine vision* as an important transdisciplinary challenge, situated at the interface of computer science and visual studies/*Bildwissenschaft*.

## 2 Deep convolutional neural networks

Our investigation looks at machine vision systems based on deep convolutional neural networks (CNNs), one of the most successful machine learning techniques within the larger artificial intelligence revolution we are witnessing today (Krizhevsky et al. 2012). CNNs have significantly changed the state of the art for many computer vision applications: object recognition, object detection, human pose estimation,

and many other computer vision tasks are powered by CNNs today, superseding “traditional” feature engineering processes.

For the purpose of this investigation, we will describe CNNs from a topological perspective rather than a mathematical perspective. In other words, we propose to understand CNNs as spatial structures. While the topological perspective certainly requires the bracketing of some technical details, it also encapsulates the historical development of AI from “computational geometry”, as Pasquinelli (2019b) reminds us.

From the topological perspective, we can describe CNNs as layered systems. In fact, the “deep” in “deep convolutional neural network” is literal rather than metaphorical (Arnold and Tilton 2019): it simply describes the fact that CNNs usually have more than two layers, with networks consisting of two layers sometimes being called “shallow” neural networks. In the simplest version of a (non-convolutional) neural network, these layers consist of neurons, atomic units that take in values from neurons in the previous layer and return some weighted sum of these values. So-called “fully connected layers” are thus not at all different from traditional perceptrons (Rosenblatt 1957; Minsky and Papert 1988), with the one exception that they only encode differentiable activation functions, i.e. the computation of the weighted sum of input values is achieved in a differentiable way, most commonly in the form of a so-called rectified linear unit.

Deep convolutional neural networks, then, introduce new classes of neurons, which perform more complex functions. Convolution operations have been used in signal processing way before neural networks regained popularity. Mathematically speaking, a convolution operation is an operation on two functions that produces a third function which measures the influence of the second function on the shape of the first. A more intuitive geometric definition is that of a kernel, a matrix, “scanning over” a second matrix to produce a third (Dumoulin and Visin 2016). A common example is a Gaussian kernel which can be used to blur an image, or a Sobel kernel that detects edges in images. The weights of such a kernel can become learnable parameters of a convolutional neural network. This means that the network, during training, will learn which kind of kernels, and thus essentially which kind of *image filters* are useful for a classification task.

Generally, as with all neural networks, learning in CNNs takes place in three different stages. A labeled input, an image, for instance, is passed through the interconnected layers of the network, until it reaches an output layer where a prediction regarding the input image is made, depending on the task set for the system. Such a task could be to classify an image according to certain categories, find the boundaries of an object in an image, or other problems from computer vision. An evaluation function (called “loss function” in machine learning) then measures how far off the prediction

of the system is. This information “flows back”<sup>3</sup> through the network, and all its internal connections are adjusted accordingly.

All of these steps take place for all images in the training set, and periodically, the system is also tested on previously unseen samples. This validation process is particularly important as it avoids a failure mode called “overfitting”, where a network learns indeed to perfectly predict the training set labels, but does not generalize to unseen data, which is the whole purpose of training.

It is because of this incremental process that often spans thousands of iterations, that CNNs are notoriously opaque. Common CNN architectures can have millions of neurons and even more interconnections between these neurons. It is thus close to impossible to infer from looking at the source code, data, weights, or any other aspect of a CNN, either alone or in conjunction, what it does, or what it has learned. Selbst and Barocas (2018) have suggested calling this opacity *inscrutability*.

Complexity, however, is not the only reason for the notorious opacity of CNNs. As Selbst and Barocas argue, CNNs are also *non-intuitive*. The internal “reasoning” of neural networks does not necessarily correspond to intuitive methods of inference, as *hidden correlations* often play an essential role. In other words, it is not only hard to infer the rationale behind a network’s decision from “looking at it” because it is inscrutable, this rationale might also be significantly non-intuitive. Selbst and Barocas have argued that non-intuitiveness could be described as an “inability to weave a sensible story to account for the statistical relationships in the model. Although the statistical relationship that serves as the basis for decision-making might be readily identifiable, that relationship may defy intuitive expectations about the relevance of certain criteria to the decision” (Selbst and Barocas 2018, 1097).

### 3 Interpretable machine learning

This problem has been widely recognized in the technical disciplines as the problem of building *interpretable machine learning* systems, also referred to as *explainable artificial*

*intelligence* systems.<sup>4</sup> Such systems, either by design or with the help of external tools, provide human-understandable explanations for their decisions, self-mitigating both their inscrutability and non-intuitiveness.

In the past 3 to 5 years, research in interpretable machine learning has matured into a proper subfield of computer science (Lipton 2016; Doshi-Velez and Kim 2017; Gilpin et al. 2018; Mittelstadt et al. 2019) and a plethora of statistical tricks (Lundberg and Lee 2017; Ribeiro et al. 2016) has been developed to ensure the interpretability of simpler models like linear regression, particularly in safety-critical or socially charged areas of machine learning like credit rating or recidivism prediction.

Beyond these technical results, however, a larger conceptual discussion has emerged in the technical disciplines as well that “infringes” on the terrain of the humanities. It is centered around attempts to find quantitative definitions for concepts that naturally emerge from the problem at hand, such as “interpretation” and “representation”, with the help of methods and concepts from disciplines as diverse as psychology, philosophy, and sociology, building a “rigorous science of interpretable machine learning”, as Doshi-Velez and Kim (2017) write.

A concrete example would be that of a “cat” and “dog” classifier. Hypothetically, we can train a standard CNN architecture on a large dataset of cat images and dog images. These images will be processed by the network as described above: the input layer of the network will transform the pixel values of an image into a multidimensional vector, which then flows “forward” through the network, until a prediction is made in the very last layer. This prediction is evaluated by the loss function—for instance by the mean squared error of all neurons—and back-propagated through the network. Weights are changed incrementally, until the performance of the network on a validation set, for instance, 10% of the dataset of cat and dog images that have been held back, stops increasing. Given a large enough dataset, we can assume that we will reach almost 99% accuracy for this task. But, how are the concepts “dog” and “cat” encoded in the system—a system that clearly somehow “knows” what these concepts are?

Research in interpretable machine learning thus requires the consideration of both technical and philosophical notions of interpretation and representation. We propose that, for machine vision systems, this inherent transdisciplinarity

<sup>3</sup> Technically, this “flowing back” of information, also called back-propagation, is enabled by the fact that convolutional neural networks, beyond the topological perspective, are just very large differentiable equations. Hence, for each neuron we can derive the slope of the loss function with respect to this specific neuron. This allows us to then adjust the weights of this neuron into the direction of the slope, thus decreasing the error of the whole function by a tiny amount. This adjustment process, can be implemented in different ways and is most commonly realized through stochastic gradient descent.

<sup>4</sup> The difference between the two terms originally was a matter of geography: “interpretable machine learning” was used in the North American context, while “explainable artificial intelligence” was used in the European context. Today, however, both terms are used interchangeably. We will stick with interpretable machine learning in the context of this paper, to emphasize its focus on machine learning (i.e. technical systems) vs. artificial intelligence (i.e. speculative technologies).

implies linking technical concepts and concepts from visual studies/*Bildwissenschaft*. In particular, it suggests understanding the interpretation of machine vision systems as an act of image-making, both literally and metaphorically. This is why, in the following, we will look at feature visualization.

#### 4 What is feature visualization?

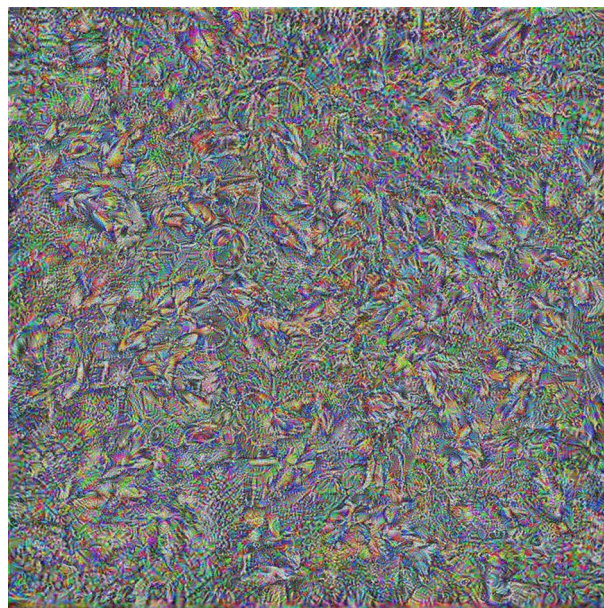
Feature visualization belongs to a range of techniques for the visual analysis of machine learning systems called *visual analytics* (Hohman et al. 2018). The idea of visual analytics is to *show how* a machine vision system perceives the world, and thus how it makes its decisions. Explanations, in visual analytics, thus, take the form of images, not of numbers or sentences. In other words, visual analytics is the visualization of machine learning systems for the sake of interpretability. Within visual analytics, feature visualization (together with attribution techniques) has become one of the most widely used methods. Originally developed by Erhan et al. (2009) and continuously improved since, it has been shown to produce remarkable results (Olah et al. 2017, 2018, 2020).

Technically, feature visualization is a straightforward optimization process. To visualize what a neuron in a deep convolutional neural network has learned, a random noise image is passed through the layers of the network up until the hidden layer that contains the neuron of interest. Normally, during the training or prediction stages, the image would be passed further on to the output layer. For the purpose of visualization, however, we are not interested in a prediction but in the “activation” of a single neuron, its individual response to a specific input image when it reaches the neuron’s layer. Hence, instead of utilizing the original loss function of the network, this response is now interpreted as its loss function. In other words, it is now the response of a single neuron that drives the “learning” process.

The important difference is that this new loss flows back through the network beyond the input layer and is used to change the *raw pixel values* of the input image. The input image is thus altered, while the network’s internal interconnections remain untouched. The altered image is then being used again as the input image during the next iteration, and so on. After a couple of iterations, the result is an image that highly activates one specific neuron.

#### 5 Perceptual bias as syntactic bias

This process, however, is called “naïve” feature visualization for a reason. In almost all cases, images obtained with it will exclusively contain very high frequencies and will thus be “illegible” in both the syntactic and semantic sense: there



**Fig. 1** Unregularized feature visualization of the “banana” class of an InceptionV3 CNN trained on the 1000 ImageNet classes in the ILSVRC2012 subset. These and the following visualizations have been generated with a custom software written by Fabian Offert in Python/PyTorch, implementing an optimized feature visualization algorithm with regularization but without natural image priors

will be no visible structure, and no recognizable content (Fig. 1).<sup>5</sup> The images may very well be the *best possible images* with regard to a specific neuron and may very well be the closest possible visualizations of what this neuron has learned. To the human observer, however, they contain no information. They are adversarial examples (Szegedy et al. 2013; Goodfellow et al. 2014a)—images that highly activate specific neurons or classes in a fully trained deep convolutional neural network, despite being utterly uninterpretable.

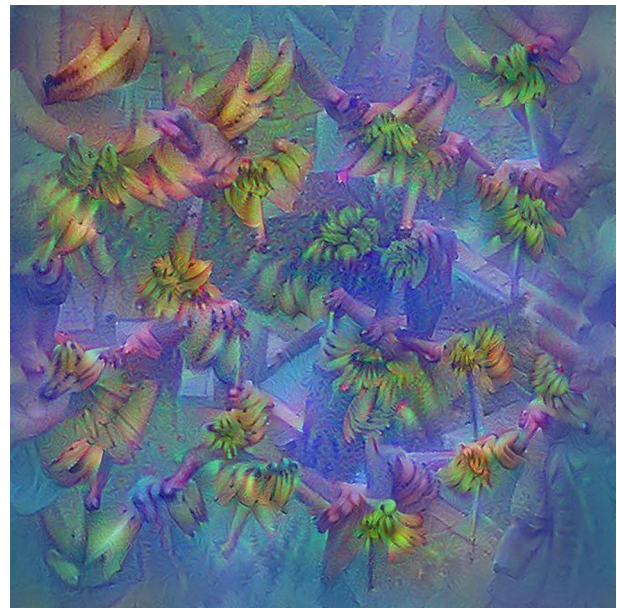
Naïve feature visualization, then, shows us a first glimpse of the peculiar perceptual topology of CNNs. Perceptual bias, here, takes the form of syntactic bias. This syntactic bias, in turn, manifests as texture bias (Geirhos et al. 2019), an inductive bias in CNNs that “naturally” appears in all common CNN architectures. Inductive biases are “general”, prior assumptions that a learning system uses to deal with new, previously unseen data.

<sup>5</sup> Concretely, to generate the feature visualization images in this paper, the following settings have been used: InceptionV3 model pre-trained on ImageNet/ILSVRC2012; stochastic gradient descent optimizer (torch.optim.SGD) with a learning rate of 0.4 and an L2 weight decay of 0.0001; optimization target: fully connected prediction layer (layer 17); 3 octaves, with 1.5×resolution increase/octave, leading to a final resolution of 672×672; 2000 iterations per octave; jitter (image is randomly shifted 32 pixels) every octave; total variation filter every 20 iterations.

At this point, it is important to note that we will not consider modifying the inductive biases of the CNN itself as a solution to the problem of perceptual bias, as, for instance, Geirhos et al. (2019) and Zhou et al. (2020) suggest. More precisely, for the purpose of our investigation, we are interested in interpretable machine learning as a narrow set of post hoc methods to produce explanations. Thus, we will also not take the field of representation learning (Bengio et al. 2013) into consideration, which is concerned with the development of mechanisms that enforce the learning of “better” representations. This restriction to the scope of our investigation has three main reasons. The first reason is the post hoc nature of the bias problem. While efforts to build resistance to bias into machine learning models exist, there is, at the moment, no clear incentive for industry practitioners to do so, except for marketing purposes. It can thus be assumed that, in real-world scenarios, the detection and mitigation of bias will be mostly a post-hoc effort. The second reason is a simple historical reason. Thousands of machine learning models based on the exact perceptual topologies under investigation here have already been deployed in the real world. Thus, it is of vital importance to understand, and be able to critique, such models and their perceptual biases. Finally, while impressive progress has been made in other areas of machine learning (Cranmer et al. 2020), in machine vision, controlling and harnessing inductive biases can still be considered an open problem. Recent research suggests that at least one established principle of gestalt theory (the law of closure) does emerge in CNNs (Kim et al. 2019; Ritter et al. 2017; Feinman and Lake 2018) as an inductive bias. Overall, however, the inductive biases of CNNs are still unclear (Cohen and Shashua 2017), and thus un-manageable.

Given these restrictions, the only option to mitigate this specific textural aspect of perceptual bias is to not change the model, but to change our image of it. In the case of feature visualization, it means adding back *representational capacity* to these images. It means introducing constraints—in other words, different biases—that allow the production of images that are images *of something*, instead of “just” images. Importantly, any such constraint, however, automatically moves the image further away from showing the actual perceptual topology of a CNN. It becomes less of a visualization, and more of a reconstruction. This trade-off is the core problem of perceptual bias: it can only be overcome by shifting towards different, “better” biases, i.e. biases that shape *our* perception of the visual world.

One strategy to “add back” the representational capacity to feature visualization is regularization (Fig. 2). Regularization, here, simply means adding additional constraints to the optimization process. This can be achieved either by adapting the loss function—for instance, using a quadratic loss function instead of just taking the mean of some values—or by applying



**Fig. 2** Regularized feature visualization of the “banana” class of an InceptionV3 CNN trained on the 1000 ImageNet classes in the ILS-VRC2012 subset

transformations to the input image in regular intervals, for instance, every few iterations in the optimization process.

Erhan et al. (2009) introduced the concept of activation maximization, the core idea of iteratively optimizing an image to highly activate a selected neuron. From there, more and more elaborate regularization techniques started to appear, each introducing concrete suggestions for signal processing operations on the input image between iterations, on top of more common regularization techniques introduced through the loss function, like L2 regularization. Among these are jitter (Mordvintsev et al. 2015), blur (Yosinski et al. 2015), total variation filters (Mahendran and Vedaldi 2015), bilinear filters (Tyka 2016), stochastic clipping (Lipton and Tripathi 2017) and Laplacian pyramids (Mordvintsev 2016). Mordvintsev et al. (2015) also introduce the idea of octave-based optimization, enabling significantly higher image resolution. What all these techniques have in common is some kind of frequency penalization, i.e. the active avoidance of input images evolving into adversarial examples, either through optimizing for transformation robustness (Mordvintsev et al. 2015) or through direct filtering (all others).

## 6 Perceptual bias as semantic bias

Despite all regularization efforts, however, feature visualizations often still present “strange mixtures of ideas” (Olah et al. 2018). Visualizing higher level neurons in particular



**Fig. 3** Left: regularized feature visualization of the “violin” class of an InceptionV3 CNN trained on the 1000 ImageNet classes in the ILSVRC2012 subset. Right: George Braque, *Violin and Candlestick* (1910)



produces ambiguous results, images that might, or might not, show proper “objects”. To learn more about the logic of representation in CNNs, we thus have to ask: what is the relation between technical and semantic units, between artificial neurons and meaningful concepts, in CNNs?

Trivially, at least for higher level neurons, individual feature visualization images must always have a degree of ambiguity that is directly correlated to the diversity of the training set. After all, the network has to be able to successfully classify a range of instances of an object with very different visual properties. In that sense, reality is “distributed”, and it is no surprise that feature visualization images will reflect different manifestations of, and perspectives on, an object, akin to Cubist paintings (Fig. 3).

But, the entanglement of concepts in the internal representations of a CNN goes beyond this “natural” ambiguity. Generally, we can state that, in all predictions of a CNN, all neurons play “a” role. Even if their role is just to stop the information flow, i.e. to pass on zero values to the next layer, these one-way streets are in no way less relevant to the classification accuracy of the whole system than all other neurons. In a way, concepts are thus “dissolved”, or “entangled”, when they are learned, and represented, by a CNN. Early work by Szegedy et al. (2013) suggests that this entanglement is inevitable and absolute. Later work by Bau et al. (2017, 2018) shows that some neural network architectures (GANs in particular) are less “naturally entangled” than others. Generally, however, significant supervision or, again, artificial inductive biases (Locatello et al. 2019) are required to “disentangle” CNNs, and arrive at a meaningful correspondence of technical and semantic units.<sup>6</sup>

Perceptual bias, here, thus takes the form of semantic bias. Other than in the case of adversarial examples/texture bias, where perceptual bias affects the formal aspects of the visualization, here, it concerns aspects of meaning. Objects,

for us, are necessarily spatially cohesive. If they are represented by CNNs, however, they lose this spatial coherence, different aspects of an object are attached to different neurons, which, in turn, get re-used in the detection of other objects. This missing coherence does not interfere with the CNN’s ability to detect or classify spatially coherent objects in images but enables it.

For feature visualization, which visualizes CNNs in their “natural”, entangled state, reaching semantic interpretability thus implies the introduction of even more constraints. These additional constraints are so called natural image priors. Just as regularization is a syntactic constraint, biasing the visualization towards a more natural frequency distribution, so-called natural image priors are a semantic constraint, biasing the visualization towards separable image objects.

To produce natural image priors, Dosovitskiy and Brox (2016) propose to use a GAN generator. Generative adversarial networks (GANs) have received a lot of attention since 2015/2016 for being able to generate realistic images in an unsupervised way from large datasets. They were originally introduced by Goodfellow et al. (2014) and have since been steadily improved and extended to other tasks beyond image generation. The term “generative adversarial network” refers to an ensemble of two convolutional neural networks that are trained together. The notion of “adversarial” describes the dynamic between the two networks, where a generator attempts to fool a discriminator into accepting its images. Whereas the discriminator is thus a regular, image-classifying convolutional neural network, the generator is a reverse CNN that outputs an image. Its input is a number from a

<sup>6</sup> Unfortunately, we cannot give a full review of the relevant computer science literature regarding representation learning and disentanglement here. Instead, we would like to refer the reader to the review articles by Bengio et al. (2013) and Locatello et al. (2019).

latent vector space, i.e. a high-dimensional space. Concretely, a random sample from this latent space is forwarded through the layers of the generator, and an image is created at its last layer. This generated image, or alternatively an image from a supplied image dataset, is then evaluated by the discriminator, which attempts to learn how to distinguish generator-generated images from images that come from the supplied dataset. Importantly, the discriminator's loss is back-propagated all the way to the generator, which allows the generator to adjust its weights depending on its current ability to fool the discriminator. Eventually, this dynamic results in a generator that has learned how to produce images that look like they come from the supplied dataset, i.e. that have similar features as the images in the supplied dataset but are not part of the dataset. Overfitting, here, is avoided by never giving the generator access to the supplied dataset. Its only measure of success is how well it is able to fool the discriminator.

Nguyen et al. (2016) turn this technique into a dedicated feature visualization method by applying the paradigm of activation maximization to the input of a GAN generator. Instead of optimizing an input image directly, i.e. in pixel space, it is thus optimized in terms of the generator, i.e. in feature space. This technique has three main benefits: (a) Optimization in feature space automatically gets rid of high frequency artifacts. The generator, trained to produce *realistic* images, will never reconstruct an adversarial example from a latent feature representation. (b) Optimization in feature space introduces a strong natural image prior. More precisely, it introduces a natural image prior that corresponds directly to the level of realism that can be attained with the generator. (c) Finally, optimization in feature space requires neither the generator, nor the network that is being analyzed to be trainable, as back propagation just passes through an image: a feature representation is fed into the generator, the generator produces an image, this image is fed into the network that is being analyzed, which in turn produces a loss with regard to the neuron being analyzed, as in regular feature visualization.

This means, however, that the images that can be produced with this feature visualization method are entirely confined to the latent space of the specific GAN generator employed. Where regularization constrains the space of possible images to those with a “natural” frequency distribution, natural image priors constrain the space of possible images to the distribution of a GAN generator. In both cases, interpretable images are the result. These interpretable images, however, do not reflect the perceptual topology of the analyzed CNN. On the contrary, they intentionally get rid of the non-humanness that defines this topology, translating it into a human mode of perception that, in this form, simply does not exist in the CNN. To be images of something, feature

visualizations have to be freed from the very mode of perception they are supposed to illustrate.

## 7 Feature visualizations as technical metapictures

As we have seen, the perceptual topology of machine vision systems, based on CNNs, is not “naturally interpretable”. It is biased towards a distributed, entangled, deeply non-human way of representing the world. Mitigating this perceptual bias thus requires a forced “making legible”. Feature visualization, as we have seen, is one possibility to achieve this forced legibility. However, feature visualization also exemplifies an essential dilemma: the representational capacity of feature visualization images is *inverse proportional* to their legibility. Feature visualizations that show “something” are further removed from the actual perceptual topology of the machine vision system than feature visualizations that show “nothing” (i.e. illegible noise).

There is thus an irreconcilable difference between the human and machine perspective. As Thomas Nagel reminds us, there is a “subjective character of experience” (Nagel 1974), a surplus generated by each specific perceptual approach to the world that can never be “translated”. Even if an external observer would be able to attain all the facts about such an inherently alien experience (analyze it in terms of “functional states”), they would still not be able to reconstruct said experience from these facts.

Feature visualizations, then, should not be understood so much as direct “traces” or “reproductions” of the perceptual topology of machine vision systems (analog to the technical images of photography) but more as indirect “illustrations”, as “visualizations” in the literal sense of forcibly making-visual (and thus making visible and subsequently making interpretable) the non-visual.

We thus propose to understand these images as *technical metapictures*, a term we adapt from W. J. T. Mitchell's picture theory (Mitchell 1995). For Mitchell, metapictures are pictures that are “deeper” than “regular” pictures, as they incorporate a form of recursion: they are representations of representation “pictures about pictures” (Mitchell 1995, 36). Mitchell identifies certain abilities of these pictures. “The metapicture [...] is the place where pictures reveal and ‘know’ themselves, where they reflect on the intersections of visibility, language, and similitude, where they engage in speculation and theorizing on their own nature and history” (Mitchell 1995, 82). They are not only self-reflective but reflective on imagery and perception. “The metapicture is a piece of moveable cultural apparatus, one which may serve a marginal role as illustrative device or central role as a kind of summary image, what I have called a ‘hypericon’

that encapsulates an entire episteme, a theory of knowledge” (Mitchell 1995, 49).

The technical metapictures that feature visualization produces realize exactly this idea of a “summary image.” They promise not a theory of images but a theory of seeing. More precisely, their promise is exactly that of interpretable machine learning: to provide an intuitive visual theory of the non-intuitive perceptual topology of neural networks. In a sense, technical metapictures, and their use in interpretable machine learning, are thus an operationalization of the notion of metapicture itself.

For Mitchell, this epistemological power of metapictures, then, equips them with a sort of agency. Metapictures “don’t just illustrate theories of picturing and vision: they show us what vision is, and picture theory” (Mitchell 1995, 57). This agency, however, is actualized only if and when it comes into contact with a viewer. To *make sense*, to actually provide the reflection on images and vision that they promise, metapictures require a viewer. In the case of feature visualization, this interpretation has to happen not only on the level of the viewer but also on the technical level, where a significant effort has to be made to translate the anti-intuitive perceptual topology of a machine vision system into human-interpretable images in the first place. This includes *adding information from the outside*, for instance in the form of natural image priors. In other words, technical metapictures manifest an implicit, technical notion of interpretation, that is inseparable from the explicit interpretation that they also require.

## 8 Learning from technical metapictures

If we understand feature visualizations as technical metapictures, that is, if we look at them not as representations of machine vision systems but as reflections on the perceptual limits of machine vision systems, we can re-imagine them as a method of critique that can detect deeper forms of bias.

Taking up the idea that “adversarial examples are not bugs, they are features,” (Ilyas et al. 2019) the peculiar ways of machine seeing described above would be utilized to detect anomalies in datasets that go beyond problematic or socially charged categories.

One of the main problems of image datasets is the diversity and heterogeneity of the real-life visual world, which is inherently difficult to capture by technical means. How is the diversity and heterogeneity of the real-life visual world to be represented in a set of images of bounded size? Historically, this problem is related to the general epistemological problem of creating taxonomies of what exists. The appearance of solutions to this problem, and their failure, can be traced back to the beginnings of the enlightenment in the seventeenth century, with Descartes, Leibniz, Wilkins and

others speculating on scientific approaches to the design of inventories of the world. Importantly, since then, taxonomies have been thought of as symbolic, i.e. as textual representations of the world.

With the introduction of images, however, and the creation of visual taxonomies, conceptual problems already present in symbolic taxonomies have been amplified by an order of magnitude: any claim to completeness is now not only a claim to a completeness of *concepts*, but of *manifestations*. Put differently: while symbolic taxonomies have to deal with an infinity of *abstractions*, visual taxonomies have to deal with an infinity of *individual objects*. Of course, this grounding of the learning problem in “real” data, in “ground truths”, to use the terminology of machine learning, is exactly the point. The hypothesis is that useful abstractions will appear automatically by means of the exposure of a system to a manifold of manifestations. This, however, increases the number of potential points of bias injection *to the size of the dataset*. In other words: it is *in every single image* that potential biases come into play, and thus the design and use of visual taxonomies becomes of crucial importance.

ImageNet (Deng et al. 2009) is a large-scale digital image dataset intended to facilitate the automatic classification of images with regard to depicted objects (object recognition). It consists of over fourteen million images in over 21,000 categories. Originally conceived and first presented at CVPR in 2009 by Fei Fei Li and others, ImageNet consolidated an existing taxonomy, WordNet (Miller 1985) and image resources freely available on the Internet. ImageNet stands out as being the first large-scale dataset using distributed labor acquired via Amazon Mechanical Turk to solve the problem of cheap data vs. expensive labels, i.e. the circular problem of first having to hand-annotate images to automate the process of annotating images.<sup>7</sup>

Since 2010, the ImageNet project has run the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a competition to identify the state of the art in image classification (Russakovsky et al. 2015). The challenge usually only requires the participating models to correctly classify images from an ImageNet subset. Importantly, because the images contained in ImageNet are “scraped” (i.e. downloaded) from the Internet and thus might be subject to copyright, the ImageNet project only supplies URLs to the images. As

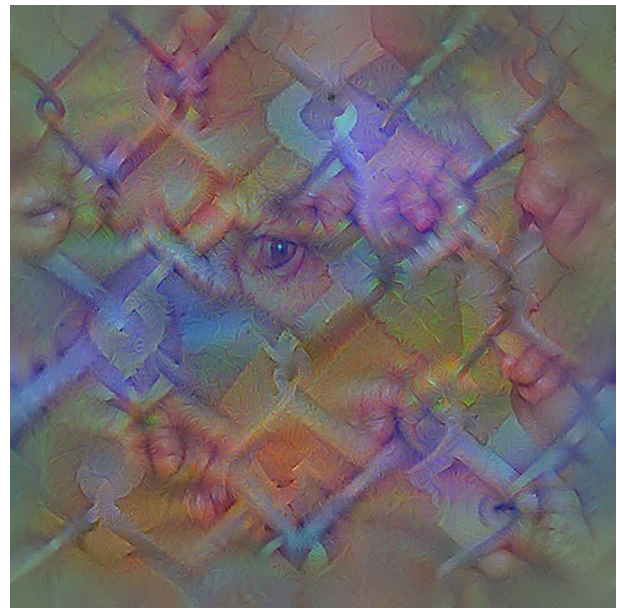
<sup>7</sup> A proper analysis of the entanglement of machine learning and labor lies outside the scope of this paper. Nevertheless it should be mentioned, as brilliantly analyzed, among others, by Daston (2018) and Pasquinelli (2019a) that machine learning is an essential Taylorist idea. As Charles Babbage writes already in 1832: “[T]he division of labor can be applied with equal success to mental as to mechanical operations, and [...] it ensures in both the same economy of time” (Babbage 2010, 295).

the project has been in existence since 2009, many of these URLs are not accessible anymore today. These two factors have facilitated the current real-life use of ImageNet: in most contemporary applications, it is actually just a subset that is being used to train and test computer vision systems, most prominently the subset created for the 2012 ILSVRC.

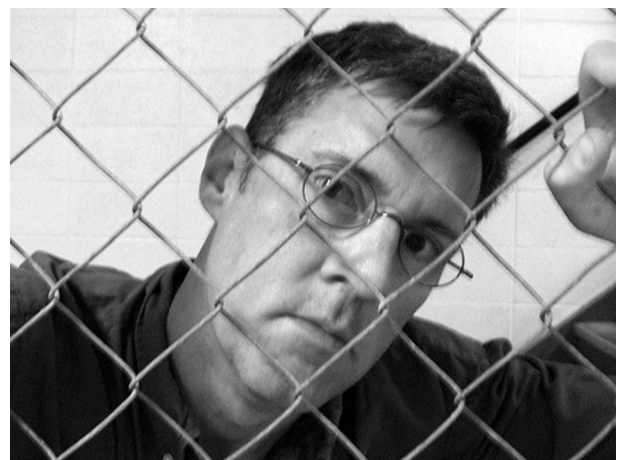
In the past few years, much has been written about biased datasets. ImageNet has been at the center of this debate (Malevé 2019), with Trevor Paglen and Kate Crawford’s “Excavating AI” project (Crawford and Paglen 2019) receiving broad media attention. What Crawford and Paglen rightfully criticized was essentially the historical debt of the dataset, which operates with a taxonomy based on WordNet. WordNet, in turn, includes many categories which are neutral as textual categories (e.g. “terrorist”), but have necessary social and political implications when “illustrated” with images. The failure of the ImageNet team to remove these and similar categories, and even stock some of them with actual images based on the aforementioned distributed micro-labor, rightfully led to a public outcry in reaction to the “Excavating AI” project, and subsequently to the removal of these categories from the dataset.

Nevertheless, there is a need for more elaborate methods of image dataset critique, and feature visualization could provide a potential basis for such a method. One example application is the detection of dataset anomalies in art historical corpora (Offert 2018). A more broad critical approach would be the analysis of highly common datasets like ImageNet, which are not only used “as is” in real-life classification scenarios but even more often used to pre-train classifiers which are then fine-tuned on a separate dataset, potentially introducing ImageNet biases into a completely separate classification problem.

To demonstrate this approach, we visualized and selected the output neurons for several classes of an InceptionV3 model (Szegedy et al. 2016) pre-trained on ImageNet/ILSVRC2012 hand-selecting visualizations that show some non-intuitive properties of the ImageNet dataset. For instance, for the “fence” class output neuron (Fig. 4) we see that the network has not only picked up the general geometric structure of the fence but also the fact that many photos of fences in the original dataset (that was scraped from the Internet) seem to contain people confined behind these fences. This can be verified by analyzing the 1300 images in the dataset class, which indeed show some, but not many scenes of people confined behind fences. Cultural knowledge, more specifically, a concrete representation of cultural knowledge defined by the lense of stock photo databases and hobby photographers, is introduced here into a supposedly objective image classifier. Importantly, this also means that images of people behind fences will appear more fence-like to the classifier. The relevance of this consequence is revealed by a Google reverse image search: for a sample image (Fig. 5)



**Fig. 4** Regularized feature visualization of the “fence” class of an InceptionV3 CNN trained on the 1000 ImageNet classes in the ILSVRC2012 subset

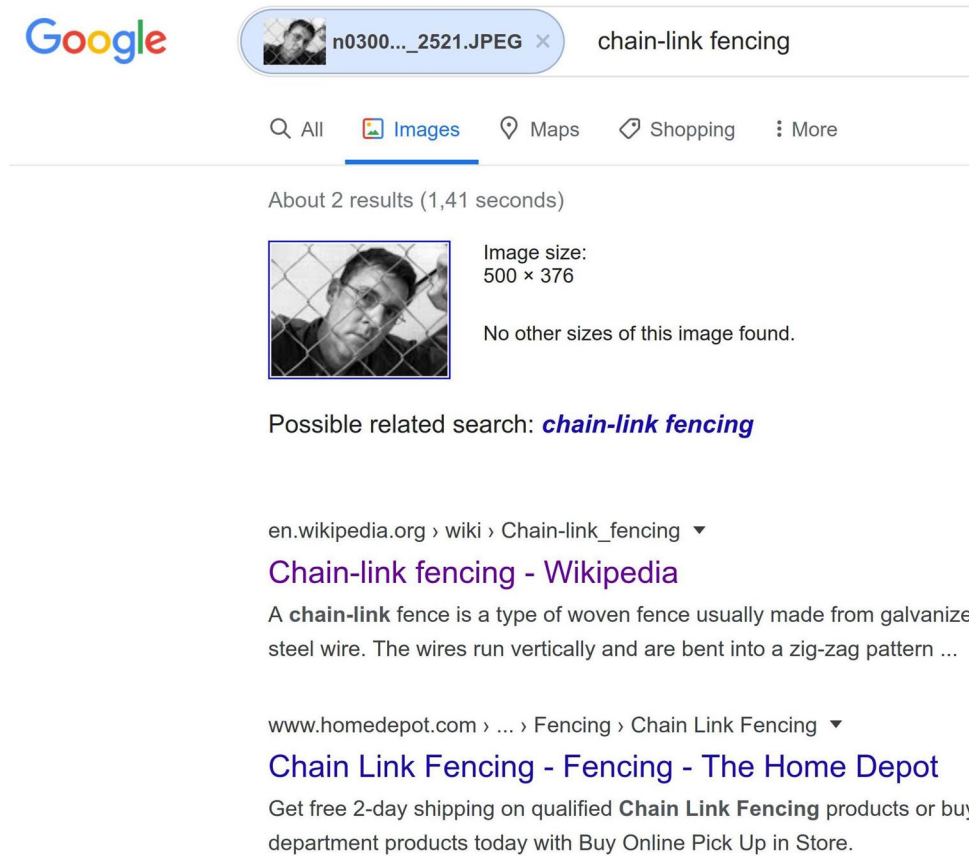


**Fig. 5** Sample image from the ILSVRC2012 “chain link fence” class. Note that there are only a few images (between 1 and 5% of the class, depending on what counts as “behind”) that show people behind fences

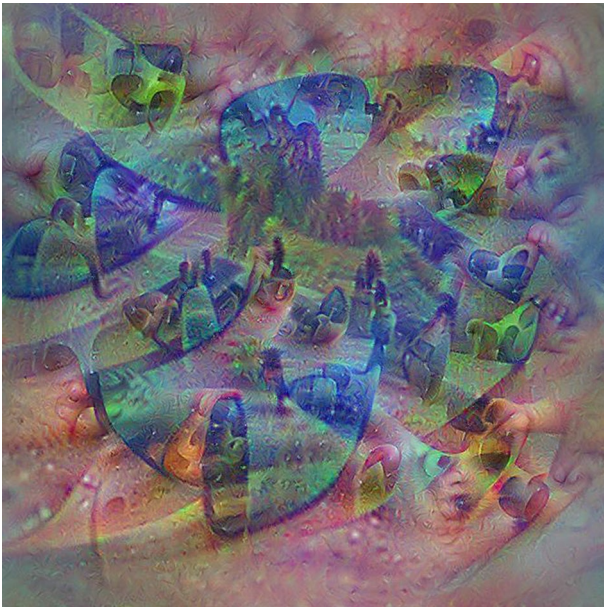
from the “fence class”, despite the prominence of the person compared to the actual fence, the search produces the Wikipedia entry for “chain link fencing” (Fig. 6), suggesting an unverifiable but likely connection between the Google image search algorithm and ImageNet/ILSVRC2012.

For the “sunglass” class (Fig. 7), the diversity that the respective output neuron has to deal with becomes obvious in the entanglement of actual sunglasses and body parts. More surprisingly is the inclusion of mirrored landscapes.

**Fig. 6** A Google reverse image search for this specific image, despite the fact that the image does not exist on the Internet anymore, and despite the prominence of the person compared to the actual fence, produces the Wikipedia entry for “chain link fencing”, suggesting an unverifiable but likely connection between the Google image search algorithm and ImageNet/ILSVRC2012. A text search for “chain-link fencing” produces no “people behind fences” scenes



The screenshot shows a Google search interface. At the top left is the Google logo. To its right is a search bar containing a small thumbnail of a person behind a fence, the filename 'n0300...\_2521.JPEG', and the search term 'chain-link fencing'. Below the search bar are navigation tabs for 'All', 'Images', 'Maps', 'Shopping', and 'More', with 'Images' selected. The results section shows 'About 2 results (1,41 seconds)'. A single image result is displayed, showing a person behind a chain-link fence. To the right of the image, it says 'Image size: 500 × 376' and 'No other sizes of this image found.' Below the image, it says 'Possible related search: *chain-link fencing*'. Further down, there are two search results: one from Wikipedia titled 'Chain-link fencing - Wikipedia' and one from The Home Depot titled 'Chain Link Fencing - Fencing - The Home Depot'.



**Fig. 7** Regularized feature visualization of the “sunglass” class of an InceptionV3 CNN trained on the 1000 ImageNet classes in the ILSVRC2012 subset



**Fig. 8** All images in the ILSVRC2012 “sunglass” class, sorted by VGG19 fc1 features and plotted with UMAP. Notice the cluster of “landscape mirror”-type images in the center

The original dataset class, as a closer investigation reveals (Fig. 8), is heavily biased towards a specific depiction of sunglasses, popular with stock photo databases and hobby photographers alike: a close-up of a pair of sunglasses that also shows parts of the surrounding landscape (and/or the photographer). The fact that ILSVRC2012 was scraped from the Internet, disregarding aspects like diversity in composition and style, again, leads to an over-specificity of the learned representation. That the mirrored landscapes are desert landscapes (as originally assumed by the authors), however, is a chimera, showing how much artificial contextualization (bringing in additional information from the outside to aid interpretation) matters in both the technical and human interpretations.

## 9 Conclusion

Analyzing and understanding perceptual bias in machine vision systems requires reframing it as a problem of interpretation and representation, for which we have adapted W. J. T. Mitchells notion of the metapicture. Technical metapictures, we have argued, mirror the act of interpretation in the technical realm: regularization and natural image priors make feature visualization images legible before any interpretation can take place. Paradoxically, however, as the representational capacity of feature visualization images is inverse proportional to their legibility, this pre-interpretation presents itself as a massive technical intervention as well, that disconnects the visualization from the visualized. Nevertheless, feature visualization can also provide a potential new strategy to mitigate bias, if the fact that technical meta-images primarily encapsulate the limitations of machine vision systems is taken into account. All of this suggests that critical machine vision is an essentially humanist endeavor that calls for additional transdisciplinary investigations at the interface of computer science and visual studies/*Bildwissenschaft*.

**Funding** Open Access funding provided by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Arnold T, Tilton L (2019) Depth in deep learning: knowledgeable, layered, and impenetrable. <https://statsmaths.github.io>. Accessed 8 July 2020
- Babbage C (2010) Babbage's calculating engines: being a collection of papers relating to them, their history and construction. Cambridge University Press, Cambridge
- Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning. <https://www.fairmlbook.org>. Accessed 8 July 2020
- Bau D, Zhou B, Khosla A, Oliva A, Torralba A (2017) Network dissection: quantifying interpretability of deep visual representations. In: IEEE conference on computer vision and pattern recognition (CVPR), pp. 6541–6549
- Bau D, Zhu J-Y, Strobel H, Zhou B, Tenenbaum JB, Freeman WT et al (2018) GAN dissection: visualizing and understanding generative adversarial networks. arXiv preprint [arXiv: 1811.10597](https://arxiv.org/abs/1811.10597)
- Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 35(8):1798–1828
- Benjamin R (2019) Race after technology: abolitionist tools for the new Jim Code. Wiley, New York
- Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency (FAT\*)
- Cohen N, Shashua A (2017) Inductive bias of deep convolutional networks through pooling geometry. arXiv preprint [arXiv: 1605.06743](https://arxiv.org/abs/1605.06743)
- Cranmer M, Sanchez-Gonzalez A, Battaglia P, Xu R, Cranmer K, Spergel D et al (2020) Discovering symbolic models from deep learning with inductive biases. arXiv preprint [arXiv: 2006.11287](https://arxiv.org/abs/2006.11287)
- Crawford K, Paglen T (2019) Excavating AI: the politics of images in machine learning training sets. <https://www.excavating.ai/>. Accessed 8 July 2020
- Daston L (2018) Calculation and the division of labor, 1750–1950. *Bull Ger Hist Inst* 62(Spring):9–30
- Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 248–255
- Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint [arXiv: 1702.08608](https://arxiv.org/abs/1702.08608)
- Dosovitskiy A, Brox T (2016) Generating images with perceptual similarity metrics based on deep networks. In: Advances in neural information processing systems, pp 658–666
- Dumoulin V, Visin F (2016) A guide to convolution arithmetic for deep learning. arXiv preprint [arXiv: 1603.07285](https://arxiv.org/abs/1603.07285)
- Erhan D, Bengio Y, Courville A, Vincent P (2009) Visualizing higher-layer features of a deep network. Université de Montréal, Montreal
- Feinman R, Lake BM (2018) Learning inductive biases with simple neural networks. arXiv preprint [arXiv: 1802.02745](https://arxiv.org/abs/1802.02745)
- Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. In: ACM conference on fairness, accountability, and transparency (FAT\*)
- Garvie C, Bedoya A, Frankle J (2016) The perpetual line-up: Unregulated police face-recognition in America. Georgetown Law, Center on Privacy and Technology. <https://www.perpetuallineup.org>. Accessed 8 July 2020
- Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W (2019) ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint [arXiv: 1811.12231](https://arxiv.org/abs/1811.12231)
- Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: An overview of interpretability of

- machine learning. In: 2018 IEEE 5th international conference on data science and advanced analytics (DSAA), pp 80–89
- Goodfellow IJ, Shlens J, Szegedy C (2014a) Explaining and harnessing adversarial examples. arXiv preprint [arXiv: 14126572](https://arxiv.org/abs/1412.6572)
- Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S et al (2014b) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
- Hohman FM, Kahng M, Pienta R, Chau DH (2018) Visual Analytics in deep learning: an interrogative survey for the next frontiers. IEEE Trans Vis Comput Graph 25(8):2674–2693
- Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A (2019) Adversarial examples are not bugs, they are features. arXiv preprint [arXiv: 190502175](https://arxiv.org/abs/1905.02175)
- Kim B, Reif E, Wattenberg M, Bengio S (2019) Do neural networks show gestalt phenomena? An exploration of the law of closure. arXiv preprint [arXiv: 190301069](https://arxiv.org/abs/1903.01069)
- Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
- Kurenkov A (2020) Lessons from the PULSE model and discussion. The gradient. <https://thegradient.pub/pulse-lessons/>. Accessed 8 July 2020
- Lipton ZC (2016) The mythos of model interpretability. In: 2016 ICML workshop on human interpretability in machine learning (WHI), New York, NY
- Lipton ZC, Tripathi S (2017) Precise recovery of latent vectors from generative adversarial networks. arXiv preprint [arXiv: 170204782](https://arxiv.org/abs/1702.04782)
- Locatello F, Bauer S, Lucic M, Gelly S, Schölkopf B, Bachem O (2019) Challenging common assumptions in the unsupervised learning of disentangled representations. arXiv preprint [arXiv: 181112359](https://arxiv.org/abs/1811.12359)
- Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Advances in neural information processing systems, pp 4765–4774
- Mahendran A, Vedaldi A (2015) Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 5188–5196
- Malev N (2019) An introduction to image datasets. Available from: <https://unthinking.photography/articles/an-introduction-to-image-datasets>
- Merler M, Ratha N, Feris RS, Smith JR (2019) Diversity in faces. arXiv preprint [arXiv: 190110436](https://arxiv.org/abs/1901.10436)
- Miller GA (1985) Wordnet: a dictionary browser. In: Proceedings of the first international conference on information in data
- Minsky ML, Papert S (1988) Perceptrons. MIT Press, Cambridge
- Mitchell WJT (1995) Picture theory: essays on verbal and visual representation. University of Chicago Press, Chicago
- Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc. <https://doi.org/10.1177/2053951716679679>
- Mittelstadt B, Russel C, Wachter S (2019) Explaining explanations in AI. In: ACM conference on fairness, accountability, and transparency (FAT\*)
- Mordvintsev A (2016) Deep dreaming with TensorFlow. Available from: <https://github.com/tensorflow/tensorflow/blob/master/tensorflow/examples/tutorials/deepdream/deepdream.ipynb>
- Mordvintsev A, Olah C, Tyka M (2015) Inceptionism: going deeper into neural networks. Available from: <https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>
- Nagel T (1974) What is it like to be a bat? Philos Rev 83(4):435–450
- Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J (2016) Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Advances in neural information processing systems, pp 3387–3395
- Offert F (2018) Images of image machines. Visual interpretability in computer vision for art. In: European conference on computer vision, pp 710–715
- Offert F (2020) There is no (real world) use case for face super resolution. [https://zentralwerkstatt.org/post\\_PULSE.html](https://zentralwerkstatt.org/post_PULSE.html). Accessed 8 July 2020
- Olah C, Mordvintsev A, Schubert L (2017) Feature visualization. Distill. Available from: <https://distill.pub/2017/feature-visualization>
- Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, et al (2018) The building blocks of interpretability. Distill. Available from: <https://distill.pub/2018/building-blocks>
- Olah C, Cammarata N, Schubert L, Goh G, Petrov M, Carter S (2020) An overview of early vision in InceptionV1. Distill. Available from: <https://distill.pub/2020/circuits/early-vision/>
- Pasquinelli M (2019a) The origins of Marx’s general intellect. Radic Philos 2(6)
- Pasquinelli M (2019b) Three thousand years of algorithmic rituals: the emergence of AI from the computation of space. eFlux. Available from: <https://www.e-flux.com/journal/101/273221/three-thousand-years-of-algorithmic-rituals-the-emergence-of-ai-from-the-computation-of-space/>
- Pasquinelli M, Joler V (2020) The Nooscope manifested: artificial intelligence as instrument of knowledge extractivism. KIM HfG Karlsruhe and Share Lab. Available from: <https://nooscope.ai>
- Ribeiro MT, Singh S, Guestrin C (2016) “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, pp 1135–1144
- Ritter S, Barrett DG, Santoro A, Botvinick MM (2017) Cognitive psychology for deep neural networks: a shape bias case study. arXiv preprint [arXiv: 170608606](https://arxiv.org/abs/1706.08606)
- Rosenblatt F (1957) The perceptron. A perceiving and recognizing automaton. Cornell Aeronautic Laboratory, Buffalo
- Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S et al (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
- Selbst AD, Barocas S (2018) The intuitive appeal of explainable machines. Fordham Law Rev 87:1085
- Selbst AD, Friedler S, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: ACM conference on fairness, accountability, and transparency (FAT\*)
- Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I et al (2013) Intriguing properties of neural networks. arXiv preprint [arXiv: 13126199](https://arxiv.org/abs/1312.6199)
- Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2818–2826
- Tyka M (2016) Class visualization with bilateral filters. <https://mtyka.github.io/deepdream/2016/02/05/bilateral-class-vis.html>. Accessed 8 July 2020
- Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. arXiv preprint [arXiv: 150606579](https://arxiv.org/abs/1506.06579)
- Zhou A, Knowles T, Finn C (2020) Meta-learning symmetries by reparameterization. arXiv preprint [arXiv: 2007.02933](https://arxiv.org/abs/2007.02933)

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Computer vision, human senses, and language of art

Lev Manovich<sup>1</sup>

Received: 1 June 2020 / Accepted: 14 October 2020 / Published online: 22 November 2020  
© Springer-Verlag London Ltd., part of Springer Nature 2020

## Abstract

What is the most important reason for using Computer Vision methods in humanities research? In this article, I argue that the use of numerical representation and data analysis methods offers a new language for describing cultural artifacts, experiences and dynamics. The human languages such as English or Russian that developed rather recently in human evolution are not good at capturing analog properties of human sensorial and cultural experiences. These limitations become particularly worrying if we want to compare thousands, millions or billions of artifacts—i.e. to study contemporary media and cultures at their new twenty-first century scale. When we instead use numerical measurements of image properties standard in Computer Vision, we can better capture details of a single artifact as well as visual differences between a number of artifacts—even if they are very small. The examples of visual dimensions that numbers can capture better than languages include color, shape, texture, contours, composition, and visual characteristics of represented faces, bodies and objects. The methods of finding structures and relationships in large numerical datasets developed in statistics and machine learning allow us to extend this analysis to very big datasets of cultural objects. Equally importantly, numerical image features used in Computer Vision also give us a new language to represent gradual and continuous temporal changes—something which natural languages are also bad at. This applies to both single artworks such as a film or a dance piece (describing movement and rhythm) and also to changes in visual characteristics in millions of artifacts over decades or centuries.

**Keywords** Computer vision · Digital humanities · Cultural analytics · Language of art

## 1 Computer vision and digital humanities

Researches in humanities research, write and argue about cultural images. They analyze and interpret content, visual style, author's intentions, audience reception, meanings, emotional effects, and other aspects of images' creation and circulation. Researchers in Computer Vision field also work with images, but their goals are very different—to teach computers to automatically understand images and enable automatic actions using visual information. The examples of these applications include their use in self-driving cars, industrial and home robots, medical diagnostics, content-based image retrieval.

What are the intellectual consequences of adopting Computer Vision methods in humanities research? What happens to humanists' understanding of images and assumptions

about how to describe and study visual cultures in this meeting? How can we bring together assumptions and goals of AI research in general and the assumptions and goals of the humanities that think of the study of cultural artifacts as their exclusive domain? (In addition to humanities fields such as art history, musicology, performance studies, cinema studies, literary studies, digital culture studies and game studies, these questions are also relevant for social science fields that deal with visual culture such as cultural anthropology, sociology, culture studies, communication and media studies.)

In this article, I will discuss the most important consequence of using Computer Vision in humanities, as I see it. Certainly, the achievements of Computer Vision such as detection of objects and scene types, people, and faces, pose estimation or optical character recognition all have their uses in art history, cinema and media studies, game studies, archeology, and so on. Working with the researchers in these fields, computer scientists also develop new tools for specific problems (Visart 2018). These applications and tools allow answering existing and generating new questions, and this work is certainly important. But in my view, they don't affect

---

✉ Lev Manovich  
manovich.lev@gmail.com

<sup>1</sup> Program in Computer Science, The Graduate Center, City University of New York, New York, USA



in a fundamental way how we see images in humanities. What does affect this is the way Computer Vision describes images, as I will explain below.

## 2 Computer vision and digital humanities

We can find examples of using computational techniques to analyze single artworks or small groups of artworks carried out already for a number of decades. In the case of visual arts, such work was aimed to help in restoration, conservation, material and structure characterization, authentication, and dating. It made a good use of Digital Image Processing techniques, but it did not challenge existing methods of classifying, describing and narrating and exhibiting art. We can find a similar logic in other fields that use computational methods such as archeology. For example, one recent paper presents a method for automatically fitting together available artifact pieces together. This is a useful application for archeology, but it does not lead to big new ideas for the field (Derech et al. 2018).

While these applications were and continue to be dominant, some researchers were also using methods from Image Processing, Computer Vision and Computer Graphics to do something new for art history—come up with mathematical descriptions of various characteristics of art images such as brushstrokes, lighting, and composition. As the key researcher in the area David G. Stork pointed out in his 2009 overview of this research, “In some circumstances, computers can analyze certain aspects of perspective, lighting, color, the subtleties of the shapes of brush strokes better than even a trained art scholar, artist, or connoisseur. Rather than replacing connoisseurship, these methods—like other scientific methods such as imaging and material studies—hold promise to enhance and extend it, just as microscopes extend the powers of biologists” (Stork 2009). Today the use of computers to mathematically describe cultural artifacts and analyze quantitatively and interpret cultural patterns based on such descriptions has become popular in some areas of humanities such as literary studies and history. However, this did not happen yet on any significant scale in art history, film and media studies, game studies or other fields that analyze visual culture. However, there have been a few inspiring researches projects done by computer scientists working together with humanists. Among them, I want in particular mention work by Impett and Moretti (2017). They carefully translate ideas of early twentieth century art historian Aby Warburg into an interactive tool while probing theoretically and critically Warburg ideas. Such work stands in contrast to more common references to books in art history, media production, graphic design and other fields in computer science research that borrow ideas from these books to build automatic systems.

Although the humanists that study the visual have been slow to make use of computers, many researchers in Computer Vision and also other areas of Computer Science, Artificial Intelligence and Machine Learning have started to do exactly this after 2006. The rise of social media platforms including Flickr and Instagram (launched in 2004 and 2010, respectively) and availability of their big data via APIs led to the burgeoning research presented in hundreds of thousands of conference papers and journal articles. While the earlier paradigm I described above subordinates the possibilities of Computer Vision to the goals and ways of working in art history discipline and museums, most of the relevant research in Computer Science does the opposite. The researchers take the goals of their field such as classification and prediction and apply them to a new type of data—large samples of visual user-generated content. So here cultural artifacts often are not approached in any different way than any other kind of data.

Both paradigms have their limitations. In this article, I take the outsider position. And this is why I start with the following question: How can we bring together assumptions and goals of two very areas of human knowledge which are fundamentally different, as opposed to subordinating one to another?

This article develops the following arguments: (1) Data representations of analog cultural artifacts used in Computer Vision, Music Information Retrieval, and Geospatial Computing give us a new and a better language for describing these artifacts in comparison to human natural languages; (2) These data representations are also closer to how human senses and central system encode analog signals. This provides another justification for the use of computer methods to analyze culture in general and using Computer Vision to see” visual culture in particular.

If Stork suggested that computers can analyze some aspects of art images better than human experts in some circumstances, I claim that computers are always more precise in their descriptions of characteristics of analog cultural artifacts. However, in the case of art historical images, the use of computation for analysis is one option because historical collections are small enough for us to study them directly. In the case of contemporary digital visual culture, using computer methods is the only way to see even small samples because of its scale. (Billions of images are shared every day on Facebook alone.)

My arguments presented in this article reflect my own practical experience of using Computer Vision with dozens of cultural datasets after I co-founded Cultural Analytics Lab at the University of California, San Diego (UCSD) in 2007. At that time, I defined “cultural analytics” as “the analysis of massive cultural data sets and flows using computational and visualization techniques” (Manovich 2007b). For about 10 years, our lab was the only one focusing on using

Computer Vision to study visual culture at large scale from the perspectives of humanities. To the best of my knowledge, the first such project done outside of our lab in the U.S. that received attention was only published in 2017 (Yale Digital Humanities Lab).

Cultural Analytics is only one among a number of research paradigms that emerged in the second part of the 2000s to take advantage of the availability of large cultural and social data. They include Digital Humanities, Computational Social Science, Social Computing, Digital Anthropology, Digital History, The Science of Cities, Urban Informatics, and Culturomics. In the same time, big cultural datasets started to be analyzed by computer scientists working in Machine Learning, Computer Vision, Natural Language Processing, Music Information Processing, Computer Multimedia, and also in Communication Studies. In the early 2010s, the “quantitative turn” begun in art history, and *International Journal for Digital Art History* was established in 2015. In 2020, the first large volume on digital art history was published:

*The Routledge Companion to Digital Humanities and Art History* (Brown 2020). In film studies, the first monograph that uses quantitative methods and data visualization to analyze works of a single film director appeared in 2019 (Heftberger 2019).

In parallel, the research in humanities using computational tools also started to grow. In 2003 it received the name Digital Humanities in 2003. In 2010s Digital Humanities kept growing, attracting more and more attention. However, the larger portion of the computational work in humanities so far focused on literary texts, historical text records and spatial data. In contrast, other types of media such as still and moving images and interactive media received relatively little attention. This situation is gradually improving but as I am writing this, analysis of visual media is still a small part of Digital Humanities (Digital Humanities Conference 2019). You can see this yourself by browsing programs of annual conferences organized by The Alliance of Digital Humanities Organizations or looking at the field journals that include *Digital Humanities Quarterly*, *International Journal of Digital Humanities*, and *Digital Scholarship in the Humanities*. The field limitations are well summarized by the title of the article published in 2017 in *Digital Scholarship in the Humanities*: “Digital humanities is text heavy, visualization light, and simulation poor” (Champion 2017).

This is surprising because computer scientists started to develop methods for the analysis of images already at the end of 1950s. Today they are implemented in numerous digital services and devices, including web image search engines, stand-alone photo cameras and cameras in mobile phones, widely used image editing software such as Photoshop, Pixelmator, Affinity Photo, and Luminar, image sharing services such Google Photos, and also available as

programming libraries (OpenCV, MATLAB). In Computer Vision and Multimedia Computing, researchers have been publishing for many years new algorithms for automatic detection of image content, artistic styles, photographic techniques, user-generated and professional video and TV programs, and photos that are more interesting, memorable, or original than others, and applying these algorithms to progressively larger datasets (Redi et al. 2017). In our lab we have been using some of these methods to analyze many types of both historical and contemporary visual media—20,000 photographs from the collection in Museum of Modern Art (MoMA) in New York, films by the pioneer of documentary filmmaking Dziga Vertov from Austrian Film Museum, sixteen million images shared on Instagram in seventeen global cities, one million Manga pages, one million artworks from popular art network DeviantArt, and other datasets.

### 3 Describing images with words and numbers

Most representations of physical, biological and cultural phenomena constructed by artists, scholars and engineers so far only capture some characteristics of these phenomena. Linear perspective represents the world as seen from a human-like viewpoint, but it distorts the real proportions and positions of objects in space. Contemporary 100-megapixel photograph made with a professional camera captures details of human skin and separate hairs—but not what is inside the body under the skin.

If the artifacts are synthetic, sometimes it is easy to represent them more precisely. Engineering drawings, algorithms, manufacturing details used to construct such artifact are already their representation in the finished state—however, we can’t predict human sensations and experiences of these artifacts only from these representations. But nature’s engineering can be so complex that even all representational technologies at our disposal can barely capture a minuscule proportion of information. For example, currently best fMRI machines can capture the brain at a resolution of 1 mm. This may look like a small enough area—yet it contains millions of neurons and tens of billions of synapses. The most detailed map of the universe produced in 2018 by Gaia (the European Space Agency craft) shows 1.7 billion stars—but according to estimates, our own galaxy alone contains hundreds of billions of stars.

And even when we consider a single cultural artifact created by humans and existing on a human scale—a photograph you took, a mobile phone you used to take it with, or your outfit consisting from items you purchased at Zara or COS—data representations of these artifacts often can only capture some of their characteristics. In the case of a digital

photograph, we have access to all the pixels it contains. This artifact consists from 100% machine data. These pixels to us will look a bit different from one display to the next, depending on its brightness, contrast, and color temperature settings, and its technology. And if we want to edit this image, what is possible is defined by particular software. (In my article *There is Only Software* (Manovich 2009) I argued that “depending on the software I am using, the “properties” of a media object can change dramatically. Exactly the same file with the same contents can take on a variety of identities depending on the software being used.”)

Digital pixel image is a synthetic artifact fully defined by only one type of data in a format ready for machine processing (e.g., an array of numbers defining pixel values). But what about physical artifacts, such as fashion designs that may use fabrics with all kinds of non-standard finishes, combine multiple materials, textures, and fabrics, and create unusual volumes? (This applies to many collections produced since the early 1990s the 1980s by Rei Kawakubo, Dres Van Noten, Maison Margiela, Raf Simons, Issey Miyake, among others, and also to many fashion designers working in countries such as South Korea today.) How do we translate clothes into data? The geometries of pattern pieces will not tell us about visual impressions of their clothes, or experience wearing them. Such garments may have unique two-dimensional and three-dimensional textures, use ornament, play with degrees of transparency, etc. And many fashion designs are only fully “realized” than you wear them, with the garment taking on particular shape and volume as you walk.

The challenge of representing the experience of material artifacts as data is not unlike calculating an average for a set of numbers. While we can always mechanically calculate an average, this average does not capture the shape of their distribution, and sometimes it is simply meaningless (Desrosières 1998). In a Gaussian distribution, most data lie close to the average, but in a binomial distribution, most data are away from it, so the average does not tell us much.

Similarly, when we try to capture our sensorial, cognitive and emotional experience of looking at or wearing a fashion garment, all methods we have available—recording heart-beat, eyes movements, brain activity, and other physiological, cognitive and affective processes, or asking a person to describe her subjective experience and fill out a questionnaire—can only represent some aspects of this experience.

But this does not mean that any data encoding automatically loses information, or that our intellectual machines (i.e., digital computers) are by default inferior to human machines, i.e. our senses and cognition. For example, let’s say I am writing about artworks exhibited in a large art fair that features hundreds of works shown by hundreds of galleries across a large space.

What I can say depends on what I was able to see during my visit and what I remembered—and therefore constrained

by the limitations of my senses, cognition, memory, and body, as well as by the language (Russian, Spanish, Indonesian, etc.) in which I write.

In the twentieth century, modern humanities, the common method of describing artifacts and experiences was to observe one own reaction as filtered by one’s academic training and use natural language for describing and theorizing these experiences. In social sciences and practical fields concerned with measuring people attitudes, taste and opinions, researchers used questioners, group observations and ethnography, and these methods remain very valuable today. Meanwhile, since the 1940s engineers and scientists working with digital computers have been gradually developing a very different paradigm—describing media artifacts such as text, shapes, audio, and images via numerical features. Humanities studies of visual art, architecture, design, video games, films, user-generated video and all other visual forms can adopt the same paradigm. Why it is such a good idea? My explanation is summarized in the next paragraph.

*Numerical measurements of cultural artifacts, interactions and behaviors give us a new language to talk about cultural artifacts and experiences. This language is closer to how the senses represent analog information* (sounds, music, colors, spatial forms, movement, etc.) The senses translate their inputs into quantitative scales, and this is what allows us to differentiate between many more sounds, colors, movements, shapes, textures than natural languages. So, when we represent analog characteristics of artifacts, interactions and behaviors as data using numbers, we get the same advantages. This is why a language of numbers is a better fit than human languages for describing analog aspects of culture.

Using natural languages was the only mechanism humanities have been using for describing all aspects of culture until the recent emergence of Digital Humanities. *Natural or ordinary language* refers to a language that evolved in human evolution without planning. While the origins of natural languages are debated by sciences, many suggest that it developed somewhere between 200,000 and 50,000 years ago. Natural languages cannot represent small differences on analog dimensions which define aesthetic artifacts and experiences such as color, texture, transparency, types of surfaces and finishes, visual and temporal rhythms, movement, speed, touch, sound, taste, etc. In contrast, our senses capture such differences quite well.

Aesthetic artifacts and experiences human species were creating during many thousands of years of their cultural history exploit these abilities. In the modern period, the arts started to systematically develop new aesthetics that strives to fill every possible “cell” of a large multi-dimensional space of all sense dimensions, taking advantage of the very high fidelity and resolution of our senses. Dance innovators from Loie Fuller and Martha Graham to Pina Bausch, William Forsythe, and Cloud Gate group defined new body

movements, body positions, compositions and dynamics created by groups of dancers or by parts of a body such as fingers or speeds and types of transitions. Such dance systems are only possible because our eye and brain abilities to register tiny differences on these dimensions of dance.

In visual arts, many modern painters developed lots of variations of a *white on white* monochrome painting—images that feature only one field of a single color, or a few shapes in the same color that differ only slightly in brightness, saturation, or texture. They include Kazimir Malevich (*Suprematist Composition: White on White*, 1918), Ad Reinhardt (“black paintings”), Agnes Martin, Brice Marden, Lucio Fontana, Ives Klein, and many others.

In the twenty-first century, works by contemporary product designers often continue the explorations that preoccupied so many twentieth century artists. For example, in the second part of 2010s top companies making phones—Huawei, Xiaomi, Samsung, Apple—became obsessed with the sensory effects of their designs. The designers of phones started to develop unique surface materials, unique colors, levels of glossiness of a finish, surface roughness and waviness. As the phone moves closer and closer to becoming a pure screen or transparent surface, this obsession with sensualizing still remaining material part may be the last stage of phone design before the phone becomes complete screen—although we may also get different form factors in the future, where small material parts become even more aestheticized (Manovich 2007c).

For instance, for its P20 phone (2017) Huawei created unique finishes each combining a range of colors. Huawei named them Morpho Aurora, Pearl White, Twilight and Pink Gold. When looking at the back of a phone at different angles, different colors would appear. (Peckham 2018). The company proudly described the technologies used to create these finishes on its website: “The Twilight and Midnight Blue HUAWEI P20 has a high-gloss finish made via a ‘high-hardness’ vacuum protective coating and nano-vacuum optical gradient coating.” (Huawei 2019) (The P30 Mate Pro I have been using during 2019 had one of these screens.)

What about minimalism that has become the most frequently used aesthetics in the design of spaces in the early twenty-first century exemplified by all-white or raw concrete spaces, with black elements or other contrasting details? From the moment such spaces started to appear in the West in the second part of the 1990s, I have been seeking them so I can work there—hotel areas, cafes, lounges. Today you can find it everywhere from but in the late 1990s they were just a few such spaces. In my book *The Language of New Media* (Manovich 2001) completed in Fall 1999, I have thanked two such hotels because large parts of that book were written in their spaces—The Standard and Mondrian in Los Angeles. While not strictly minimalist in a classical way (they were not all white), the careful choice of textures, materials, and

elimination of unnecessary details was certainly minimalist in its thinking. (Later in 2006–2007 I have been spending summers in Shanghai working on a new book and moving between a few large minimalist cafes—at that point, Shanghai had more of them than Los Angeles. Today, a city like Seoul probably has over 100,000 such cafes, each unique in its design.)

On first thought, such spatial minimalism seems to be about overwhelming our perception—asking us to stretch our limits, so to speak, to take in simultaneously black and white, big and tiny, irregular and smooth. I am thinking of famous Japanese rock gardens in Kyoto (created between 1450 and 1500), an example of *kare-sansui* (“dry landscape”): large black rocks placed in the space of tiny grey pebbles. In 1996 a store for Calvin Klein designed by London architect John Pawson opened in New York on Madison Avenue around the 60th Street, and it became very influential in the minimalist movement. Pawson was influenced by Japanese Zen Buddhism, and an article in the New York Times called his store “Less is Less.” (Goldberger 1996). The photographs of the store show a large open white space with contrasting with dark wood benches (Pawson 2020). So what is going on with these examples?

I think that minimalist design uses both sensory extremes for aesthetic and spatial effect, and small subtle differences that are our senses are so good at registering. The strong contrast between black and white or smooth and textured, or wood and concrete, and so on helps us to better notice the variations in the latter—i.e., the differences in shapes of tiny pebbles in Kyoto Garden, or all white parts of the 1996 Calvin Klein store space which all have different orientations to the light coming from very large windows.

The famous early twenty-first century examples of minimalist design are all white and or silver-grey Apple products designed by Jonathan Ive in the 2000s. The first in this series was iPod in 2001, followed by PowerBook G4 in 2003, iMac G5 in 2004, and iPhone in 2007. In his article “How Steve Jobs’ Love of Simplicity Fueled A Design Revolution,” Walter Isaacson quotes Jobs talking about his Zen influence: “I have always found Buddhism—Japanese Zen Buddhism in particular—to be aesthetically sublime,” he told me. “The most sublime thing I’ve ever seen are the gardens around Kyoto” (Isaacson 2012). In the most famous Kyoto garden, which I was lucky to visit, the monochrome surface made from small pebbles contrasts with a few large black rocks. In Apple products of the 2000s, the contrast between all-white object and the dark almost black screen when the device is turned off made from different material works similarly. It makes us more attentive to the roundness of the corners, the shadows from the keys, and other graduations and variations in tone and shape of the device.

In general, minimalism is everything but minimal. It would be more precise to call it “maximalism.” It takes

small areas on sensory scales and expands it. It makes you see that between two grey values there are, in fact, many more variations than you knew (I call this aesthetics common today in Korea “50 shades of grey”); that the light can fall on a raw concrete surface in endless ways; that the edge in the textured paper cut into two parts by hand contains fascinating lines, volumes, and densities. Our senses delight in these discoveries. And this is likely to be one of the key functions of aesthetics in human cultures from prehistory to today—giving our sense endless exercises to register some small differences, as well as bold contrasts. And to clean visual, spatial and sound environment from everything else, so we can attend to these differences. To enjoy “less is less.”

#### 4 Human senses, numbers and the arts

For thousands of years, art and design have thrived on human abilities to discriminate between very small differences on analog dimensions of artifacts and performances, and to derive both pleasure and meaning from this. But natural languages do not contain mechanisms to represent such nuances and differences. Why? Here is my hypothesis for why this is the case.

Natural languages emerged much later in evolution than the senses—to compensate for what the latter cannot do—represent the experience of the world as categories. In other words, *human senses and natural languages are complementary systems*. Senses allow us to register tiny differences in the environment, as well as nuances of human expressions (face expressions, body movements, etc.), while languages allow us to place what we perceive into categories, to reason about these categories and communicate using them.

Evolution had no reason to duplicate the already available functions, and that is why each system is great at one thing and very poor at another. The senses developed and continued to evolve for billions of years—for instance, the first eyes developed around 500 million years ago during the Cambrian Explosion. In comparison, the rise of human languages with their categorization capacities is a very recent development (sometime between 200,000 and 50,000 years ago).

When we use a natural language as a *metalanguage* to describe and reason about an analog cultural experience, we are doing something strange: *forcing it into small number of categories which were not designed to describe it*. In fact, if we can accurately and exhaustively “put into words” an aesthetic experience, it is likely that this experience is an inferior one. In contrast, using numerical features instead of linguistic categories allows us to much better aspects of an analog experience.

Our sensors and digital computers can measure analog values with even greater precision than our senses. You

may not be able to perceive a 1% difference in brightness between two image areas or 1% difference in the degree of smile between two photos of people, but computers are able to measure these differences. For example, for *Selficity* we used online computer vision service that measured the degree of smile in each photo on 0–100 scale. I doubt that you will be able to differentiate between smiles on such a fine scale.

Consider another example—representation of colors. In the 1990s and 2000s, digital images often used 24 bits for each pixel. In such format, each pixel can encode grayscale using 0–255 scale. This representation supports 16 million different colors—while human eyes can only discriminate between approximately 10 million colors. As I am writing this, many imaging systems and image editing software use 30, 36 or 48 bits per pixel. With 30 bits per pixel, more than 1 billion different colors can be encoded. Such precision means that if we want to compare color palettes of different painters, cinematographers, or fashion designers using digital images of their works, we can calculate it with more than sufficient accuracy. Certainly, this precision goes well beyond what we can do with small number of terms for colors available in natural languages (Gibson and Conway 2017). Certainly, some natural languages have more terms for different colors than other languages, but no language can represent as many colors as digital image formats.

In summary, a data representation of a cultural artifact or experience that uses numerical values or features computed from these values can capture analog dimensions of artifacts and experiences with more precision than a linguistic description. However, remember that a natural language also has many additional representation devices besides single words and their combinations. They include the use of metaphors, rhythm, intonation, stream of consciousness and other strategies that allow us to describe experiences, perceptions and psychological states in ways that single words and phrase can't. So, while natural languages are categorical systems, they also offer rich tools to go beyond the categories. Throughout human history poets, writers, and performers using speech (and best hip-hop and spoken word artists) today create exceptional works by employing these tools.

Not everybody can invent great metaphors. Numerical features allow us to measure analog properties of the scale of arbitrary precision and do this automatically at scale using computers. But this does not mean that data representations of aesthetic artifacts, processes, and performances that use numbers can easily capture everything that matters.

In the beginning of the twentieth century, modern art rejected figuration and narration, and decided instead to focus on the sensorial communication—what Marcel Duchamp referred to as “retinal art.” But over the course of the twentieth century, as more possibilities were fully explored and became new conventions, *artists started to*

create works that are harder and harder to describe using any external code, be it language or data. For example, today we can easily represent flat geometric abstractions of Sonia Delaunay, František Kupka, and Kasimir Malevich as data about shapes and colors and sizes of paintings and drawings; and we can even encode details of every visible brushstroke in these paintings. (Computer scientists have published many papers that describe algorithmic methods to authenticate the authorship of paintings by analyzing their brushstrokes.) But this becomes more difficult with new types of art made in the 1960s–1970s: light installations by James Turrell, acrylic 3D shapes by Robert Irvin, “earth-body” performances by Ana Mendieta, happenings by Alan Kaprow (to mention only most canonical examples), as well as works of thousands of other artists in other countries, such as Движение art movement in USSR. Their works included *Cybertheatre* staged in 1967 and described in their article published in *Leonardo* journal (Nusberg 1969). The only actors in this theatre performance were 15–18 working models of cybernetic devices (referred as “cybers”) capable of making complex movements, changing their interior lighting, making sounds, and omitting color smoke. For something less technological, consider *Imponderabilia* by Marina Abramović and Ulay (1977): for 1 h, the members of public were invited to pass through the narrow “door” made by naked bodies of the two performers.

The experience of watching documentation left after an art performance is different from being present at this performance; and what can we measure if an artwork is designed to deteriorate over time or quickly self-destructs like Jean Tinguely’s “Homage to New York” (1960)? Similarly, while the first abstract films by Viking Eggeling, Hans Richter, and May Ray made in the early 1920s can be captured as numerical data as easily as geometric abstract paintings by adding time information, how do we represent Andy Warhol’s *Empire* (1964) that contains a single view of the Empire State Building projected for 8 h? We certainly can encode information about every frame of a film, but what is crucial is the physical duration of the film, its difference from the actual time during shooting, and very gradual changes in the building appearance during this time. The film was recorded at 24 frames per second, and projected at 16 frames per second, thus turning physical 6.5 h into 8 h and 5 min of screen time. (Very few viewers were able to watch it from beginning to end, and Andy Warhol refused to show it in any other way.)

## 5 Conclusion

In this article, I have argued that the use of numerical representation and data analysis and visualization methods offers a new language for describing cultural artifacts, experiences

and dynamics. The human languages such as English or Russian that developed rather recently in human evolution are not good at capturing analog properties of human sensorial and cultural experiences. These limitations become particularly worrying if we want to compare thousands, millions or billions of artifacts—i.e. to study contemporary media and cultures at their new twenty-first century scale. When we instead use numbers, numerical summaries such as Computer Vision features and also data visualization, we can better capture small differences between a few or very many artifacts. The methods of finding structures and relationships in large numerical datasets developed in statistics and machine learning such as cluster analysis, dimension reduction, and other fields such as network science allow us to extend the analysis to very big datasets of cultural artifacts. Equally importantly, numbers, features and data visualization also give us a language to represent gradual and continuous temporal changes—something which natural languages are also bad at.

Having a better language to describe the analog dimensions of visual culture including single images, video, or a dance performance is invaluable. Digital computers that work on numerical representations are better at capture many dimensions which natural languages can’t describe in enough detail, such as motion or rhythm. We can now describe the characteristics of cultural processes which are hard to capture linguistically—for example, gradual historical changes in any visual culture over long periods, changes in visual form over the career of an artist, changes in cinematography over the course of a feature film or a music video.

And this is what the phrase “language of art” in the title of this article refers to. In the twentieth century, many artists, filmmakers, architects and theorists—especially within semiotics paradigm—were proposing that different arts and culture areas have their own languages comparable to human natural languages (Barthes 1997). In my view, these explorations did not reach satisfactory results partly because these theories were using natural languages to try to describe analog dimensions of art and culture. And as I argued here, such an attempt is inherently problematic.

I don’t want to argue for or against the idea that painting, fashion, food or space design communicate like languages. In fact, works by Goodman (1968), Sonesson (1989) and by other theorists developed more precise and productive concepts and theories that describe about the differences between languages and various art and cultural forms. What I did claim here is that now we can use digital computers to capture analog dimensions of artifacts and our aesthetic experiences as numbers. This numbers can use continuous scales that allows us to capture tiny differences between artifacts and details of artifacts with as much precision as we want. And we do can this for arbitrary large numbers of artistic and cultural artifacts.

In other words, we now possess a new language for describing and talking about art and culture. In my view, this is very important because being able to describe any phenomenon more precisely than we could earlier is the first step for expanding our knowledge in any domain.

**Funding** No funding supported writing this article.

## Compliance with ethical standards

**Conflicts of interest** All authors declare that they have no conflict of interest.

## References

- Barthes R (1997) *Elements of semiology*. Hill and Wang, New York. Originally published in France in 1962
- Brown K (2020) *The Routledge companion to digital humanities and art history*. Routledge, London
- Champion E (2017) Digital humanities is text heavy, visualization light, and simulation poor. *Digital Scholarship Humanities* 32, issue supplement 1: 25–32. [https://academic.oup.com/dsh/article/32/suppl\\_1/i25/2957402](https://academic.oup.com/dsh/article/32/suppl_1/i25/2957402). Accessed 1 July 2020
- Derech N, Tal A, Shimshoni I (2018) Solving archeological puzzles. <https://arxiv.org/pdf/1812.10553.pdf>. Accessed 1 July 2020
- Desrosières A (1998) *The politics of large numbers: a history of statistical reading*. Harvard University Press, Cambridge
- Digital Humanities Conference (2019) <https://dh2019.adho.org>. Accessed 1 July 2020
- Gibson T, Conway BR (2017) The world has millions of colors. Why do we only name a few? *Smithsonian Magazine*. <https://www.smithsonianmag.com/science-nature/why-different-languages-name-different-colors-180964945/>. Accessed 1 July 2020
- Goldberger P (1996) On Madison avenue, sometimes less is less. *The New York Times* October 27, 1996
- Goodman, N (1968) *Languages of art: an approach to a theory of symbols*. Bobbs-Merrill, Indianapolis
- Hefberger A (2019) *Digital humanities and film studies*. Springer, Berlin
- Huawei (2019) Huawei P20. [consumer.huawei.com. http://consumer.huawei.com/en/phones/p20/](http://consumer.huawei.com/en/phones/p20/). Accessed 1 July 2020
- Impett L, Moretti F (2017) Totentanz. Operationalizing Aby Warburg's Pathosformeln. *Stanford Literary Lab*. <https://litlab.stanford.edu/LiteraryLabPamphlet16.pdf>. Accessed 1 July 2020
- Isaacson W (2012) How Steve jobs' love of simplicity fueled a design revolution. *Smithsonian Magazine*, September 24, 2012. <http://www.smithsonianmag.com/arts-culture/how-steve-jobs-love-of-simplicity-fueled-a-design-revolution-23868877/>. Accessed 1 July 2020
- Manovich L (2001) *The language of new media*. The MIT Press, Cambridge
- Manovich L (2007b) Cultural analytics: about. *Software Studies Lab*. <http://lab.softwarestudies.com/p/overview-slides-and-video-articles-why.html>. Accessed 1 July 2020
- Manovich L (2007c) Information as an aesthetic event. *Receiver*, n.p. <http://manovich.net/index.php/projects/information-as-an-aesthetic-event>. Accessed 1 July 2020
- Manovich L (2009) There is only software. In: Lee Y, Henk Slager H (eds) *Nam June Paik reader—contributions to an artistic anthropology*. NJP Art Center, Yongin, pp 26–29
- Redi M, Liu FZ, O'Hare NK (2017) Bridging the aesthetic gap: the wild beauty of web imagery. In: *ICMR'17: proceedings of the 2017 ACM international conference on multimedia retrieval*. ACM, New York, pp 242–250
- Nusberg L (1969) *Cybertheater*. *Leonardo* 2: 61–62. [http://monoskop.org/images/a/af/Nusberg\\_Lev\\_1969\\_Cybertheater.pdf](http://monoskop.org/images/a/af/Nusberg_Lev_1969_Cybertheater.pdf). Accessed 1 July 2020
- Pawson J (2020) Calvin Klein Collections Store. <http://www.johnpawson.com/works/calvin-klein-collections-store>. Accessed 1 July 2020
- Peckham J (2018) Huawei P20 and P20 pro colors: what shade should you buy? *Techradar*. <http://www.techradar.com/news/huawei-p20-and-p20-pro-colors-what-shade-should-you-buy>. Accessed 1 July 2020
- Sonesson G (1989) *Pictorial concepts: inquiries into the semiotic heritage and its relevance to the interpretation of the visual world*. Lund University Press, Lund
- Stork D (2009) Computer vision and computer graphics analysis of paintings and drawings: an introduction to the literature. In: Xiaoyi J, Nicolai P (eds) *CAIP'09: proceedings of the 13th international conference on computer analysis of images and patterns*. Springer, Berlin, pp 9–24
- VISART IV (2018) 4th workshop on computer vision for art analysis, 9th September 2018, Munich, Germany. <https://visarts.eu/past-workshops/2018>. Accessed 1 July 2020
- Yale Digital Humanities Lab (2017) Yale DHLab—robots reading vogue. <http://dhlab.yale.edu/projects/vogue/>. Accessed 1 July 2020

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# On machine vision and photographic imagination

Daniel Chávez Heras<sup>1</sup> · Tobias Blanke<sup>2</sup>

Received: 1 May 2020 / Accepted: 14 October 2020 / Published online: 17 November 2020  
© The Author(s) 2020

## Abstract

In this article we introduce the concept of *implied optical perspective* in deep learning computer vision systems. Taking the BBC's experimental television programme “Made by Machine: When AI met the Archive” (2018) as a case study, we trace a conceptual and material link between the system used to automatically “watch” the television archive and a specific type of photographic practice. From a computational aesthetics perspective, we show how deep learning machine vision relies on photography, its technical regimes and epistemic advantages, and we propose a novel way to identify the *latent camera* through which the BBC archive was seen by machine.

**Keywords** Computational aesthetics · Philosophy of photography · AI television · Computer vision · Deep learning · Dataset archaeology

## 1 Introduction

Is that a person or a reflection? A man or a woman? Is the woman holding a mobile phone, or is it, rather, the statue of an ancient Egyptian king? Is the man wearing a shirt, or is it an elephant? Or a stuffed animal holding a banana...? (Figs. 1, 2, 3).

These are some of the mislabellings produced when a small team of technologists and researchers set a computer vision system to “watch” thousands of hours of British television for the project “Made by Machine: When AI met the Archive” (MbM), whose outputs were eventually packaged and broadcast on BBC Four as an experimental “AI TV” programme in 2018.<sup>1</sup>

In line with the public purposes of the British broadcaster (BBC 2018), one of the main goals of the programme was to show to a wider audience some of the possibilities and limitations of AI, and in particular deep learning approaches that underlie many contemporary computer vision systems. From a research perspective, the project was also designed as prompt to explore just how exactly computers are said to be “seeing”. What type of knowledge is produced by computer

vision and how does it inform the ways we understand and give currency to audio–visual media more generally?

In related work, such questions have generally been approached by focussing on training datasets and how they are assembled as well as how the resulting AI systems represent or fail to represent different sectors of society. Exemplary of this approach are the works of Kate Crawford and Adam Harvey:

“Training sets, then, are the foundation on which contemporary machine-learning systems are built. They are central to how AI systems recognize and interpret the world. These datasets shape the epistemic boundaries governing how AI systems operate, and thus are an essential part of understanding socially significant questions about AI.” (Crawford and Paglen 2019)

“A photo is no longer just a photo when it can also be surveillance training data, and datasets can no longer be separated from the development of software when software is now built with data.” (Harvey and LaPlace 2019)

Research using this approach has shown datasets to be deeply problematic, both in their politics of assemblage and of representation. The experience in MbM was no exception here. A team from BBC R&D implemented an automated dense captioning system pre-trained on the Visual

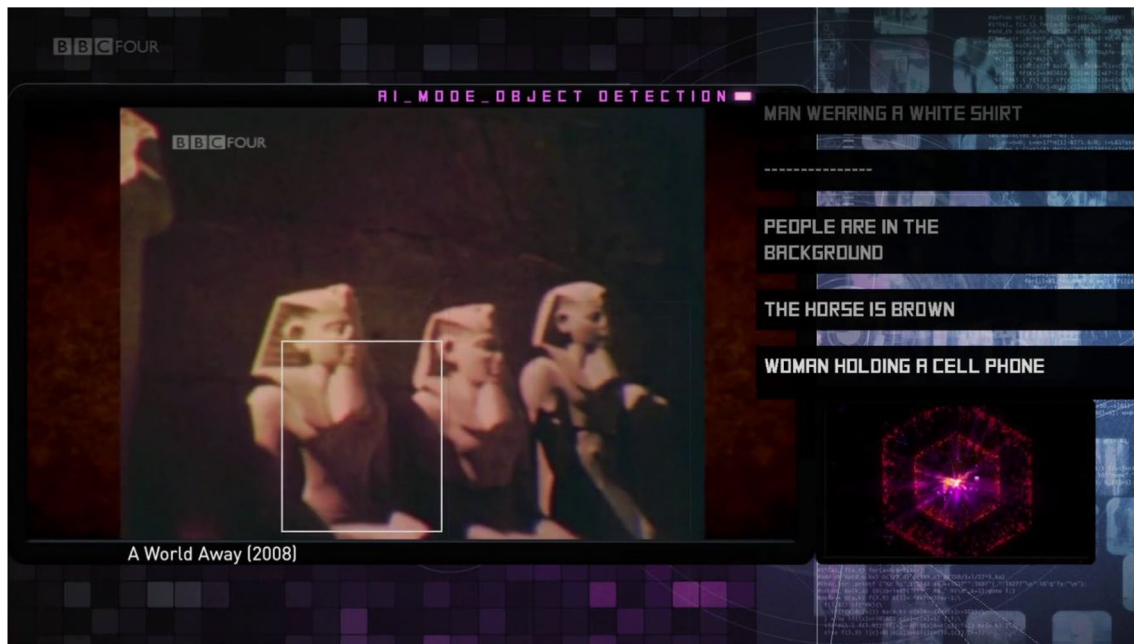
✉ Daniel Chávez Heras  
daniel.chavez@kcl.ac.uk

<sup>1</sup> Department of Digital Humanities, King's College London, London, UK

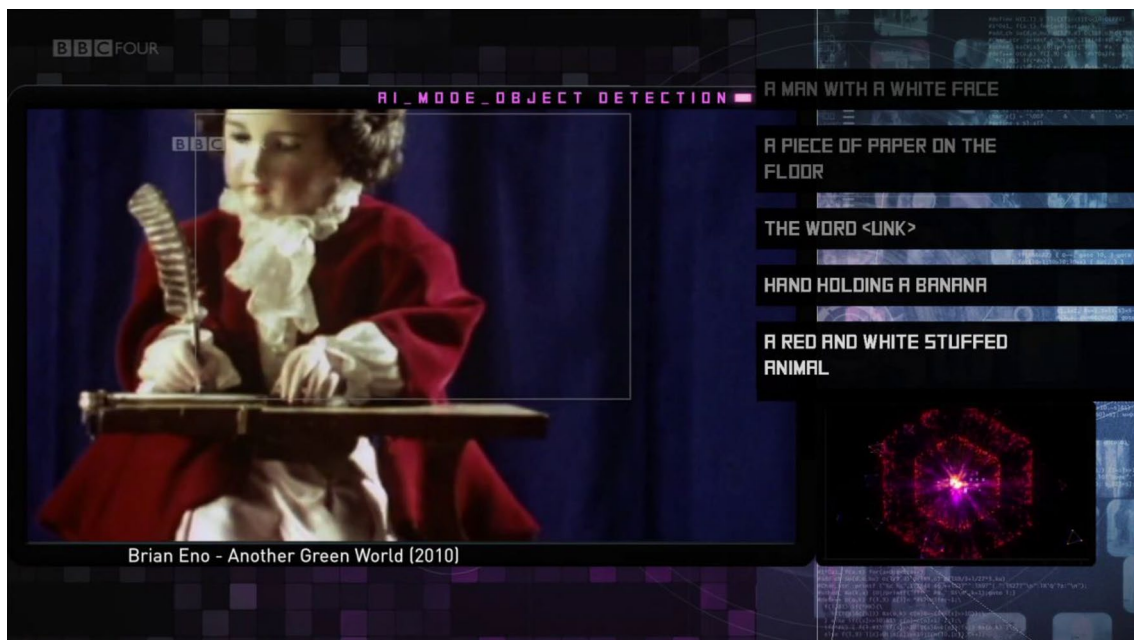
<sup>2</sup> University of Amsterdam, Amsterdam, The Netherlands

<sup>1</sup> <https://www.bbc.co.uk/programmes/b0bhwk3p>





**Fig. 1** MbM still frame with predicted label ‘woman holding a cell phone’

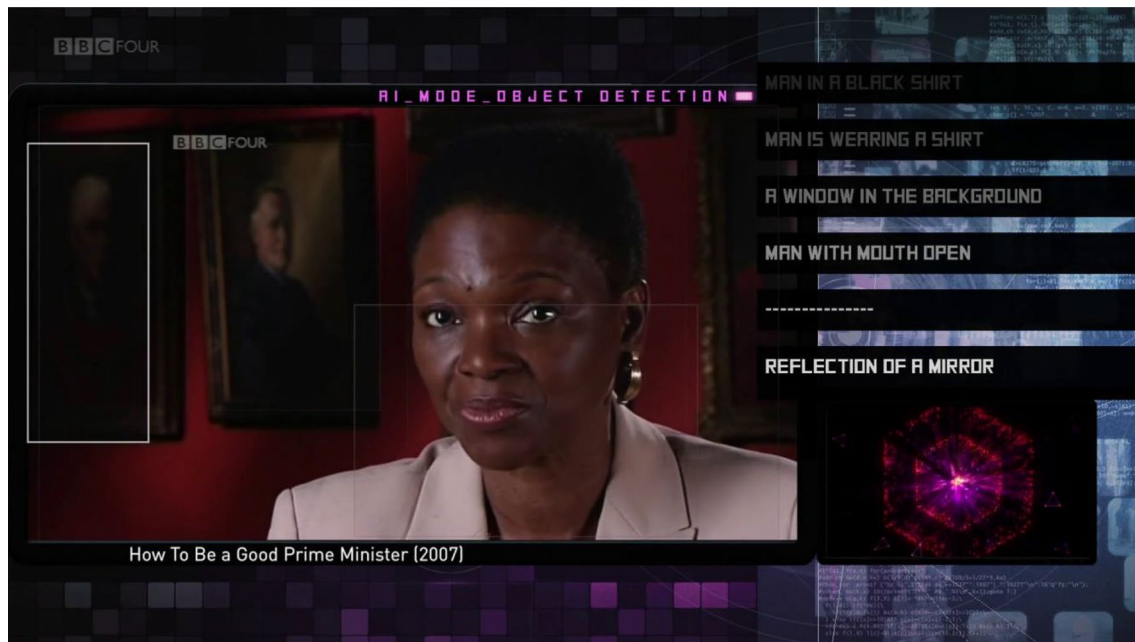


**Fig. 2** MbM still frame with predicted label ‘red and white stuffed animal’

Genome dataset<sup>2</sup>; comprised of little over 108,000 images downloaded from Flickr and annotated by 33,000 Amazon Mechanical Turk workers, 61% of whom were under 35 and 93% from the USA (Krishna et al. 2017, 43). This captioning system implementation was used to automatically annotate

several hundred hours of television from the BBC archive, and these annotations used as metadata in an associational engine that concatenated new sequences of related television clips (Cowlshaw 2018; Chávez Heras et al. 2019).

<sup>2</sup> <https://visualgenome.org/>



**Fig. 3** MbM still frame with predicted label ‘reflection of a mirror’

One of the results immediately observed in these machine-generated clips was the system’s propensity to identify men in shirts. Although imbalance in gender representation is a known issue in television, including the BBC (Cumberbatch et al. 2018), this effect was so pronounced that it prompted a closer inspection of the training dataset. “white” is the most common attribute, “man” most common object (with twice as many instances as “woman”) with “shirt” also among the top ten (Krishna et al. 2017, 50–53, 63). Biases such as these have been consistently found in this and other large image training sets used in deep learning computer vision, the localisation and disaggregation of such biases being one of the main goals of Crawford’s proposed archaeology of datasets (Crawford and Paglen 2019).

However, by focusing on labelling images as the main layer of human intervention, i.e. the principal source of data, bias and error, such an archaeology very often overlooks other significant areas of human subjectivity encoded in these systems, namely the nature of the images themselves. These labels are produced not over direct observations of the world, but over photographic images of it; images that are technically and socially mediated in powerful ways that create and sustain specific regimes of visibility and with which we hold a complex relationship before they become digital artefacts.

Following this line of thought, we set out to look at computer vision through the (figurative and literal) lens of photography. Through a technical genealogy of photographic lenses, the types of images they afford and the social functions given to these images, we show how AI is materially

and conceptually connected to optical regimes of visibility. Based on this connection, in the last section we assemble a dataset with which to analyse photographic practice, and then use it to train a bespoke focal length classifier as a proof-of-concept for a system designed to investigate the optical perspectives implied in MbM.

## 2 Epistemic discontents

An often unexamined fact about of deep-learning computer vision is that the millions of pictures it algorithmically mobilises are, for the most part, photographs. Perhaps one of the most revealing accounts about how this choice came to be seen as obvious comes from Fei-Fei Li, one of the creators of *ImageNet*<sup>3</sup> and a key figure in the shift towards deep learning over the last decade. Li was asked recently about the choice of using photographs for *ImageNet* during an event celebrating the tenth anniversary of the dataset: “That’s a great question.” —she replied— “We didn’t really stop to think much about it (...). I suppose we wanted as a realistic representation of the world as possible” (Li 2019).

Li is not alone in her assumption about realism and photography. A widely shared intuition about photographic images is that “the camera does not lie” or that in any case it lies less than other methods of depiction. In an often-cited passage of his influential *Ontology of the Photographic*

<sup>3</sup> <https://www.image-net.org/>

*Image*, André Bazin (1960[1958]) wrote that the invention of photography and cinema “satisfy, once and for all and in its very essence, our obsession with realism. No matter how skillful the painter, his work was always in fee to an inescapable subjectivity. The fact that a human hand intervened cast a shadow of doubt over the image.” (7). What he meant exactly by “realism” has been the subject of much debate, since but this view of photographs as trusted visual renderings of the world due to their alleged automatic mode of production has proved remarkably enduring.

Li’s response earnestly voices just such a view, where unlike for example drawings or paintings, which are inextricably bound to the mental states and technical abilities of their authors (as well as the embodied command of these states and abilities), photographs appear to be pictures produced through mind-independent processes. They capture whatever is in front of the camera regardless of what the photographer believes about what is in front of the camera, i.e. cameras can only show what is there to be seen. In philosophy, this idea of mind-independence mechanical process has served to explain the epistemic advantage of photographs over other types of images (Cavell 1979; Cohen and Meskin 2004; Walden 2005; Abell 2010).

One formulation of this argument proposed by Gregory Currie (1999) is that we treat photographs as *traces* as opposed to *testimonies*. The former are counter-factually dependant on nature, like a footprint, in a way that the latter are not, like the tale of how someone once took a step in the mud. For the footprint to be any different, Currie would argue, the sole of their shoe would have had to differ accordingly, while a description of the step taken, however, rich or detailed, necessarily implicates the intentions of the describer and belongs, therefore, to a different epistemic register altogether. According to this view, the social credibility lent to photographs makes them more akin to *light detections* captured as a result of a mind-independent mechanical process, while paintings and other pictures made by hand tend to be seen as *someone’s depiction* of a scene, this is, as the result of an embodied cognitive and creative process.

This is the dominant logic that underwrites computer vision too, at least in its current form and insofar as it is powered by photographic images, from which it inherits, exploits and amplifies their epistemic advantage founded on this mind-independent conjecture about the photographic process. When millions of photographs are aggregated into large datasets and used to train machine learning systems, the representational powers of photography and computation compound, to the point, where predicted labels are also seen as traces, face *detection* and not face *depiction*. The predictive tokens produced by computer vision are thereby presented as the counter-factually-dependent *and* mind-independent detections of something or someone. This person

or this object was seen automatically and, therefore, *had* to be there to be seen.

Recently, however, this view has been put under mounting pressure by the so-called new theory of photography, whose proponents argue for an expanded view of the photographic event as a multi-staged process of which only some parts can be said to occur automatically (Maynard 1997; Phillips 2009; Lopes 2016; Costello 2017). Costello (2017), for example, identifies a contradiction in ascribing epistemic value to photographs on the basis of their supposed mind-independence processes while simultaneously characterising these processes as automatic. A process cannot be both natural and automatic, he argues, without separating humans from nature (42). Automatic processes are causally-dependant but not spontaneous according to Costello. For a process to be called automatic it should be possible to specify it in terms of the labour that is being delegated to a mechanism, one that serves human ends and, therefore, necessarily involves human minds. Costello asks:

“just what is it exactly that is supposed to ‘happen by itself’? [...] In photography, almost everything that expresses comparable choices [to painting] happens *off* the support—the choice of lens, distance, lighting, moment of exposure, point of view, etc.[...] This can give those who have no idea, where to look the impression that the photographer has done very little, or that the mechanism is responsible. But this is plain ignorance. The fact that so many of the acts take place prior to the image appearing ‘all at once’ does not negate the photographer’s responsibility for what then appears. Merely noting depth of field markers in an image already tell us much about what the photographer was after. One needs to be a competent judge in photography as in any other domain; and this requires a basic grasp of the internal relations between focus, depth of field, and exposure that most Orthodox theorists fail to evince.” (45) [italics in original].

From this perspective, photographs are seen as faithful visual representations by virtue of their mechanisms no less than by the ways in which such mechanisms are controlled and regulated by photographers. Following Costello, we need to also consider that our intuition of what is a “realistic” image rests in the case of photography as much on *what* it shows than on *how* it shows it, which is to say that the photographic image is granted its privileged epistemic position in society by adding, not subtracting layers of subjectivity; not “without the creative intervention of man”(7), as Bazin would have it, but precisely because of it.

If the depicted is indeed inextricable from the process of depiction, we must then ask why should we not care about this process when it comes to computer vision?

### 3 The glass computer

One of the reasons photographic cameras appear to record the world automatically is that many of the calculations needed to render space visible are pre-programmed in the photographic devices themselves, most significantly in photographic lenses. When photographers *pull* an image into focus by adjusting the focus ring, the lens is doing some of the heavy lifting in terms of the calculations necessary to harness light convergence and render a slice of space visible in a specific manner. This is not to say the lens itself thinks, but rather that thought has been put into the lens, quite literally crystallised in its design, and that the photographer is able to interface with it through the camera controls.

From a cognitive standpoint, like the Sumerian abacus or the medieval volvelle, photographic lenses can be thought of as a type of analogue computer: a system that allows its user to actively externalise memory to a programmable calculating object and in this way distribute the cognitive load required to perform a specific task. Configured in this manner, user and object enter into an interaction feedback loop, creating “a coupled system that can be seen as a cognitive system in its own right” (Clark and Chalmers 1998). That photographers need not perform optic calculations to mobilise their effects is one of the most salient affordances of photography as a technology, and connects it to computer vision in their shared overarching project to automate visual labour through the computation of space (Pasquinelli 2019), with the obvious difference that in the case of photographic lenses such computation is analogue.

If we have not traditionally thought of photographic lenses as machinery with which to calculate<sup>4</sup> is perhaps because their inputs and outputs are presented as images and not numbers or letters. We do not know, for example, whether an image is in focus if presented as a matrix of pixel values (or a tensor); we need to see the results as an image “all at once” to evaluate it. However, the intermediate steps of interaction involved in producing photographic images are in fact heavily mediated by numerical parameters and standardised metrics: focal length, exposure, aperture and ISO. These are all given as numbers that describe the internal relations of the photographic mechanism, and from this point of view a key aspect of photographic practice as an imaging technology is to understand, control and harness different permutations of these relations for a variety of

<sup>4</sup> However there are a lot of optical calculations crystallised in photographic lenses. One notable example that links optics with computing is how in 1840 Joseph Petzval, an Austrian mathematician, employed several human computers to aid in the design a new four-element lens capable of under-one-minute exposures: the famous *Petzval Portrait* (See: Peres 2007, 159).

lenses. An equivalent process in machine learning would be the understanding and control of hyper-parameters.

Take focal length as an example. Today, it is widely used as a standard measure for lens classification, since it correlates with the size of the image plane<sup>5</sup> and the aperture<sup>6</sup> of the camera to define, among other things, the field of view, i.e. how much of a given scene fits into the frame, and the depth of field, i.e. how much of it is in focus at any given time. For a full-frame format,<sup>7</sup> a wide angle lens (e.g. 28 mm) will cover a wider field of view and have a deeper focus range, while a telephoto (e.g. 300 mm) will magnify to a narrower area and have a shallower focus range. In between we find a 50 mm, often called a “normal” lens.

Over time, the effects produced by different focal distances get thematised and are attached to specific social narratives. Long telephotos tend to be used in sports and nature photography, where subjects are often moving at a distance and backgrounds can be out of focus. Wide lenses, on the other hand, privilege field of view and focused scenes instead of magnification. Depicted through a different lens the same subject can be made to look *in fraganti* in a leaked mobile phone picture (28 mm) or like a model fit for the cover of a fashion magazine (175 mm) (Wieczorek 2019).

Different lenses contribute in this way to our understanding of what pictures are *about*. Our argument is that it is precisely this “aboutness” of vision that we feel to be conspicuously absent or compromised in the tokens of prediction produced by systems like the one used to machine-see the BBC television archive, a type of computer vision which only points to what images are *of*.<sup>8</sup> Drawing from this distinction, we can clearly see how lens aesthetics are not incidental to photography but rather a fundamental dimension of its epistemic advantage insofar as they enable distinct relations between the see-able and the know-able; between knowledge and the appearance of knowledge. That images are seen to be *about* something inasmuch as *of* something suggests that we put our faith in photography not because it offers undistorted images of the world, but because we

<sup>5</sup> Usually given in millimetres as the diagonal measure of a rectangular projection surface or screen onto which an image is formed when reflected light is projected through a lens.

<sup>6</sup> Known as *f*-stop or, somewhat confusingly, *f*-number (*N*), calculated using the formula:  $N=f/D$  (where *f* is focal length and *D* the diameter of the iris or pupil that allows light into the lens).

<sup>7</sup> 35 mm (36 mm × 24 mm frame) film negative or equivalent digital sensor. Many digital cameras have smaller sensors, thereby modifying (cropping) the field of view of lenses. The smaller the sensor is in relation to a full-frame the larger its crop factor. Conversely, by knowing the crop factor of a given sensor, one can estimate a lens’ full-frame equivalent focal length. Mobile phones, for example, have much smaller sensors and lenses, an iPhone X has a 4 mm lens, 28 mm equivalent in a full frame camera.

<sup>8</sup> For a discussion on this distinction see (Maynard 1997).

believe that photographic distortions are meaningful. Computer vision, we argue, gains its powers by treating photographs not as detections of the world, but as measurements of these beliefs, and in doing so it assumes an implied optical perspective.

## 4 A machine made of images

Let us now return to MbM and ask what optical perspective is implied in it. What lens or lenses are encoded in the computational gaze we set upon the BBC Television archive?

We know the Visual Genome uses images originally sourced from *Flickr* (Krishna et al. 2017, 47) and that the photo platform hosts many of its images along with their EXIF data,<sup>9</sup> which is an international metadata standard for digital images and sound that includes tags for camera settings and lens information.

EXIF is far from perfect. Its metadata structure is borrowed from TIFF files and is now over 30 years. A notable drawback to working with this type of data is, therefore, its inconsistency, given the quick pace at which digital cameras changed over the last decades and the many differences in how they used the standard over time, even among cameras from the same manufacturer. What is more, some manufacturers like Nikon use custom format fields not common to any other brand and encrypt the metadata contained in them. This makes it very difficult to extract, disaggregate and process EXIF.<sup>10</sup> Finally, this type of metadata is not usually available for not-born-digital photographs, i.e. taken with analogue cameras or images that were scanned.<sup>11</sup>

These caveats notwithstanding, EXIF is still the most widely used metadata standard for photography and as such a key resource to research the equipment and technical practice that underlies the creation of photographic images in the digital age. And precisely because of its longevity and pervasive use, it is one of the few ways to trace a technical lineage from lenses to computer vision. It is quite possibly the only, where such a lineage can be done at a larger scale, given the size of the collections of images used in deep learning.

<sup>9</sup> Developed in 1998 by the Japan Electronic Industries Development Association (JEIDA), eventually absorbed by the Japan Electronics IT industries association (JEITA) and the Camera & Imaging products association (CIPA). (See: JEITA Standards).

<sup>10</sup> <https://exiftool.org/TagNames/Nikon.html#LensData01>

<sup>11</sup> A scanned photo, for example, will sometimes include EXIF data from the scanner, but obviously no lens or other information about the camera with which was originally taken. It is possible, however, to manually write or re-write EXIF tags of a digitised photograph (or nearly any other digital image for that matter). For the most authoritative source on working with EXIF see Phil Harvey's Exif Tool: <https://exiftool.org/>

**Table 1** Example of Exif tags extracted

Tag	Value
Camera manufacturer	Canon
Camera model	Canon EOS 7D
Exposure	1/1600
Aperture (F-number)	2.8
Focal length	145.0 mm
Lens info	0EF70-200 mm f/2.8L IS II USM

We extracted EXIF metadata from all the images whose Flickr IDs matched the ones present in the Visual Genome. The metadata standard is comprised of over twenty thousand tags, but we only selected tags that were general enough so as to be reported by most cameras. Within these, we only focused on the ones containing data about the parameters over which photographers tend to have more choice and control, namely their choice of camera and lens, as well as the aperture, exposure and focal length settings. Table 1 shows a list of the tags that were queried and an example of the values extracted. Table 2 shows an overview of the extraction results.

The extraction process yielded a relatively dense distribution, with over 83% of accessible images returning metadata in at least one of the five of the queried tags. The one exception was <Lens info>, for which only 10% of accessible images returned values. In light of this, we decided to consolidate data for all tags except <Lens info>, which was kept separately for later analysis. We also parsed over apertures and focal lengths to bin them into categories: twelve bins corresponding to full *f*-stops for apertures—from *f*1 to *f*45,<sup>12</sup> and seven focal distance bins corresponding to a commonly used classification<sup>13</sup>:

<sup>12</sup> There are wider and narrower apertures, for example *f*/0.95 of the Shenyang Zhongyi Mitakon Speedmaster 50 mm. However these are very rarely encountered and were not observed in our dataset.

<sup>13</sup> These categories are not policed or enforced by any particular institution, as the boundaries are seen as irrelevant in most areas of photographic practice. They are, rather, more of a tacit agreement among photographers, lens and camera manufacturers. Of the tags queried, focal distance was the most challenging because, as we noted earlier, these values are relative to the size of the sensor. The standard “full frame” sensor was adopted as an equivalent of 35 mm film stock, but as digital cameras shrunk in size so too did their sensors. The effect is particularly stark in mobile phones, whose sensors are particularly small, so in order to compare their focal length to that of larger cameras one needs to multiply their reported Exif value by a crop factor so as to obtain a 35 mm equivalent. This crop factor is different across models and manufacturers, for example many Apple iPhone models have a sensor crop factor of 7.6, if the focal length in their Exif metadata is 4.3, the 35 mm equivalent is a little over 30 mm. If one were to accurately measure focal length one

**Table 2** Overview of the extraction results

Category	Count	Percentage
Total number of IDs processed	103,077	100.00%
Unavailable URL request (500 error)	18,521	18.00%
Available image but with no data in the queried tags	443	0.40%
Available images with data in at least one queried tag	84,113	81.60%

**Table 3** First five observations of our working data frame, shaped 68,085 rows  $\times$  5 columns

Camera manufacturer	Camera model	Exposure	Aperture	Focal length
Canon	Canon PowerShot S2 IS	1/640	4.0	72.0 mm
Panasonic	DMC-FX9	1/13	3.6	9.9 mm
Canon	Canon EOS 20D	1/250	11.0	560.0 mm
Nikon	NIKON D50	1/250	5.0	125.0 mm
Canon	Canon PowerShot SD600	1/320	2.8	5.8 mm
...	...	...	...	...

- Ultra wide (< 24 mm)
- Wide (24–35 mm)
- Normal (35–85 mm)
- Short telephoto (85–135 mm)
- Medium telephoto (135–300 mm)
- Super telephoto (+ 300 mm)

We also parsed over exposures to remove faux entries (e.g. a small number of older mobile phones reported infinite or zero values for exposure), and manually matched some camera manufacturers names (e.g. ‘NIKON’ and ‘Nikon Corporation’). The consolidated data frame includes all values in all remaining tags for a total 68,085 entries, which is

66% of all images that comprise the Visual Genome (v1.2). An example of our working data frame is shown in Table 3.

Our analysis of EXIF data shows the clear dominance of DSLR over other types of equipment, with Canon and Nikon being the two major manufacturers combining for over 64% of all cameras, more than eight times the share of the third largest manufacturer, Sony, at 8% (Fig. 4).

From these, the ten most popular camera models all correspond to Canon EOS and Nikon DX systems, with the only exception of the Apple iPhone 4, at number nine. The most common camera in our dataset is Nikon’s D90, an entry-level DSLR released in 2008, and the first model with video-recording capabilities. The second most popular is the semi-professional Canon 5D Mark II, released the same year, closely followed by the 7D also from Canon, released in 2009.

In terms of how these cameras were used, our analysis identifies large apertures  $f$  2.8, 4, and 5.6 as the most popular, accounting together for 74% of photographs (Fig. 5).

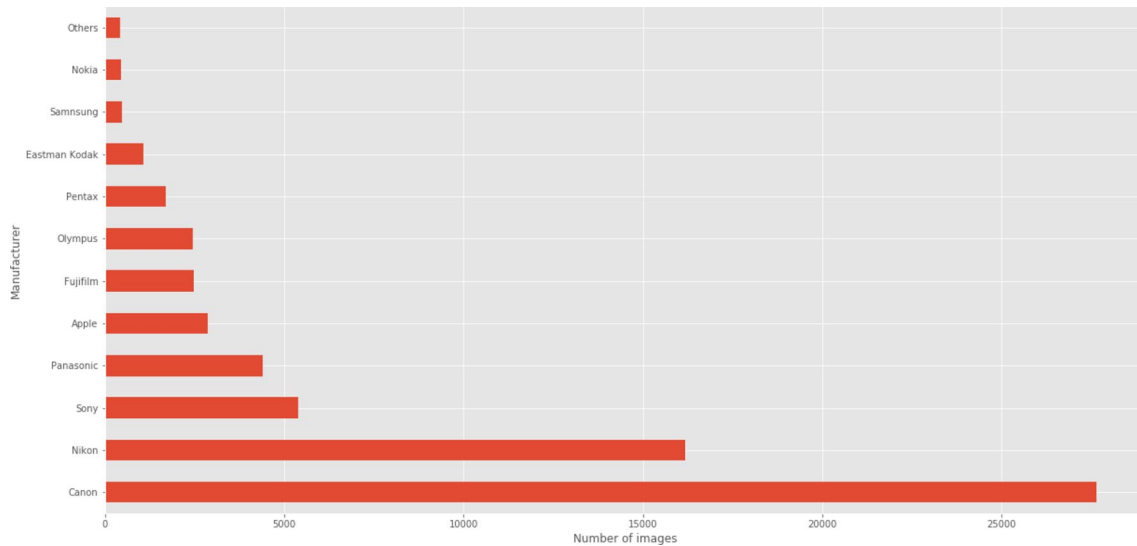
For focal length, lenses between 35–85 mm are the most common, accounting for 50.7% of the images, with the least popular being the super telephoto, only used to take 1.6% of the photos in our dataset (Fig. 6).

Exposure was more evenly distributed between the extremes with the notable exception of 1/60, identified as significantly more popular than all other shutter speeds. This is possibly due to the common belief that this is the slowest shutter speed one can expose without needing a tripod.<sup>14</sup>

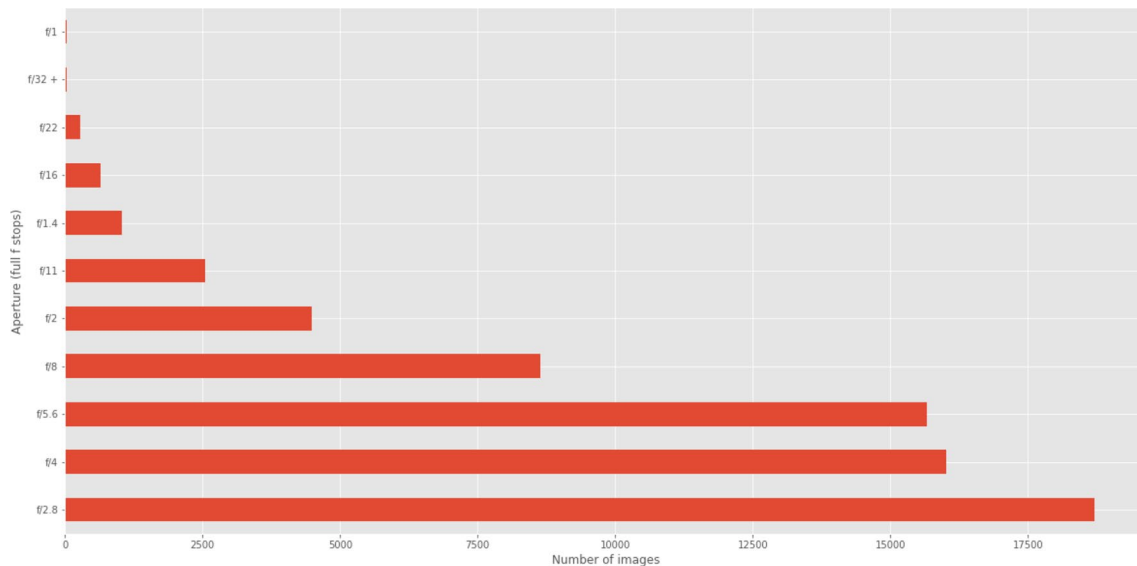
<sup>14</sup> This is commonly known as the “1/focal length rule”. According to it, for a 50 mm lens (75 mm in most APS-C sensors), the longest exposure that still produces sharp images with hand-held cameras would be approximately 1/60. This is only a guideline, as many other factors impact sharpness: ISO, time of day, weather, and indeed how much one’s hand shake. Still, as rule of thumb for DSLR aficionados, it might contribute to explain the popularity of this particular shutter speed.

Footnote 13 (continued)

would need to extract the size of the sensor from Exif (assuming this is not given as the 35 mm equivalent), calculate the crop factor for each individual camera model, and then match it to the corresponding entry in the dataset. We did not have the time or resources to do this. However, through controlled manual sampling we identified entries that reported focal lengths consistent with two types of cameras widely available at the time these pictures were taken: 3G Mobile phones (~ 15 k entries, e.g. iPhone 4 to 5), compact and ultra compact point and shoot cameras (~ 12 k entries, e.g. Canon Powershot and Pentax Optio series). Based on this we compensated for these two groups by applying a weighted average crops factor of 7.6 and 4.8, respectively. From a similar sampling at the other end, it was apparent that this process not necessary for long focal lengths, which were mostly taken with full frame or APS-C or APS-H cameras, which magnify the image even more.



**Fig. 4** Camera manufacturers of images in the Visual Genome



**Fig. 5** Distribution of apertures in images from the Visual Genome

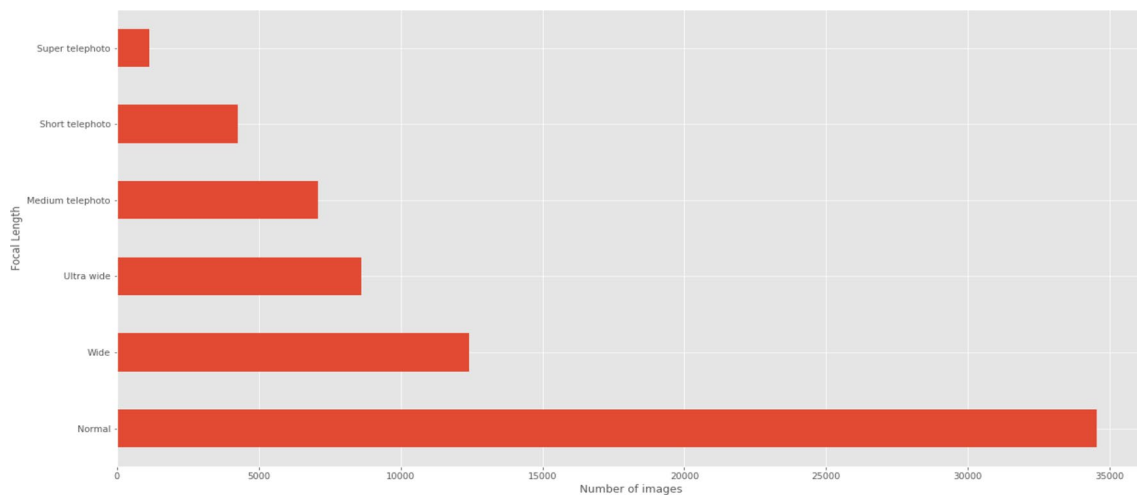
Figure 7 shows the ten most common combinations of aperture, focal length and exposure parameters in images in the Visual Genome, all of which are under direct control of their photographers.

<sup>15</sup> Single-Lens Reflex cameras (both digital and analogue). This type of camera allows for interchangeable lenses and has a mirror system that allows the photographer to use a view finder to see through the camera lens in order to compose their photographs. When the shutter is pressed the mirror flips and the sensor or film stock gets directly exposed to light coming in through the lens. The acronym is often used to differentiate these cameras from point-and-shoot models, which are much smaller and have fixed (often retractile) lens, or from

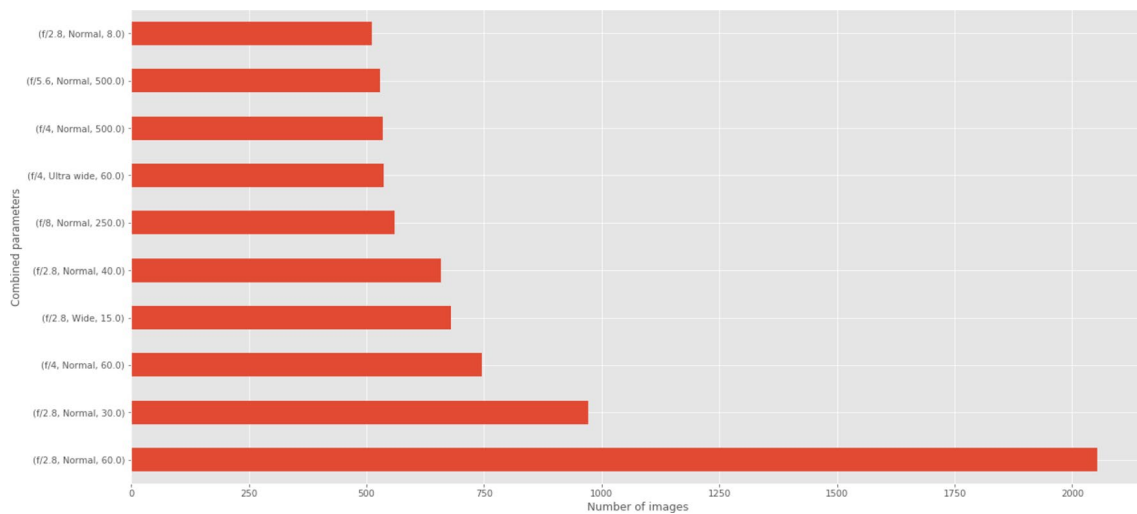
These findings are consistent with the practices of a “proficient consumer” community of photo enthusiasts working with DSLR equipment.<sup>15</sup> These are generally non-professional photographers who nevertheless are willing to invest in a bulkier and more expensive camera and take the time to learn how to operate it manually. Users of the Nikon D90 are often recent converts migrating upwards from the

Footnote 15 (continued)

so-called mirrorless models, which do admit different lenses but do not have a mirror.



**Fig. 6** Focal length categories in images from the Visual Genome



**Fig. 7** Combined aperture, focal length and exposure\* of images in the Visual Genome. \*Exposure is given as the denominator of a fraction of a second, e.g. 250 is equivalent to 1/250, or 0.004 s

common point-and-shoot photography. Or they might also be more established and committed users of a Canon 7D, who probably own a few lenses already and might be close to going professional. This grouping is also supported by our smaller sample of lens data from the <Lens Info> tag, which shows inexpensive lenses that come bundled with cameras to be very popular, e.g. the 18–55 mm  $f/3.5$ – $5.6$ L included in both Nikon and Canon starter kits (camera body + lens), but also include a few more expensive lenses (particularly on longer focal lengths, e.g. the 100–400 mm  $f/4.5$ – $5.6$ L or the EF70–200 mm  $f/2.8$ L, both by Canon).<sup>16</sup> We believe these

lenses overlap with professional practice and were probably acquired as second or third lenses for purpose-specific photography, specifically wildlife or sports, both of which featured heavily in a manual sampling we conducted over images taken with these two models.<sup>17</sup>

By identifying the dominant photographic practices of this community of DSRL enthusiasts in the Visual Genome, we show the *implicit optical perspective* mobilised in MbM. If one were to ask not about the accuracy in detecting

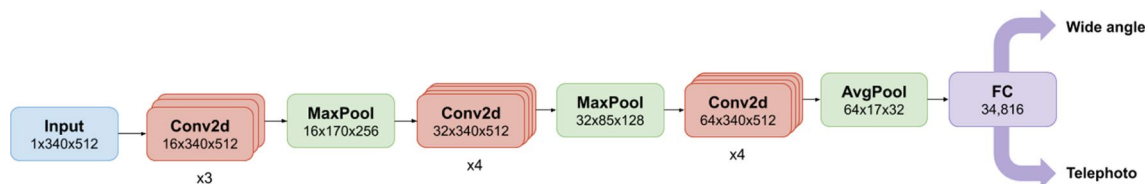
<sup>16</sup> We believe it is possible that Canon lenses on this range make more consistent use of the <Lens Info> tag.

<sup>17</sup> We looked at about 10% of the images taken with these two lenses.





**Fig. 8** Batch of four samples of inputs and labels



**Fig. 9** 15-layer Convolutional neural network architecture

what is depicted, but about the *latent camera* of this particular computer vision system, we could now reply with some degree of confidence that this perspective falls within the focal range of a 18–55 mm lens on a APS-C or APS-H camera; apertures between  $f3.5$  and  $f5.6$ , and a likely exposure  $1/60$  s. Casting aside some of the other complexities of MbM for a moment, we could say that in general terms this was the lens through which the BBC archive was seen.

Today, DSLR photography of this kind is a somewhat dying practice, as sales of this type of camera have been steadily declining over the past decade (CIPA 2019). Everyday photographs are now taken with mobile phones and circulated through social media (Herrman 2018). However, while the equipment and the communities that supported this visual regime recede into history, lens aesthetics are anything but history. On the contrary, the standard of photography set by DSLR practitioners is now being reimagined under the logic of digital computation and mobile phones,<sup>18</sup> pursued through software and through AI (See for example: Yang et al. 2016; Ignatov et al. 2017).

With this in mind, we suggest turning computer vision to itself and asking whether it is possible to engineer a machine

that tells us about the becoming of images; not only *what* they depict but *how*. If we concede that the “aboutness” with which we invest photographic images—including their epistemic advantage—is a function of the depicted no less than of the depiction modality, such an aesthetic machine, we argue, is as justified as one that distinguishes cats from dogs, or hot-dogs from other sandwiches. Could we not train machines to learn about optical perspectives as well as what these perspectives are used for at given times in history?

To close this article, we offer a prototype along these lines as a proof-of-concept, which is purposefully designed to be blind to what photographs are of; a type of vision that cares nothing about recognising objects, people or scenes, and is instead programmed to learn only about how its images were made and the visual perspectives they embody, in this case the focal distance of the lenses with which they were taken.

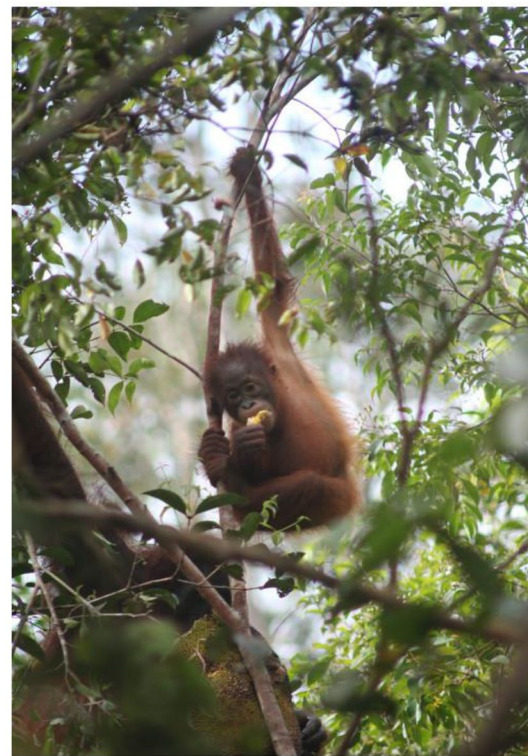
Using the EXIF dataset we assembled and the images from the Visual Genome, we trained a Neural Network to classify focal lengths and to distinguish between photographs taken with a wide angle lens from those taken with a telephoto. The class boundaries are drawn at under 24 mm for the former and over 135 mm for the latter. Each class was given little over 12,000 training samples (Fig. 8). The model was trained from scratch using a VGG-based convolutional neural network (Fig. 9).

Our results show test accuracy of 83% after fourteen epochs of training. We manually tested the model at this checkpoint by running inference on several photographs not contained in the Visual Genome to confirm it performed as expected in evaluating out of sample images. But we are only at the beginning of our work here. Without probing further into the model and conducting more systematic tests, it is difficult to know what exactly the neural network has

<sup>18</sup> Consider how large phone manufacturers like Huawei and Nokia partnered over the last decade with camera and lens manufacturers Leica and Zeiss, respectively, to produce multi-camera devices. In the case of the Huawei P series, such partnership was overtly marketed with the slogan “rewrite the rules of photography”, in direct reference to its capacity to reproduce and control lens effects such as shallow depth of field and bokeh, which occur at longer focal distances and narrower apertures, and had until recently been the sole province of photographic cameras, most notably professional and semi-professional SLRs. See: <https://consumer.huawei.com/uk/campaign/rewritetherules/>



class: 'wide angle'



class: 'telephoto'

**Fig. 10** Photo on the left was taken with a mobile phone (28 mm); photo on the right with a DSLR (340 mm)

learned from these images. One of our working hypotheses is that there are low-level features like the texture of bokeh or warmer green tones which might correlate strongly to longer focal lengths, since both the field view and speed of many of these lenses favours their outdoor use. In any case, our initial results already suggest that, with some exceptions such as irregularly shaped images from elongated panoramas, grainy images or images captured with optical zoom, the predictions of our classifier were reasonably accurate for photographs taken with either very long or very wide lenses. Figure 10 shows a comparison of two successfully classified images using this method. For the casual observer who sees these two images all at once, instead of counting them pixel by pixel, there are many apparent differences: one is the Shard in London, the other a baby orangutan in Borneo; one is a landscape, the other a portrait; one is a night scene, the other was taken in broad daylight. However, when it comes to the type of lens used to render these scenes visible, a *posteriori* knowledge might in fact be a task for which computer vision is much better suited. In particular deep convolutional networks can help with their progressive and content-agnostic abstraction of pixel relations.

Going back to MbM, we used our focal length classifier on frames from one of the mislabelled sections mentioned at the beginning (Fig. 11). Comparing the predictions

outputted by the two systems, our 'telephoto' classification seems intuitively more accurate than MbM's 'reflection in a mirror'. This might be an extreme example but it points us to a fundamental problem that is sometimes overlooked in machine learning. Which prediction tells us more about the image? What kind of knowledge is implied by each, and when or why would we prefer one kind over the other?

## 5 Conclusions and future research

Initial results suggest that with more data, extended training and fine-tuning, a much more sophisticated lens classifier is possible. To our knowledge this has not been tried before, and therefore, we approached the design of the system with naive confidence, in the hope that others might be intrigued and improve upon it. Similarly, we believe there are many other possibilities beyond a binary classifier, even using this relatively small dataset, for example by looking at exposure as one of the immanent temporalities of computer vision systems—photographers not only manipulate the shape of light but also its speed. These need not be isolated dimensions nor indeed separate from existing approaches aimed at naming objects or people.



**Fig. 11** Predicted class of (the whole) image on the left using our focal lens classifier prototype. Predicted label of (a region) of the image on the right in MbM

Our contribution is, rather, an initial and tentative answer to a much larger question: how can computer vision evolve from systems designed to name what is in the picture, to systems that approximate more precisely what we see in the picture? In this paper, we show from a computational aesthetics perspective how diversity in photographed subjects ought not to be confused with visual diversity, nor indeed bias with error. To be clear, our argument is not that Crawford’s archaeology of datasets is not necessary, or that Harvey is mistaken when he states that a photo is not just a photo any more. It is rather that “just a photo” includes a whole field of meanings and technical mediations that are also encoded, abstracted and mobilised through machine learning and deep learning in particular. This is not to deny the digital dimension of these images, nor the latent computational powers of datasets, but to say that these powers are largely derived from the representational powers of photography. Therefore, our relationship with the photographic image underwrites our relation with computer vision more generally. From this perspective, a critical programme of computer vision, insofar as it is powered by this type of images, necessitates a techno-aesthetics of photography to explain how these images afford knowledge by distorting perspective, and how they can be seen as faithful representations beyond the factual events they depict. So here we must simply insist in incorporating insights from the study of the photographic and cinematic image in the technical milieu of AI to argue that meaning is not something that can be extracted from pictures alone, but that is instead co-constructed with their audiences through usage; photography is an imaging no less than an imagining technology, and so too must we learn to understand computer vision.

**Acknowledgements** Daniel Chávez Heras is funded by Mexico’s National Science and Technology Research Council (CONACYT).

## Compliance with ethical standards

**Conflict of interest** The authors hereby declare to not have any conflict of interest or competing interests. The dataset and code produced for this article is publicly available under MIT licence from: [https://github.com/chavezheras/shape\\_of\\_computervision](https://github.com/chavezheras/shape_of_computervision). A training dashboard of the classifier model described in this article is publicly available at: <https://cutt.ly/4taTWmt>

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abell C (2010) The epistemic value of photographs. in philosophical perspectives on depiction, ed. Catharine Abell and Katerina Bantinaki. Oxford University Press, Oxford
- Bazin A (1960) The Ontology of the Photographic Image. Translated by Hugh Gray Film Quarterly 13:4–9. <https://doi.org/10.2307/1210183>
- BBC (2018) BBC—Public Purposes. <https://www.bbc.co.uk/corporate2/insidethebbc/whoweare/publicpurposes> Accessed 02 Oct 2018
- Cavell S (1979) The world viewed: reflections on the ontology of film. Harvard University Press, Cambridge
- Chávez Heras D, Blanke T, Cowlshaw T, Man D, and Herranz Donnan, A (2019) Seen by machine: computational spectatorship in the BBC television archive. In ADHO Proceedings. Utrecht, Netherlands
- CIPA (2019) Camera and imaging products association: statistical data report. Sales and shipment report. Digital Cameras
- Clark A, Chalmers D (1998) The extended mind. *Analysis* 58:7–19

- Cohen J, Meskin A (2004) On the epistemic value of photographs. *J Aesth Art Crit* 62:197–210
- Costello D (2017) *On photography: a philosophical inquiry*. Routledge
- Cowlishaw T (2018) Using artificial intelligence to search the archive. BBC R&D Blog. <https://www.bbc.co.uk/rd/blog/2018-10-artificial-intelligence-archive-television-bbc4> Accessed 28 Sept 2018
- Crawford K, and Paglen T (2019) Excavating AI: The politics of images in machine learning training sets. AI Now Institute. <https://www.excavating.ai/> Accessed 10 Sept 2019.
- Cumberbatch G, Bailey A, Lyne V, Gauntlett S (2018) On-screen diversity monitoring BBC One and BBC Two. Media Monitoring. Cumberbatch Research group
- Currie G (1999) Visible traces: documentary and the contents of photographs. *J Aesth Art Crit* 57:285–297. <https://doi.org/10.2307/432195>
- JEITA Standards. Exchangeable image file format for digital still cameras: Exif
- Harvey A, LaPlace J (2019) MegaPixels: origins, ethics, and privacy implications of publicly available face recognition image datasets. <https://megapixels.cc/> Accessed 18 Apr 2019
- Herrman J (2018) It's almost 2019. Do you know where your photos are? *The New York Times*
- Ignatov A, Kobyshev N, Timofte R, Vanhoey K, Van Gool L (2017) DSLR-quality photos on mobile devices with deep convolutional networks. arXiv:1704.02470 [cs]
- Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vision* 123:32–73. <https://doi.org/10.1007/s11263-016-0981-7>
- Li F-F (2019) ImageNet 10th birthday party, september 21. The Photographers Gallery, London
- Lopes DI (2016) *Four arts of photography: an essay in philosophy*. Wiley, Hoboken
- Maynard P (1997) *The engine of visualization: thinking through photography*, 1st edn. Cornell University Press, Ithaca
- Pasquinelli M (2019) Three thousand years of algorithmic rituals: the emergence of AI from the computation of space. *e-flux* 101
- Peres MR (2007) *The focal encyclopedia of photography: digital imaging, theory and applications, history, and science*. Taylor & Francis
- Phillips DM (2009) Photography and causation: Responding to Scruton's scepticism. *The British Journal of Aesthetics* 49. Oxford University Press: 327–340. <https://doi.org/10.1093/aesthj/ayp036>
- Walden, Scott. 2005. Objectivity in Photography. *The British Journal of Aesthetics* 45. Oxford Academic: 258–272
- Wieczorek M (2019) What I think about when I think about Focal Lengths. Medium. Accessed 28 Dec 2019
- Yang Y, Lin H, Yu Z, Paris S, Yu J (2016) Virtual DSLR: High quality dynamic depth-of-field synthesis on mobile platforms. *Electron Imaging* 18:1–9. <https://doi.org/10.2352/ISSN.2470-1173.2016.18.DPMI-031>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# The brain, the artificial neural network and the snake: why we see what we see

Carloalberto Treccani<sup>1</sup>

Received: 29 July 2019 / Accepted: 18 August 2020 / Published online: 12 September 2020  
© Springer-Verlag London Ltd., part of Springer Nature 2020

## Abstract

For millions of years, biological creatures have dealt with the world without being able to see it; however, the change in the atmospheric condition during the Cambrian period and the subsequent increase of light, triggered the sudden evolution of vision and the consequent evolutionary benefits. Nevertheless, how from simple organisms to more complex animals have been able to generate meaning from the light who fell in their eyes and successfully engage the visual world remains unknown. As shown by many psychophysical experiments, biological visual systems cannot measure the physical properties of the world. The light projected onto the retina is, in fact, unable to specify the physical properties of the world in which humans and other visually ‘intelligent’ animals behave; however, visual behaviours are habitually successful. Through psychophysical evidence, examples of the functioning of Artificial Neural Networks (ANNs) and a reflection upon visual appreciation in the cultural and artistic context, this paper shows (a) how vision emerged by random *trial and error* during evolution and lifetime learning; (b) how the functioning of ANNs may provide evidence and insights on how machine and human vision works; and (c) how rethinking vision theory in terms of trial and error may offer a new approach to better understand vision—biological and artificial—and reveal new insights into *why we like what we like*.

**Keywords** Human vision · Machine vision · Brain · Artificial neural network · Visual appreciation

## 1 Introduction

It is possible to consider metaphors as a way to attribute human characteristics to an animal, object or any other subject. At the same time, it is important to consider that abstract (non-physical) phenomena are understood through the attribution of physical features to the phenomena. For instance, the sentence *inflation rose in July* is interpreted using two concrete physical phenomena: inflation (an increase in size) and rising (a change in position) (Pinker 2013).

Metaphors are used in science and philosophy to explain the unknown, as well as a tool to generate new knowledge and provide a better understanding of many phenomena (Zarkadakis 2015). Every century, perhaps even every decade, has its own metaphors, and “when the use of a specific metaphor ceases and a new metaphor takes its place, we have

a ‘paradigm shift’ in the way science explains the world” (Zarkadakis 2015).

One main source of metaphors is the human brain and body. In the Book of Genesis Adam, for instance, is created out of dust and then life is infused into it—interestingly, the word human comes from the Latin *humus*, which means ground or earth. Later on, in the third century BCE, the invention of hydraulic and pneumatic systems provided a new paradigm to understand the human body as a “dynamically moving fluid within a mechanical body” (Zarkadakis 2015). In the sixteenth century, again, René Descartes described the human body as a complex machine, comparing muscles and bones to cogs and pistons.

The arrival of computer technology, however, changed the paradigm once more. The brain started to be compared to a computer, as it processes the information much similar to how a computer does. The computer–brain analogy perhaps finds its roots in the fact that information is transmitted

✉ Carloalberto Treccani  
carloalberto.t@my.cityu.edu.hk

<sup>1</sup> School of Creative Media-City, University of Hong Kong, Hong Kong, Hong Kong SAR

through electrical signals. Neurons,<sup>1</sup> like electronic components, indeed transmit information through a voltage change that much resembles the binary logic—0 and 1—used by computers.

From the 1950s, however, with the advent of Artificial Neural Networks (ANNs), this metaphor has begun to be revised. Unlike a computer, ANNs operate empirically, based on trial and error and on what worked best in the past, with no rigid rules or specific steps to follow—this distinction, which may seem to be of no particular importance, is fundamental to the understanding of this paper.

Using the new brain–ANN analogy, this paper introduces a new understanding of visual perception. Section 2 provides some essential information on the structure and functioning of the nervous system and briefly discusses its advantages. Important terminology is defined to better understand the entire text. Section 3 provides a basic understanding of the functioning of the ‘visual brain’ and introduces the idea that vision should be understood in empirical terms arising from trial and error during evolution and life-long learning. Section 4 provides a basic understanding of the functioning of ANNs. Section 5 clarifies the position of culture in visual understanding. Additionally, it provides insights into comprehend *why we like what we like*. Lastly, Sect. 6 presents some implications and possible consequences of ‘accepting’ the brain–ANN paradigm.

## 2 The nervous system

*Trichoplax adhaerens* is a marine animal roughly one millimetre in diameter without a nervous system (Senatore et al. 2017). However, despite this lack, as shown by Senatore and colleagues, *Trichoplax* is capable of a variety of behaviours typically found in animals with nervous systems. *Trichoplax* can track space, communicate through cellular transmission and exhibit different feeding behaviours: it can arrest its ciliary movement, used for locomotion, when algae are detected, showing that it has a sensory system able to detect nutrients and to communicate over short distances via chemical secretion. Plants also lack a nervous system but, like bacteria and protists, they can use environmental information to generate behaviours that enable them to survive and reproduce.

What are the evolutionary advantages of having a brain, then? Answering this question risks falling into some kind of hierarchical division, with organisms with a nervous system at the top of the pyramid of life and others at the bottom. It is certainly true that having a nervous system provides

numerous advantages. Nevertheless, having a brain is not a fundamental requisite for an ‘intelligent’ life form: most past and extant organisms are without a nervous system, and this absence does not seem to have caused them any particular problems.

No one knows when the first nervous system appeared; however, the need to survive in a constantly changing environment seems to have benefited those organisms ‘gifted’ with a nervous system (Purves 2019). The key distinction between non-nervous system organisms and those with a nervous system seems thus a *quantitative* distinction in the range of possible behaviours. The appearance of nervous systems and later on of central nervous systems enabled biological creatures to respond to the external environment in more sophisticated and useful ways (Robson 2020).

The nervous system is composed of specialised cells known as neurons and glia.<sup>2</sup> It is described for convenience as the central nervous system (brain and spinal cord) and the peripheral nervous system—both areas are of course in continuity. The task of the nervous system is to carry sensory information (e.g., heat) from the periphery to the brain to promote a behaviour (e.g., remove your hand). Each neurons signals via a bioelectrical signal, known as an action potential, which travels along the nerve cell and communicates the information to another neuron—the synapse.<sup>3</sup> Each cell remains independent and separate. In the case of vision, when photons stimulate photoreceptors cells in the retina, the sensory input is transformed into a neural signal and sent to specific areas within the brain. The information is then *processed*, and a behavioural response occurs.

Is certainly true that creatures gifted with *excitable cells* able to perceive and convey information about the outside world have evolutionary advantages compared to creatures without such cells. For instance, in the case of vision, the ability to create a visual representation of the world—e.g., identify a predator. Nevertheless, as previously stated, the majority of living beings that have existed and that currently exist lack a nervous system. Thus, having cells that can ‘represent’ the world and coordinate such representation with a behaviour (e.g., movement) should be seen as increasing the possibilities to representing the outside environment (Churchland 1989) rather than a fundamental requirement. Although the behavioural catalogue of different creatures varies tremendously, living organisms with and without nervous systems have developed strategies to pair sensory

<sup>1</sup> A neuron is a specialised cell for the transmission of electro-chemical signals.

<sup>2</sup> Glia cells are non-neural cells in the nervous system. Their main role is to provide structural support to neurons. They are not directly involved in the transmission of signals.

<sup>3</sup> Synapses are connections between neurons and a target cell. The role of synapses is to allow communication by receiving or transmitting chemical signals.

inputs with useful behavioural answers (Purves 2019). In summary, the advantage of having a nervous system seems that of having a richer behavioural repertoire.

### 3 Vision

According to Andrew Parker's *light switch theory* (2004), the change in atmospheric conditions during the Cambrian period and the subsequent increase in light has triggered the sudden evolution of vision and the consequent evolutionary benefits. Despite the fascinating idea proposed by Parker and the consequent belief of the supremacy of vision over other senses, this theory does not explain how organisms have been able to usefully pair a visual stimulus<sup>4</sup> (e.g., the 'image' of a prey) with a useful behaviour (e.g., eat the prey). After all, at that time, vision was a new and, therefore, unknown source of information. How have simple organisms and more complex animals been able to develop useful behaviours in response to visual stimuli, or put differently, how *visually gifted creatures*<sup>5</sup> have been able to create meaning from the light that enters their eyes?

The plain individual is probably convinced to see objects with a specific shape, because those objects have those certain physical features or to see them at a certain distance, because they are actually at that distance (Kanizsa 1997). However, as further explained in this paper, despite the apparent simplicity, how vision works, is still largely unknown—which is why, from their first appearance more than 50 years ago, computer vision systems remain to date, under certain circumstances, inaccurate, unreliable and easily deceived. Indeed, the need to further understand and replicate vision has only recently arisen, with the dream and need to build '*visually intelligent machines*' (Treccani 2018).

The conventional idea is that light carries with it information about the world that somehow resembles the world as it is (e.g., I see a house because the light input received by the sensory apparatus carries with it some 'houseness' information, such as shape). This belief is probably based on old theories of vision such as *intromission theories*, which see vision as light rays emitted by objects, and *extromission theories*, which instead understand vision as the emission of light rays from the eyes towards external objects. This idea is, however, incorrect, because light does not carry with it any information other than energy. Furthermore, the human perceptual apparatus cannot measure the physical parameters of the world, because it lacks the necessary instruments

and thus cannot retrieve the real properties of the world (as described more in detail later in this section). Besides, as the world cannot be assessed, even the idea of a representation of the world close enough to reality must be incorrect.

According to Parker's theory, the advantages appearing from the evolution of vision gave rise to numerous new animal behaviours. Seeing in colour, for example, is undoubtedly an immediate asset. As pointed out by Purves, "a visual system that can identify object boundaries based on the spectral distribution of light energy will, therefore, be more successful in responding to images" (Purves 2019). Here, Purves refers to the possibility of perceiving boundaries given by colours. Achromatic animals, for instance, cannot distinguish boundaries between two objects with a spectral difference but the same luminosity (i.e., brightness). Colours, a fundamental source of information that animals use to identify predators and poisonous plants or for reproductive purposes, are, however, a brain construction.

Colours are defined by variations in the wavelength (or frequency) of light. Red, for instance, has a wavelength of between 700 and 635 nm, whereas blue is between 490 and 450 nm. However, light wavelengths themselves do not correspond to any colours. The variation in the frequency of the wavelength, commonly understood as colour variation, is a vibrational variation (moving photons) in the amount of energy in a light wave. In addition, light waves that reach the retina always entail a combination of illumination, reflectance, and transmittance (Fig. 1a), and there is no analytical method to unravelling how these factors provide visually appropriate answers (Purves and Lotto 2003).

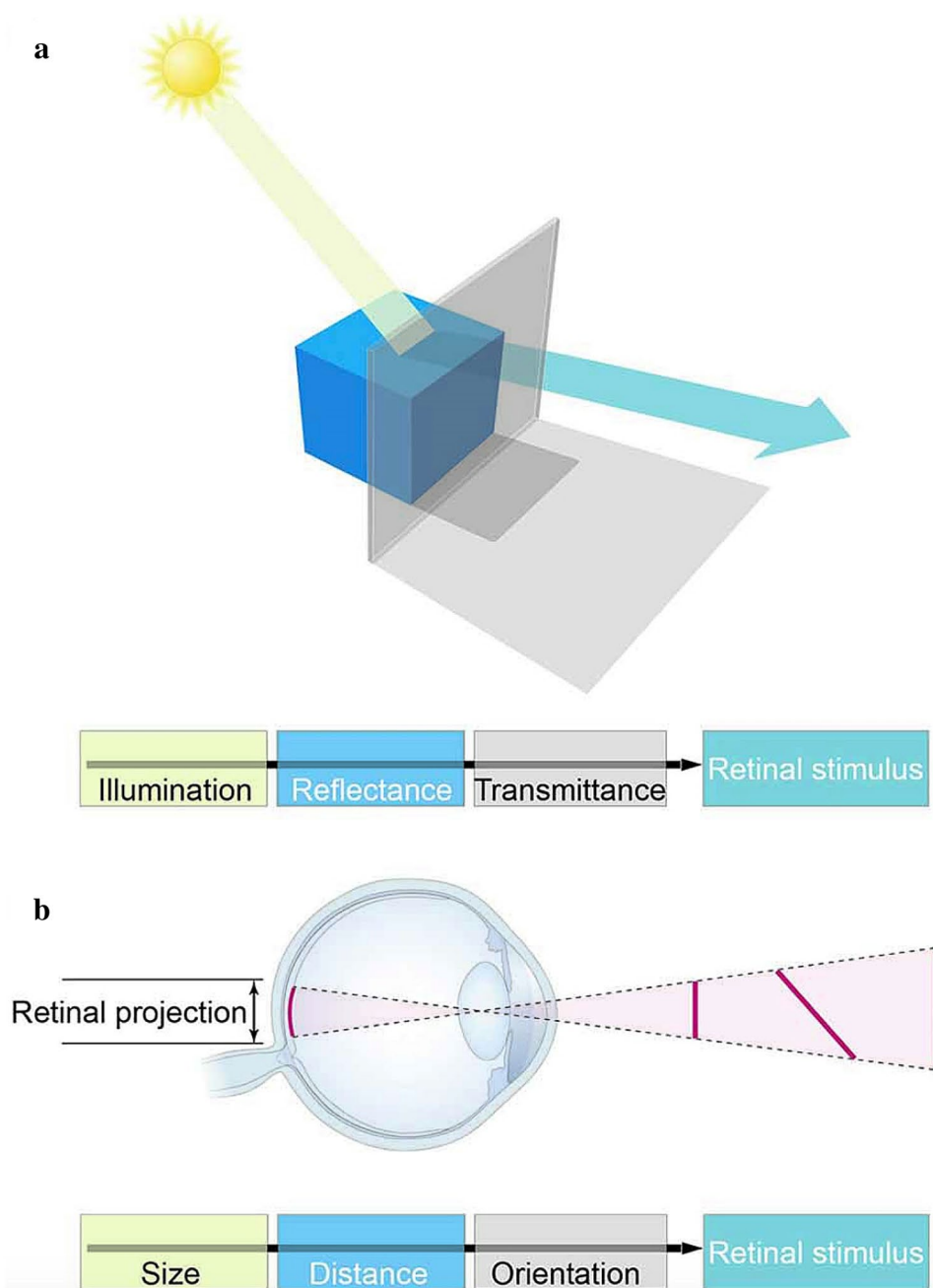
How do humans perceive colours? The capacity of the human eye to discriminate spectral variations is based on the sensitivity of retina cells to different light frequencies. Humans have two types of photoreceptor cells in the retina: rods and cones. Rods seem to play a minor part in colour detection. The three different types of cones are each characterised by a different type of photopigment. Each type of cone is sensitive to a different light frequency (i.e., colour). Yet this explanation does not seem to fully explain how and why colours are perceived. Additionally, the way geometrical properties are perceived shows the inaccuracy of the human vision in discerning reality once more. The human visual system, in fact, cannot retrieve the geometrical properties of the world. As shown in Fig. 1b, objects with different inclinations, sizes and at different distances can indeed generate the same retinal image. As human eyes, and more generally, the entire perceptual system, cannot retrieve the 'real' properties of the world, it is clear that visual perception must be a *generated perception* (Purves et al. 2014).

To explain the operation of visual perception, Dale Purves proposed the idea that vision should be understood in *empirical terms* "in which perceptions reflect biological utility based on past experience rather

<sup>4</sup> In biology an event, in the form of energy, that provokes and activates a receptor cell.

<sup>5</sup> Biological creatures *gifted* with photoreceptors able to transmit information about the outside world.

**Fig. 1** **a** The conflation of illumination, reflectance and transmittance. Many combinations of these objective parameters in the real world can generate the same values of luminance at the retina. **b** The conflation of physical geometry. The same image on the retina can be generated by objects of different sizes, at different distances from the observer, and in different orientations (Purves 2019).



than objective features of the environment” (Purves et al. 2015). Retinal images,<sup>6</sup> continue Purves “conflate the physical properties of objects, and therefore cannot be used to recover the objective properties of the world. Consequently, the basic visual qualities we perceive – e.g., colours, form, distance, depth and motion – cannot specify reality” (Purves et al. 2015). Visual perceptions, therefore,

<sup>6</sup> Images should be intended both as the light pattern detected by the retina, as well as the visual result of the processed light-pattern, meaning its visual representation—e.g., the image of a cat.

must emerge independently from any measurement of the world, as these measurements are not reliable. To paraphrase Purves, the perceptual information—the visual world—we experience, is determined by the frequency of light pattern<sup>7</sup> and its consequent importance in terms of survival. The association between the frequency of

<sup>7</sup> It is important to highlight that a light pattern does not resemble the physical world, nor any ‘image’ of the world produced by the human brain. A light pattern should be intended as the light configuration that is perceived, determined by its frequency of occurrence and its consequent usefulness (for more, see Purves et al. 2015., Yang and Purves 2004, Rao et al. 2002).



occurrence of a light stimulus (light pattern) and its consequent useful (successful) behaviour thus arises from trial and error during evolution and life-long learning.

If, as previously stated, animals' visual perceptual apparatus cannot assess the physical properties of the world, how are then images of the world formed? For Purves, if a given light stimulus occurs often, its value will be high. At first, different and random behaviours will appear in response to this stimulus. However, over time—evolutionary time and individual lifetime—an automatic link between the stimulus and the most successful behaviour will arise. According to Darwin and neo-Darwinian theory (the integration of Darwin's theory of evolution by natural selection and Mendel's theory of genetics as the basis for inheritance), when a mutation occurs and is randomly associated with a neural response that promotes survival/useful behaviour, this mutation and its consequent neural activity will tend to be passed to subsequent generations. In other words, if the image that arises from a given luminous-configuration and its consequent behaviour proves to be useful for evolutionary purposes, a link will be created. This successful creation will be disseminated to the next generation following evolutionary processes (thus, a non-useful behaviour will *not* tend to be passed to future generations).

Purves, even if referring here to the process of hearing, gives an explicative example that can be likewise transferred to vision: “This strategy works because it establishes an objective-subjective association in biological machinery that does not depend on the measurements of sound sources in physical reality. As before. The role of the physical world in this understanding of sensory neurobiology is simply an arena in which neural associations are empirically tested according to survival and reproductive success” (Purves 2019).

As humans perceive light “categorically—usefully but not at all accurately—because it is an extremely efficient way of perceiving visual stimuli that allow us to save brain cells to devote to the neuroprocessing of our other senses” (Lotto 2017), vision should be seen as a fast answer to a visual stimulus. Consider, for instance, the knee-jerk reflex. This reflex is simply an evolutionary ‘answer’ that establishes a successful behaviour. The activation of the nervous system, which is rapid and precise and without the need for any ‘brain computation’, promotes a response (i.e., extend the leg) that is useful (i.e., extend the leg to avoid falling if an object hits you). In this sense, vision should be considered a *reflex*: an automatic, quick and useful answer to a visual stimulus that does not require any measurement of the world (Purves et al. 2015). The role of vision, and more generally of the nervous system, is, therefore, to promote a biological advantage rather than to reveal *how the world looks like* (Purves 2019). In short, the role of vision is to promote what was *useful* to see in the past.

## 4 Artificial neural networks (ANNs)

Evidence of neuroscientific practices has been found in ancient societies all around the world. However, only with the advent of electronic technologies did scientists start to try to replicate the functioning of the brain. In the 1940s (following Turing's work in the 1930s), McCulloch and Pitts, respectively a neurophysiologist and a mathematician, began to investigate the possibility of neural computation. In 1943, they published a paper on the operational functioning of neurons and the construction of ANNs that could compute logical functions (McCulloch and Pitts 1943). With the idea of building a machine able to solve any logical operation, McCulloch first, with the help of Pitts later, embarked on the design of a mathematical model—a network—of brain functioning.

Towards the end of the 1940s, Donald Hebb published *The Organization of Behavior* (2002), in which he demonstrated that the more a neural pathway is used, the stronger it becomes. This concept is of fundamental importance: it shows that connections between neurons can change their *synaptic weight*, i.e., the strength of the connection. The stronger the connection, the more likely an association (stimulus-behaviour) will be to appear in the future. It is essential to understand this idea, as it is critical to the way humans and machines learn to see the world. In contrast, if seeing a stimulus—input—in a particular way shows itself not to be useful, it is less likely that the same neural pathway will appear.

In the past 10 years, ANNs seem to have gained the upper hand over algorithmic processing. Indeed, the “abilities to recognize patterns, make inferences, form categories, arrive at a consensus ‘best guess’ among numerous choices and competing influences, forget when necessary and appropriate, learn from mistakes, learn by training, modify decisions and derive meaning according to context, match patterns and make connections from imperfect data” (Greenwood and Bartusiak 1992) have made ANNs particularly successful—this is also why, all around the world, neuroscientists use ANNs to simulate the brain functioning. As example, a paper presented in 2018 by the DeepMind group (Google) (Silver et al. 2018) shows how the AlphaZero system beat the best human Go player in the world. Instead of using the *force* of an algorithm—following fixed procedures—AlphaZero learned, *tabula rasa*, how to win by playing against itself millions of time via a process of trial and error. By random trial and error (i.e., reinforcement learning), the system learns from wins and losses (i.e., empirical evidence) to adjust the parameters (i.e., synaptic weight) of the network, making it more likely to choose useful moves in the future. By ranking the frequency of winning moves, AlphaZero quickly learns how to become the strongest player in Go history.

In other words, differently from traditional game engines like IBM's Deep Blue<sup>8</sup> that rely on thousands of rules, fixed procedures and heuristic moves chosen by human experts, AlphaZero learns how to become the best player, *without in-built knowledge*—except the rules of the game—but by random self-play—trial and error (Hassabis et al. 2018).

Even if first created more than 50 years ago, it is only recently that ANNs have achieved satisfying results. ANNs are now used for image recognition tasks, to automatically identify objects, people, actions and places in a given image. Image recognition technologies are used by companies and governments to execute tasks such as guiding robots and vehicles, content searching in images, image labelling and medical diagnosis. Facebook, for instance, uses ANNs to help visually impaired users to identify people or objects in a photograph. Self-driving cars use ANNs to locate pedestrians and vehicles, and airports use facial recognition technologies as a biometric confirmation to allow entry to a country.

For the purposes of this paper, it is of great importance to clarify the driving principles of an ANN's functioning, as their understanding seems to provide insights into comprehend how the brain solve visual issues.

In his book *Intelligence Emerging*, Keith L. Downing explains the operating of an ANN as follows: “Take a brief pause from reality and imagine that you are the first-year coach of a professional basketball team. It is the first game of the season, and the score is tied with only seconds remaining. You call a timeout, consider your multitude of strategic options, and then decide to set up a play for the guy that everyone calls C. The play begins, C gets the ball, and though double-covered by two hard-nosed defenders, gets off a long shot. Swish! The ball goes through the net, you win the game, and C is carried off the court on the shoulders of his/her teammates. You have now learned some valuable information that will help throughout the season: C is a clutch performer” (Downing 2015). Let us also imagine a few other scenarios.

In a second case, a player known as D gets the ball and shoots, but the ball does not reach the basket, and the match is lost. The newspapers will attribute the defeat to D.

In a third scenario, as described by Downing, C passes the ball to B, who shoots and wins the match. In this case, both C and B will be glorified by the media and acknowledgement will be given to both—indeed, as Downing notes, basketball statistics include assists as evidence of a player's value.

In a fourth new scenario, “C passes to B who passes to (the guy everyone just calls) A, who makes the winning shot. A gets the points and the shoulder ride to the locker room, B gets the assist, but does C get (or deserve) anything? It could be the case that, before passing to B, C faked a pass or dribble attack that froze A's defender in place. This made it easier for A to come free to get the ball and shoot the winning basket. Such a contribution by C would not show up on a statistics sheet, though you may notice it. You may even praise C more than B or A afterwards in the locker room, since his/her fake-then-pass was obviously the key to the whole play. He set up a situation that then became routine for B and A” (Downing 2015).

After many games, it becomes clear that C plays a critical role during the final minutes of each game. In fact, the value of C is proved by the statistical significance of his performances. The proved value (i.e., usefulness) of a player during the closing minute of a match is essential to plan the strategy and organisation of the team through the rest of the season. Moreover, it is necessary to highlight that all shots, faults, mistakes, defensive impact, points, passes and all the events that precede the win are equally important when assigning value to a player or a particular team configuration. Was L fundamental for the winning of the game during his/her 2 min on the court? Was C a valuable player during the first part of the tournament but less valuable during the last part of the season? Was D a valuable player throughout the season despite the shooting error in the last game?

In short, the sum of all the *trials* and *errors* of the coach's choices and the consequent accumulation of experiences describe the essence of the functioning of an ANN—*reinforcement learning*.<sup>9</sup>

The more the trial and error search space is extended, the more likely is that a useful strategy will be found, although the incredibly large size of the search space and the difficulty of taking into account all of the possible variables do not allow the success of a decision to be determined with certainty. In the case of an *object recognition task*, for instance, it is difficult for a machine vision system to identify all the possible variables—e.g., size, shape, colour, orientation—of an object—e.g., a chair. Furthermore, an object can be partially hidden by other objects (a chair partially hidden by a table) or reflected into a mirror, creating even further difficulties (Treccani 2018). However, the ability of an ANN

<sup>8</sup> On 11 of May 1997, after a 6-match game, Deep Blue beat the world-best chess player—Garry Kasparov. The Deep Blue's win had an important symbolic significance in the advancement of 'intelligent' machine-artificial intelligence.

<sup>9</sup> Reinforcement learning, in machine learning's context, refers to the ability of an agent to learn by trial and error without supervision. Reinforcement learning, unlike supervised learning, does not require any labelled data. In the context of machine vision, a labelled data, for instance, might be a photograph in which the objects represented are specified—e.g., a cat, an apple or a house. Labels are usually obtained by asking annotators—humans—to make judgments about the content of a given image.

to successfully solve a task—for instance, detect every chair present in a given picture—increases over time. Increasing the exploration of the search space for the possible solution increases the possibility of a successful result.

Like Downing’s first-year coach at the season’s start, an ANN begins with a series of random decisions (e.g., to choose player D instead of C)—exploration. Over time and through numerous attempts, the network uses the information gathered to craft more useful decisions—exploitation. As the ANN explores the space of possible choices, it learns that certain decisions or actions lead to a reward while others lead to negative consequences.

In short, an ANN solve *the problem of vision* by millions of random trial and error or, in other words, through a well-indexed database of past experiences and not through logical procedures. Just as an ANN system learns how to recognise faces among millions of other faces, distinguish different dog breeds or recognise and describe a scene, a biological system learns how to see by countless trials and errors during evolution and individual lifetimes. In light of this, a new, *wholly empirical understanding* of the way humans and machines see based on trial and error is therefore needed (Purves et al. 2015).

## 5 The snake

The idea that vision emerges empirically, by trial and error, during evolution, life-time experience and training period—in the case of an ANNs—was previously presented in Part II and III. Extending this hypothesis to the cultural and aesthetic realm, part IV, will suggest that also visual appreciation should be understood as a useful behaviour emerging by trial and error.

Culture plays an important role in visual understanding. Nevertheless, culture has to be understood in evolutionary and biological terms, which means emphasising the evolutionary advantages of developing cultural traits—ideas, technologies and behaviours—that can be transmitted from an individual to another via parenteral and social learning. This idea implies that culture can be seen as a biological extension or, as noted by Creanza, as “the extension of biology through culture” (Creanza et al. 2017).

Relevant in this regard is the case of *snake detection theory*, which holds that “humans and other primates can detect snakes faster than innocuous objects” (Van Le et al. 2013). In *The Fruit, The Tree, and The Serpent: Why We See So Well*, Lynne Isbell argues that “When snakes (the Serpent) appeared, a particularly powerful selective pressure... favored expansion of the visual sense” (Isbell 2009). Isbell argues that the environmental pressure caused by the appearance of snakes—as competitors and predators—and their threat to survival was a possible trigger for the complexity of

the primate visual brain, its enlargement and the particular sensitivity towards snakes. “Across primate species, ages, and (human) cultures, snakes are indeed detected visually more quickly than innocuous stimuli, even in cluttered scenes. Physiological responses reveal that humans are also able to detect snakes visually even before becoming consciously aware of them” (Van Le et al. 2013).

Interesting is the reference that the author makes to the events of the Garden of Eden described in the Old Testament. Eve’s mistake was, in fact, noticing the snake. If she had not noticed the animal, she probably would not have eaten the apple. Snake references are found not only in Judeo-Christian confessions but also in other religions and cultures. Snake representations are found in pre-Christian societies, on Sumerian amulets, Iranian boxes, Greek and Chinese mythology. The presence of snakes in different cultures and the fear of snakes (ophidiophobia), seems thus to have an evolutionary explanation—a visual reminder that snakes are dangerous. As Isbell noted, “ophidiophobia may go way, way back, to at least 30–35 million years ago when the first Old World monkeys and apes, the so-called catarrhine primates, are thought to have appeared. Ophidiophobia may even extend farther back to 60 million years ago when the first generalised simian primates, the anthropoids, are thought have appeared. If so, this timeline might help explain the shared ophidiophobia of all anthropoids, including humans” (Isbell 2009).

The enlargement of the visual brain in primates, for Isbell, seems to have provided a fast and *automatic* ‘predator detection system’ (Isbell 2009). “The ancestral environment of primates uniquely affected them to link vision with automatic, fast, accurate, and adjustable reaching and grasping, and to improve upon vision as a way to detect and avoid predators” (Isbell 2009). As also pointed out by Purves (see Sect. 3), Isbell seems to refer to vision as an *automatism*. The advantages of understanding vision as a *reflex* are in fact: to provide a fast, direct and accurate response to the external environment that does not require any computation or processing, as the ‘computation’ have “already been accomplished by laying down connectivity instantiated by feedback from empirical success over evolutionary and individual time” (Purves et al. 2015). The capacity of humans to visual discriminate an object, for instance, is in fact in the order of tens of milliseconds—this may also provide insights on why, when looking a particular painting, for instance, we are immediately captured by it.

Snakes visual sensitivity and ophidiophobia thus appear to be an evolutionary aid: those who did not respond to snakes with a proper behaviour (e.g., running away) would have fewer chances to survive compared to animals with a more useful behavioural answer (Isbell 2009). The presence of snakes in human artefacts and religions can be explained through the behavioural advantages that arise from the

ability to visually detect snakes. In other words, a particular *visual sensitivity* towards snakes seems to have had cultural repercussions that justify the presence of this animal and its *visual appreciation*<sup>10</sup> in artistic productions all around the world.

It is certainly important to understand the value of culture in the transmission of useful behaviour, for instance, of techniques related to the construction of a tool. It is well known that several non-human species exhibit cultural transmissions. Chimpanzees and macaques, for instance, can build tools like hammers to open nuts using stones. However, even the value of transmitting visual appreciation skills should not be underestimated, it should be read instead as an evolutionary aid in the same fashion of a chimpanzee's ability to build a hammer, although with more degrees of separation.

As shown by Michael Baxandall in *Painting and social experience in fifteenth century Italy* (1988), having the ability to appreciate a painting was a necessary social skill to master for the upper middle and aristocratic Italian classes of the Renaissance. Possessing these *visual skills* gave access to a series of benefits in the form of social connections. In this fashion, artistic visual competence can be seen as an evolutionary aid. These *visual capacities*, although not immediate, must be understood as useful answers to the environmental and social 'selective pressure'. Every given period, in fact, has its own ecological 'selective pressure' and its consequent *useful visual behaviour*.

To trace the biological value of visual appreciation is surely complicated. However, exploring in this direction can reveal meaningful insights into *why we like what we like*. Furthermore, understanding vision and visual appreciation arising from trial and error during evolution and life-long learning can provide a new understanding to study human perception and culture.

## 6 Conclusion

The idea that the visual brain may operate like an ANN, and that vision works in empirical terms poses considerable difficulties both in the science and the humanities; however, as previously shown, there is much evidence that justifies this new analogy. Reconsidering vision in terms of trial and error implies the necessity to revise some of the ideas proposed in this regard, particularly in the fields of psychology and theory of perception studies. However, visual perception issues also need to be rethought in the perspective of brain–ANNs

analogy, since it seems to provide some insight into how human see and visually understand the world.

As demonstrated in Sects. 3 and 4, there are clues that suggest that vision emerges empirically as an automatic answer to a stimulus. The link of the frequency of a light stimulus and consequent useful behaviour determines what is seen. In the same fashion, by exploring all the possible behaviours (e.g., actions, choices) and changing the relative synaptic weight until the best 'move' is found, ANNs successfully learn how to solve a visual task (e.g., object classification, scene classification and image segmentation), possibly as biological creatures have done during evolution. The analogy should then be clear: the way humans and other animals build their way to visually understand the world closely resembles the operational functioning of an ANN—and vice versa.

This idea may seem suspect to many and generate a certain antipathy in others, but the possibility of exploring the vision (and, more generally, the functioning of the whole brain) in terms of trial and error can provide a first model of *why we see what we see*, to date still missing. Although the theories of the last century, especially in the field of psychology (such as Gibson's bottom–up theory and Gregory's top–down theory) and later in neuroscience (such as Marr's computational theory), had the merit of providing further elements of understanding of how the visual system may work. Likewise, these theories were not able to provide a clear explanation of *how*, and most importantly, *why* the 'visual world' is constructed (i.e., the relation between the physical properties of the world, the visual stimulus that falls onto the eyes and the consequent mental image).

As previously discussed in this paper, the physical properties of the world cannot be measured, and the degree of discrepancy between reality and perceived reality must be substantial. Thus, the idea that a subject is able to respond to a stimulus, based exclusively on the features of the stimulus itself (bottom–up theory) appears insufficient. Furthermore, the idea of seeing as knowing (top–down theory) seems equally insufficient, since the state of the world is unknown. However, by random trial and error, by ranking the frequency of the appearance of a light stimulus, the success of the response to that particular stimulus and the consequent neural wiring, animals, and more recently machines, seem to have been able to create a useful visual representation of the world. The role of vision then "is not to reveal the physical world, but to promote useful behaviours" (Purves et al. 2015).

In trying to understand visual perception and its guiding principles, it has to be clear that the human brain evolved from earlier brains, and that its capacity and limitations have a historical basis (Churchland 1989). The selective pressure that developed visual perception—and more generally, all perceptions—did not arise rationally and logically but

<sup>10</sup> Visual appreciation refers not only to the aesthetic appreciation of a work of art but also to the ability to analyse, describe, interpret and make connections between works of art and their cultural context.

instead from the need of living beings to successfully deal with the surrounding world. As the world cannot be conquered by the visual sensory system, animals—including humans—have learned to *usefully represent* and respond to the physical world through trial and error during evolution and individual learning. As an ANN learns how to solve a visual task by randomly trying the possible ‘moves’, so the brain has learned how to successfully respond to the world outside itself.

Comprehending how visual understanding was won by simple organisms first, more complex animals later and—partially—by machines recently, may be of great help as it may help to unhinge wrong assumptions and test the validity of new ones. Abandoning the nest of common-sense conceptions about vision—both biological and artificial—will certainly lead to further complications that nevertheless deserve to be investigated to better understand and provide a first, albeit imperfect, answer to the relationship between *what we see and what we know* (Berger 2008).

**Acknowledgements** Many thanks to Olli Tapio Leino for his continuous support, to Patricia Armati for her comments and to Dale Purves for his guidance in writing this paper.

**Funding** The writing of this paper was possible thanks to the financial support of the School of Creative Media-City University of Hong Kong and the University Grants Committee, Hong Kong, SAR, China.

## References

- Baxandall M (1988) *Painting and social experience in 15th century Italy*. Oxford University Press, Oxford
- Berger J (2008) *Ways of seeing*. Penguin Classics, London
- Creanza N, Kolodny O, Feldman MW (2017) Cultural evolutionary theory: how culture evolves and why it matters. *Proc Natl Acad Sci* 114(30):7782–7789. <https://doi.org/10.1073/pnas.1620732114>
- Churchland PS (1989) *Neurophilosophy toward a unified science of the mind-brain*. A Bradford Book, Cambridge (MA)
- Downing KL (2015) *Intelligence emerging*. Adaptivity and searching in evolving neural systems. MIT Press, Cambridge
- Greenwood A, Bartusiak MF (1992) *Neural networks: computational neuroscience: a window to understanding science at the frontier takes you on a journey how the brain works?*. The National Academies Press, Washington, Science at the Frontier
- Hassabis D, Hubert T, Schrittwieser J, Silver D, et al (2018) AlphaZero: shedding new light on chess, shogi, and Go. DeepMind Blog. [https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-andgo?fbclid=IwAR0j8BhGOWaLBPRu00pVBEX0g4TgztWwTxA4\\_J3ozYsBeymnG5\\_QQWoqMXg](https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-andgo?fbclid=IwAR0j8BhGOWaLBPRu00pVBEX0g4TgztWwTxA4_J3ozYsBeymnG5_QQWoqMXg). Accessed 28 January 2020
- Hebb DO (2002) *The organization of behavior: a neuropsychological theory*. Psychology Press, London
- Isbell LA (2009) *The Fruit, the tree, and the serpent: why we see so well*. Harvard University Press, Cambridge
- Kanizsa G (1997) *Grammatica del vedere Saggi su percezione e Gestalt*. Il Mulino, Bologna
- Lotto B (2017) *Deviate: the science of seeing differently*. Hachette Books, New York
- McCulloch WS, Pitts WH (1943) A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 5:115–133
- Parker A (2004) *In the blink of an eye*. Basic Books, New York
- Pinker S (2013) *Stylish Academic writing*. <https://www.youtube.com/watch?v=IE-TTz13P7w>. Accessed 18 January 2020
- Purves D, Monson BB, Sundararajan J, Wojtach WT (2014) How biological vision succeeds in the physical world. *Proc Natl Acad Sci USA* 111(13):4750–4755
- Purves D, Morgenstern Y, Wojtach WT (2015) Perception and reality: why a wholly empirical paradigm is needed to understand vision. *Front Syst Neurosci* 9(November):1–10. <https://doi.org/10.3389/fnsys.2015.00156>
- Purves D, Lotto B (2003) *Why we see what we do: an empirical theory of vision*. Sinauer Associates, Sunderland MA
- Purves D (2019) *Brains as engines of association: an operating principle for nervous systems*. Oxford University Press, Oxford
- Rao RPN, Olshausen BA, Lewicki MS (2002) *Probabilistic models of the brain: perception and neural function*. MIT Press, Cambridge MA
- Robson D (2020) A brief history of the brain. NewScientist. <https://www.newscientist.com/article/mg21128311-800-a-brief-history-of-the-brain/>. Accessed 26 December 2019
- Senatore A, Reese TS, Smith CL (2017) Neuropeptidergic integration of behavior in trichoplax adhaerens, an animal without synapses. *J Exp Biol* 220(18):3381–3390. <https://doi.org/10.1242/jeb.162396>
- Silver D, Hubert T et al (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362(6419):1140–1144. <https://doi.org/10.1126/science.aar6404>
- Treccani C (2018) How machines see the world: understanding image annotation. *NECSUS. Eur J Media Stud* 7(1): 235–254. <https://doi.org/10.25969/mediarep/3425>
- Van Le Q, Isbell, et al (2013) Pulvinar neurons reveal neurobiological evidence of past selection for rapid detection of snakes. *Proc Natl Acad Sci* 110(47):19000–19005. <https://doi.org/10.1073/pnas.1312648110>
- Yang J, Purves D (2004) The statistical structure of natural light patterns determines perceived light intensity. *Proc Natl Acad Sci* 101(23):8745–8750
- Zarkadakis G (2015) *In Our Own Image*. Rider Books, London

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Memo Akten's *Learning to See*: from machine vision to the machinic unconscious

Claudio Celis Bueno<sup>1</sup> · María Jesús Schultz Abarca<sup>2</sup>

Received: 6 July 2020 / Accepted: 18 August 2020 / Published online: 23 September 2020  
© Springer-Verlag London Ltd., part of Springer Nature 2020

## Abstract

This article uses Memo Akten's art installation *Learning to See* (Akten <https://www.memo.tv/portfolio/learning-to-see/>, 2017) to challenge the belief that machine learning and machine vision are neutral and objective technologies. Furthermore, this article follows Bernard Stiegler to contend that not only machine vision but also human vision is the result of constant training processes that rely directly on technology (understood as a technical surface of inscription). From this perspective, human vision is always already technical. Likewise, in an age dominated growingly by machine learning technologies, it is possible to speak not only of machine vision but also of a machinic imagination and a machinic unconscious, two notions that can be illustrated through Akten's art installation.

**Keywords** Machine vision · Machine bias · Bernard Stiegler · Algorithmic art · Black box

## 1 Introduction

The notion of vision machines was first introduced by Paul Virilio in the 1980s. According to him, the latest stage in the history of industrialisation would be the automation of visual perception. Following the shift in manual labour from the artisan to the assembly line, and the standardisation of intellectual activity by the culture industry, the rise of artificial vision would now “delegate the analysis of objective reality to a machine” (Virilio 1994: 59). Since the years Virilio anticipated this mutation, the development of machine learning algorithms has turned machine vision into a concrete technology with multiple real-world applications. Assembly lines, drones, logistics and transportation, CCTV, border control, are some of the domains where the automation of visual perception is being used to detach the production and circulation of commodities from human labour. Back in the nineteenth century, Marx forecast that modern industry would reduce the human worker to the role of a “watchman and regulator of the production process” (1973: 705). With

the automation of visual perception, however, even the task of supervising the production process is being delegated to machines, accelerating the exodus of living labour from the sphere of commodity production.

Echoing Virilio, Trevor Paglen (2016) has introduced the term “invisible images” to refer to those types of images “made by machines for other machines”. According to Paglen (2016), “the overwhelming majority” of images today belong to this invisible domain. Furthermore, the expansion of invisible images and vision machine technologies

is starting to have profound effects on human life, eclipsing even the rise of mass culture in the mid-twentieth century. Images have begun to intervene in everyday life, their function changing from representation and mediation, to activations, operations, and enforcement. Invisible images are actively watching us, poking and prodding, guiding our movements, inflicting pain and inducing pleasure (Paglen 2016).

As mentioned above, machine vision has been made possible mainly due to significant progress in the field of machine learning algorithms. The complexities of visual perception and image recognition make it almost impossible for rule-based algorithms to achieve these tasks since computer engineers would need to know in advance every possible visual input (Greenfield 2017: 214; Fry 2018: 11). Machine learning algorithms, instead, operate by identifying

---

✉ Claudio Celis Bueno  
ccelis@academia.cl

<sup>1</sup> Universidad Academia de Humanismo Cristiano, Santiago, Chile

<sup>2</sup> Universidad Adolfo Ibáñez, Santiago, Chile

complex patterns that are later used to match a given image to a specific label. These patterns, however, are not pre-defined by human programmers, but are instead produced through a training process in which thousand or even millions of labelled images are fed to an optimization algorithm that will progressively improve its efficiency by adjusting its weights (Greenfield 2017: 215). Put differently, “machine learning is the process by way of which algorithms are taught to recognise patterns in the world, through the automated analysis of very large data sets” (Greenfield 2017: 216).

This article uses Memo Akten’s art installation *Learning to See* (2017) to challenge the belief that machine learning and machine vision are neutral and objective technologies. Furthermore, this article follows Bernard Stiegler (2010, 2011, 2016) to contend that not only machine vision but also human vision is the result of constant training processes that rely directly on technology (understood as a technical surface of inscription). From this perspective, human vision is always already technical. Likewise, in an age dominated growingly by machine learning technologies, it is possible to speak not only of machine vision but also of a machinic imagination and a machinic unconscious, two notions that can be illustrated through Akten’s art installation.

## 2 Black boxes and biased machines

Two major issues in the field of machine learning are the ‘black-box effect’ and ‘machine bias’. These two issues are symptomatic of some of the political and ethical dilemmas of the several uses of machine vision. Given the scale of the training datasets and the speed of the training process, many machine learning algorithms become real “mysteries” to their programmers. This means that in many cases, the way these algorithms identify patterns and take decisions remains obscure to a human observer (Fry 2018: 11; Mordvintsev and Tyka 2015). As machine learning becomes “more complicated and their workings more inscrutable to users, it may become increasingly difficult to understand how autonomous systems arrive at their decisions” (Danks and London 2017: 4691). Frank Pasquale has forged the notion “black-box society” (2015) to emphasise how the contemporary world is relying more and more on algorithmic judgements that are extremely difficult to explain in human terms. This black-box effect is reinforced by corporate secret and intellectual property rights that keep algorithms undisclosed from public access. Several authors (Boulamwini and Gebru 2018; Danks and London 2017; Fry 2018; Pasquale 2015; Zarsky 2011) advocate for more transparency in the way machine learning algorithms function to avoid conflicts that could arise from the application of this technology by government agencies or private corporations. The European Union, for

example, produced in 2018 a *Data Protection Regulation* that includes not only the ‘right to be forgotten’ but also the ‘right to explanation’ whereby “a user can ask why an algorithmic decision was made about him or her” (Garcia 2016: 115). Nonetheless, the appeal to transparency is still limited by the issue of scale. Even if corporations and governments made their algorithms public, the scale and speed of the training process make it extremely difficult for a human observer to identify the logic behind the decision tree. Machine learning technologies are so complex and obtuse that in many cases, their mode of operation is not intelligible to human agents (Garcia 2016: 116; Goodman and Flaxman 2016).

The second issue is that of ‘machine bias’. Megan Garcia (2016: 112) defines this phenomenon as the process in which “seemingly innocuous programming takes on the prejudices either of its creators or the data it is fed”. According to her,

more and more computers are tasked with making crucial decisions, often on the basis of their perceived impartiality. For example, police use algorithms to target individual populations, and banks use them to approve loans. In both instances, computer results have been discriminatory—a reminder that learning how to account for algorithmic bias is increasingly important as more financial and legal decisions are driven by artificial intelligence (Garcia 2016: 113).

The issue of machine bias “is particularly worrisome for autonomous or semi-autonomous systems, as these need not involve a human being ‘in the loop’ (either active or passive) who can detect and compensate for biases in the algorithm or model” (Danks and London 2017: 4691). Since machine vision depends largely on machine learning algorithms, the issue of machine bias is a key aspect of the political and ethical questions surrounding the automation of perception. On the one hand, there is a group of authors who call for an improvement of the datasets to train less biased algorithms. For them, using a more representative training set would create more ‘objective’ algorithms. Boulamwini and Gebru (2018), for example, argue that “since computer vision technology is being utilised in high-stakes sectors such as health-care and law enforcement, more work needs to be done in benchmarking vision algorithms for various demographic and phenotypic groups” (Boulamwini and Gebru 2018). Similarly, Tal Zarsky (2011: 312) argues that machine bias is not an ‘inherent feature’ of algorithmic technologies, but is rather the result of ‘poor training data’. This means that actual cases of algorithmic bias can be solved by improving the training data and/or the training process. If data mining and training process are “sufficiently transparent”, Zarsky (2011: 312) argues, algorithmic bias can be “effectively overcome”. Furthermore, Zarsky claims that since a well-trained algorithm has the potential to become more objective

and less biased than human agents and institutions, this technology should be considered an essential aspect of fairer societies (2011: 311).

On the other hand, it could be argued that since machine learning algorithms are by default the result of a concrete training process, the possibility of an ‘objective judgement’ must be ruled out completely. It has been mentioned above that machine learning algorithms are taught to recognise patterns through a training process involving large datasets. According to some authors, class, gender and racial structures are transferred from the datasets to the algorithm, and in that process these algorithms are in fact “automating inequality” (Angwin et al. 2016; Eubanks 2018). Hence, critics of machine vision have highlighted that instead of mathematically neutral, these algorithms are ‘biased’ in their very constitution. From this perspective, the ideological justification of machine vision as a neutral technology could be seen as a reification and reinforcement of existing social asymmetries which are contained in the training datasets and codified into the algorithm. Trevor Paglen (2016), for example, contends that

there is a temptation to criticize algorithmic image operations on the basis that they are often “wrong”. These critiques are easy but misguided. They implicitly suggest that the problem is simply one of accuracy, to be solved by better training data. Eradicate bias from the training data, the logic goes, and algorithmic operations will be decidedly less racist than human–human interactions. Program the algorithm to see everyone equally and the humans they so lovingly oversee shall be equal.

Furthermore, Paglen (2016) adds that “ideology’s ultimate trick” is to “present itself as objective truth”, that is, to “present historical conditions as eternal” and “to present political formations as natural”. Automated systems tend to denote neutrality and objectivity, concealing that they may work as “immensely powerful levers of social regulation that serve specific race and class interests while presenting themselves as objective” (Paglen 2016). In this sense, machine vision should be understood as “a kind of hyper-ideology that is especially pernicious because it makes claims to objectivity and equality” (Paglen 2016).

The possibility of objective judgement has been a territory of vivid dispute since long before the appearance of machine learning and machine vision. According to Richard Rorty (1991), there are two main ways of defining objectivity: as correspondence or as agreement. He calls the first ‘realism’ and the second ‘solidarity’. In the case of objectivity as correspondence, a given statement is ‘objective’ if it is in conformity with ‘reality’. To achieve an objective judgement, then, a person must rule out any preconceived idea inherited from his or her community and attempt

to access ‘reality as such’. In the case of solidarity, truth appears not as a direct and unmediated relation to reality, but as an agreement between the members of a community. For Rorty, contemporary social sciences are the “heirs of the objectivist tradition, which centres around the assumption that we must step outside our community [and the opinion of its members] long enough to examine it in the light of something which transcends it” (1991: 22). The scientific search for ‘underlying structures’, ‘culturally invariant factors’, or ‘biologically determined patterns’ in society is an example of the predominance of the ‘realist’ conception of objectivity (Rorty 1991: 22).

Machine learning, through its immense capacity for pattern recognition, has become the latest instrument for pursuing pure objectivity. Apologists of this technology claim that thanks to machine learning and big data, theories will no longer be needed to explain the world (Anderson 2008). Instead, a real-time picture of reality will be permanently accessible offering “a whole new way of understanding the world” (Anderson 2008). From this perspective, algorithms will finally achieve the realists’ old dream of an unmediated image of the world free from preconceived ideas, opinions, and theories. The problem is that since machine learning depends directly on the training process, these preconceived assumptions are transferred from the datasets to the algorithm. In Rorty’s terms, it could be said that machine learning algorithms take one form of truth as ‘agreement’ and objectify it, presenting it no longer as agreement but as correspondence with ‘reality’. From a critical perspective, it must be emphasised that, at least in the case of machine learning, the ‘ruling out’ of prejudices that underlies the ‘realist’ conception of objectivity is structurally impossible since the system itself requires preconceived elements in order to train itself. Hence, a machine learning algorithm capable of ‘objective’ and ‘neutral’ judgements that ‘bracket off’ preconceived ideas would be a contradiction in terms.

### 3 Memo Akten’s *Learning to See*

*Learning to See* is an ongoing series of works based on machine learning technologies developed by artist Memo Akten. The series is composed of both an interactive art installation and a number of videos. This article will focus mainly on the interactive installation which was part of the 2019 Barbican exhibition “AI: More than Human”. It will also refer briefly to one of the videos from the series titled *We are Made of Star Dust #2*. According to Memo Akten (2017), the *Learning to See* interactive installation consists of “a number of neural networks” which “analyse a live camera feed pointing at a table covered in everyday objects” (Fig. 1). “Through a very tactile, hands-on experience”, Akten (2017) adds, “the audience can manipulate the

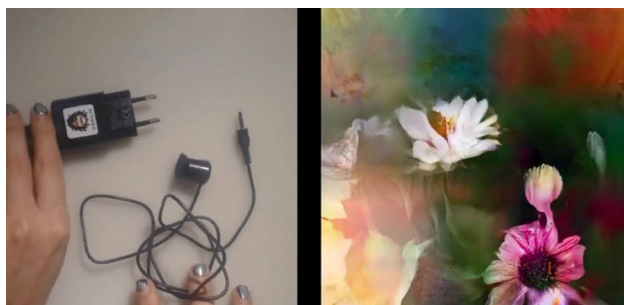




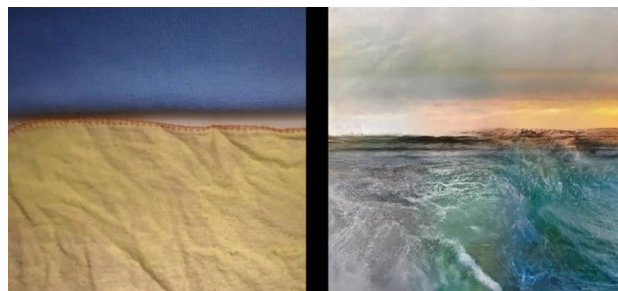
**Fig. 1** Memo Akten, *Learning to See* [Installation view], 2017 (<https://www.memo.tv/works/learning-to-see>)

objects on the table with their hands, and see corresponding scenery emerging on the display, in real-time, reinterpreted by the neural networks”.

Two images are projected on the wall of the gallery: on the left, a ‘live feed’ of the objects on the table as captured by the camera while, on the right, a ‘reinterpretation’ of the ‘live feed’ by a neural network (Figs. 2, 3). The image on the left comes across as ‘objective’, since it resembles what we have become accustomed to perceive through photo and video cameras. This effect of objectivity is enhanced by the real-time nature of the live feed: participants can identify



**Fig. 2** Memo Akten, *Learning to See* [Video view A], 2017 (<https://www.memo.tv/works/learning-to-see>)



**Fig. 3** Memo Akten, *Learning to See* [Video view B], 2017 (<https://www.memo.tv/works/learning-to-see>)

their own hands as they manipulate the objects on the table. The image on the right, instead, is a reinterpretation of the first image by machine learning algorithms trained with very specific datasets. Akten (2017) explains that

every thirty seconds the scene changes between different networks trained on five different datasets: the four natural elements: ocean and waves (representing ‘water’), clouds and sky (representing ‘air’), fire, flowers (representing earth and life); and images from the Hubble space telescope (representing the universe, cosmos, ether, void or God).

Every object that enters the camera’s visual field alters the output of both images. The participant’s hands manipulating the objects on the table are ‘seen’ by two different forms of machine vision: a traditional type of video image and a novel type of neural network image. The layout of these images—one next to the other—stresses how a single input produces two different outputs depending on the type of vision technology. This creates a tension between the ‘objective’ and ‘neutral’ image of the live feed on the left and the biased image on the right in which the algorithm can only see “through the filter of what it already knows” (Akten 2017). In this sense, Akten’s art installation functions as a concrete illustration of the core aspects of machine bias: by opposing an objective image to a biased one, *Learning to See* is addressing how the training datasets directly modify the output of the neural network.

One interesting aspect of this work is the fact that the objects on the table are objects of everyday life such as phone chargers, cleaning cloths, headphones, car keys, glasses, pens, USB cables, etc. All these objects are so ingrained in our daily life that they usually become invisible: they are not meant to be looked at, but rather to be used in an automatic and inattentive manner. They become visible only when their immediate use fails or is interrupted. In this sense, the objects selected by Akten in this artwork recall Martin Heidegger’s (2008) distinction between ‘present-at-hand’ [*vorhanden*] and ‘ready-to-hand’ [*zuhanden*].

The first term refers to a theoretical disposition in which one is detached from the object and can hence objectify it as an object of contemplation, scientific observation, and thorough analysis. For Heidegger, Western science and philosophy have been grounded mainly on this form of relation to non-human beings. The second refers to a more primordial relation to the world which is previous to the process of objectification. The ‘ready-to-hand’ relation is exemplified in our day-to-day use of tools and devices that make up our most immediate world. As Hubert Dreyfus (1991: 61) puts it, Heidegger “proposes to demonstrate that the situated use of equipment is in some sense prior to just looking at things and that what is revealed by use is ontologically more fundamental than the substances with determinate, context-free properties revealed by detached contemplation”. Since this relation to the world is prior to the theoretical disposition that objectifies it, Heidegger claims that ‘ready-to-handness’ should become a methodological guideline and the starting point for a new phenomenological philosophy. In *Learning to See*, Akten chooses objects that we commonly relate to as being ‘ready-to-hand’: not objects of contemplation but objects of daily use that remain ‘hidden’ behind our ordinary relation to them. A cable becomes visible when it gets tangled, a cleaning cloth when instead of cleaning it makes something dirty, a pen when it stops writing. Even most significant is the example of glasses, that specific piece of equipment which allows seeing other objects as long as it remains invisible. Akten’s art installation collects all these ‘invisible objects’ and makes them visible through a very particular vision machine that shifts our ‘ready-to-hand’ relation to them towards the contemplative domain of the ‘presence-at-hand’.

#### 4 Immersion or coexistence?

A first impression may suggest that Akten’s interactive art installation produces an immersive experience. The piece is set up in a dark room with just three sources of light: the table with the objects and the two projected screens. These light sources work as attention points that stand out from the fading background. The participant can, hence, transit between three different experiences: the ‘direct’ perception and manipulation of the objects on the table, the contemplation of those objects on the live feed camera projected on the left screen, and the interactive construction of images on the right screen. The room’s darkness and the interactivity between the table and the images seem to reinforce a sense of immersion in which reality is blurred out or ‘parenthesised’. In particular, the relation between the direct manipulation of objects on the table and the construction of interactive algorithmic images on the right screen could be seen as a sort of ‘live-painting’ software that stimulates an immersive

experience. Immersion can be described as the situation in which one perceptive experience is replaced by another one (Boyer 2015). Pure immersion, if possible, requires the complete replacement of one for the other. Accordingly, any conflict between the two would interrupt the immersive experience. In the case of *Learning to See*, it could be suggested that had Akten just kept the table and the neural network screen on the right, the installation would have offered the conditions of possibility for an immersive experience in which the neural network image, capturing the participant’s gaze, replaces his or her direct perception of the objects on the table. However, Akten opted for a different format which included also the live feed from the camera, hence halting the immersive experience. Unlike a videogame or virtual reality software, Akten’s art installation does not simply aim at a pure immersive experience. Instead, the piece contrasts both images projected on the wall and interrupts the possibility of immersion by claiming back the participant’s attention. Rather than an immersive experience, the installation’s set up accentuates an experience of coexistence in which one perceptual content does not substitute another: in *Learning to See* the three described experiences coexist and compete for the participant’s attention.

According to Edmund Husserl (1983), the correct phenomenological method requires a process of ‘reduction’ [*epoché*], that is, a type of ‘parenthesizing’ or ‘bracketing off’ of everything that does not belong to ‘originary perception’. This phenomenological reduction is the condition of possibility for an analysis of a pure act of perception that is not contaminated by cultural prejudice and, therefore, the condition of possibility for objective judgement. This led Husserl to distinguish between “simple founding acts” and “complex founded acts” (1983: 277). The former refers to an experience of pure, originary, perception. It is a founding act since all other forms of experience are grounded on it. The latter refers to secondary experiences that are founded on primary perception. Founded acts include experiences of representation such as memory and imagination, as well as any form of relation to external representations (images, paintings, screens, etc.). From a Husserlian perspective, *Learning to See* offers both types of experiences: a founding act of perception of the objects on the table and two founded acts of vision of the images projected on the wall. Furthermore, the image on the left gives the impression of objectivity since it is closer to the originary perception of the objects on the table, while the image on the right gives the impression of artificiality since it diverges from and transforms it. Furthermore, Akten’s algorithm can be considered ‘biased’ for the simple fact that it relies on memory (training datasets) and is structurally incapable of ‘parenthesising’ this founded act.

Once again, the question of objectivity is reduced to the act of bracketing off perception from preconceived ideas.

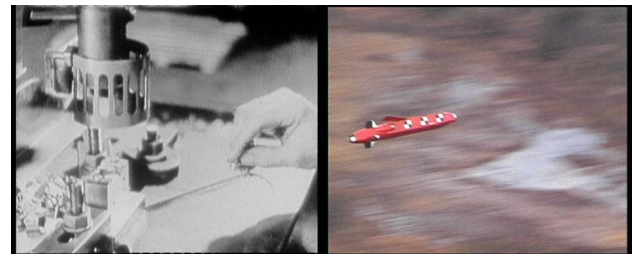
The fact that the image on the left comes across as ‘objective’ implies that it is free from preconceived ideas, opinions and beliefs. It is closer to the founding act of perception. The image on the right, instead, comes across as biased because of its inability to bracket off the restricted dataset it was trained with. Nonetheless, this reading tends to naturalise our relation to the live video camera. Akten’s installation, through the coexistence of three different experiences of the same objects, aims not only at a critique of computer vision. It also attempts halting the ‘ready-to-hand’ relation to technical images and digital technologies (which has become so naturalised by our everyday use of them). By interrupting this initial relation to the image on the left (just like it interrupts our usual relation to the objects chosen), *Learning to See* challenges its alleged ‘objectivity’ in opposition to the ‘artificiality’ of the one on the right, putting into question the sheer possibility of a non-biased judgement.

## 5 Soft-montage and technical images

The experience of coexistence is achieved in Akten’s art installation by placing the two screens side by side. This layout reminds us of Harun Farocki’s ‘soft-montages’ (2009). Used first in Farocki’s 1995 video installation *Interface* (Fig. 4), the double projection was meant to produce a two-fold sense of simultaneity and succession: “the relationship of an image to the one that follows as well as the one beside it; a relationship to the preceding as well as to the concurrent one” (Farocki 2009: 71). Soft-montage aims at an experience of coexistence in which one image operates as the commentary or interruption of the other. Between 2000 and 2003, Farocki used the soft-montage technique in his *Eye/Machine* series (Fig. 5). Inspired by the work of Paul Virilio, these video installations explore the birth of machine vision. Farocki (2009: 72) puts forth the thesis that the automation of visual perception for industrial and military purposes was first achieved by a technical system that compared a pre-defined template (what he calls a ‘pre-image’) to an actual



**Fig. 4** Harun Farocki, *Interface*, 1995 (<https://www.harunfarocki.de/installations/1995.html>)



**Fig. 5** Harun Farocki, *Eye Machine*, 2000 (<https://www.harunfarocki.de/installations/2000s/2000/eye-machine.html>)

image provided by a live camera feed. In this primitive form of machine vision, “contours and significant details” were stored in the autonomous system and then “compared to with the actual item” (Farocki 2009: 72). Farocki (2009: 74) writes that one of the key concerns when creating this art installation was that the “mode of presentation” (soft-montage) would be “justified by the subject matter itself” (machine vision). For him, the double projection in his soft-montages replicated the act of comparison that took place in the first forms of machine vision between a ‘pre-image’ and an ‘actual image’. This was a primitive form of machine vision in which a stored pattern was used to classify images from the real world. The difference with contemporary forms of machine vision is that the pattern had to be defined in advanced by a programmer, which meant that it could only work for basic shapes in controlled environments. In any case, the connection between the mode of presentation (soft-montage) and the subject matter (machine vision) in Farocki’s *Eye/Machine* series is replicated in Akten’s *Learning to See*.

In Akten’s art installation, the comparison between the two screens reinforces reading both images projected on the wall as the result of a technical process. Hence, the image on the left should not be perceived as a ‘natural’ image opposed to the ‘artificial’ image on the right. Instead, by placing them side to side, *Learning to See* reveals that both images share the same technical input (a video camera placed on top of the table) and output (a light projection on the wall). Additionally, both images appear as the result of a specific technical process: digital video coding on the left and algorithmic video processing on the right. Like Farocki’s soft-montage in the *Eye/Machine* series, the double projection works as a metaphor of the inner workings of machine vision. To fully grasp how *Learning to See* deploys a critique of machine vision in both screens and, hence, prevents regarding the image on the left as ‘objective’, it is useful to recur to what Vilém Flusser (2000) has called ‘technical images’.

In his *Philosophy of Photography*, Flusser defines the technical image as “an image produced by an apparatus” (2000: 14). According to him, traditional images were produced in the imagination of the artist, who transferred

them by means of the tool (paintbrush, chisel, etc.) onto a material surface (2000: 15). The invention of the technical image, instead, involved an externalisation of the faculty of imagination. In this process, the capacity to produce images is transferred from the black box of the imagination to the black box of the apparatus (2000: 16). Additionally, Flusser claims that an apparatus is defined by its programme, that is, by the set of instructions that, given a certain input, produce a specific output (2000: 26). This defines a finite number of outputs that each apparatus can produce. Every time an apparatus is put to work, it actualises the programme contained in it and reduces by one the amount of possible outputs (2000: 26). Photography, film, video and digital images all belong to the category of ‘technical images’ as defined by Flusser. With the massification of these technologies, human beings are becoming more and more dependent on apparatuses. This phenomenon, in turn, is demanding a new philosophy of the technical image (or a “philosophy of the black box” as Flusser sometimes defined it). To fully grasp the significance of technical images, this new philosophy must begin by opening up the apparatus’ black box and ‘elucidate’ the inner working of its programme (Flusser 2000: 16).

Traditionally, technical images were considered objective because they were produced by an apparatus that captured the world without interference from a human hand (Zylinska 2017: 70). It could be argued, however, that by removing the human from the process, technical images overcome the representationalist domain in which the question of objectivity can be even posed. Joanna Zylinska (2017) suggests that technical images are the result of a non-human process that bypasses the mediation between a subject and the world. As it was mentioned, objectivity depends on the possibility of either correspondence or agreement. In both cases, representation works as an element of human mediation: either between the subject and reality, or between the subject and the community. Once the representationalist perspective is replaced by the non-human perspective of technical images, however, the question of objectivity can no longer be posed and must be replaced by the non-human problem of effectivity.

Akten’s art installation is a hands-on exploration of the black box behind machine vision. Specifically, this piece highlights how machine vision relies on training datasets to interpret the world. According to Akten (2019), “the work exposes and amplifies the learned bias in artificial neural networks, demonstrating how critical the training data is to the predictions that the model will make” (Akten 2019). This experience is emphasised by the fact that the trained algorithm changes every thirty seconds, showing how the same input generates entirely different outputs depending on the training datasets. Hence, “when the trained network looks out into the world via the camera, it can only see what it already knows” (Akten 2019).

Machine bias is, in this sense, essential to machine vision. At the same time, however, Flusser’s notion of technical image makes it possible to suggest that since both images in *Learning to See* are technical, and since technical images demand moving away from a representationalist approach, it is no longer possible to evaluate either of them using the distinction between ‘objectivity’ and ‘bias’.

## 6 Inceptionism

It is important to point out that the technique employed by Akten is different to that of style transfer, “in which the ‘style’ of one image is transferred onto another image, which provides the ‘content’” (Akten 2019). Style transfer produces “an output image via an optimization process”, while *Learning to See* uses a neural network that produces “a predictive model” (Akten 2019). This predictive model “contains knowledge of the entire dataset, hundreds of thousands of images” (Akten 2017).

Akten (2015) learns about the difference between style transfer and predictive algorithms through the analysis of Google’s *Deep Dream* (Fig. 6). According to Alexander Mordvintsev and Mike Tyka (2015), “one of the challenges of neural networks is understanding what exactly goes on at each layer”. With this aim in mind, these Google engineers created *Deep Dream*, and application which inverts the image recognition algorithm in order to see its internal operations. As Mordvintsev and Tyka (2015) put it,

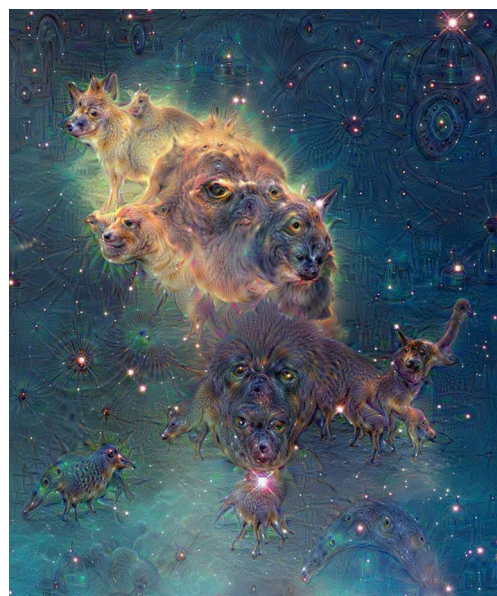


Fig. 6 Google’s Deep Dream ([https://www.vice.com/en\\_us/article/zngm94/no-they-dream-of-puppy-slugs-0000703-v22n8](https://www.vice.com/en_us/article/zngm94/no-they-dream-of-puppy-slugs-0000703-v22n8))

One way to visualise what goes on is to turn the network upside down and to ask it to enhance an input image in such a way as to elicit a particular interpretation. Say you want to know what sort of image would result in ‘banana’. Start with an image full of random noise, then gradually tweak the image towards what the neural net considers a banana [...] We call this technique ‘inceptionism’.

Considering that Google’s search engine is a constant flow of nearly six billion searches per day around the world, Google’s *Deep Dream* can be literally read as the dreaming machine for the world’s social unconscious. As Hito Steyerl puts it, “in a feat of genius, inceptionism manages to visualize the unconscious of prosumer networks: images surveilling users, constantly registering their eye movements, behaviour, preferences” (Steyerl 2017: 57). In this context, she adds, “Walter Benjamin’s ‘optical unconscious’ has been upgraded to the unconscious of computational image divination” (Steyerl 2017: 57). Back in 1930, Benjamin suggested that photography and cinema were optical devices capable of rendering visible previously unconscious elements, just like psychoanalytic techniques would allow objectifying unconscious elements of the psyche (2008: 287–79). Machine vision, likewise, can help rendering visible a new unconscious: that of the social structures engrained on the training dataset. If one overcomes the illusion of objectivity, machine vision can be read symptomatically as the ‘index’ of an unconscious content present on the dataset.

In 2015 Akten wrote an article on Google’s *Deep Dream* where he described it as “mind-blowing”. In particular, he was fascinated by how ‘inceptionism’ would iterate the algorithm’s initial intuition to explore the neural network’s hidden layers:

What was an initial “maybe I see inklings of little lizard-like features over here” on a deep sub-conscious level, starts to become “yea, I think that might be lizard-like features”, to “oh definitely, that’s a lizard-skin puppy-slug” at a well-defined, visible high level. These activations are now strong enough not to dissipate and disappear in the depths of the network, and can propagate to higher levels, potentially even affecting the final output or decision. (Akten 2015)

By iterating seemingly insignificant elements, Google’s *Deep Dream* algorithm

creates a positive feedback loop, reinforcing the bias in the system. Building confidence with each iteration. Transforming what was subtle, unnoticeable trends deep within the network, to strong, visible, defining biases that affect the decisions of the network. (Akten 2015)

The issue is that previous to the emergence of machine learning technologies, this unconscious content hidden in the dataset would remain invisible, just like the optical unconscious remained concealed to the naked eye before the invention of photography.

In the case of *Learning to See*, the unconscious content of machine vision is made visible by explicitly narrowing the training dataset. Furthermore, it could be argued that *Learning to See* uses the technique of soft-montage to contrast two different forms of machinic unconscious: an optical unconscious on the left and an algorithmic unconscious on the right. In this sense, both images appear as the result of a machinic imagination, that is, a machinic black box responsible for the production of images. This becomes more evident in a video version of *Learning to See* titled *We Are Made of Star Dust #2* (Fig. 7). In this version, the screen on the left shows a high-definition and highly detailed close-up of the artists’s face captured with a microscopic lens. The screen on the right, instead, shows a reinterpretation of these microscopic images by a neural network trained exclusively with images produced by the Hubble telescope. This creates an interaction between the microscopic images of the artist’s face and the telescopic images of the universe. Microscope and telescope are two concrete examples of what Benjamin would call, like photography and cinema, technical sources of an ‘optical unconscious’. This optical unconscious is contrasted with the algorithmic unconscious rendered visible by the image on the right. An analogy, emphasised by the title of the artwork, is then forged between the microscopic particles that compose our body and those that compose the universe. The soft-montage technique in this piece reinforces this idea.

## 7 The technicity of human perception

But *Learning to See*, Akten (2017, 2019) tells us, is not only about how “an artificial neural network looks out onto the world, trying to make sense of what it’s seeing, in context

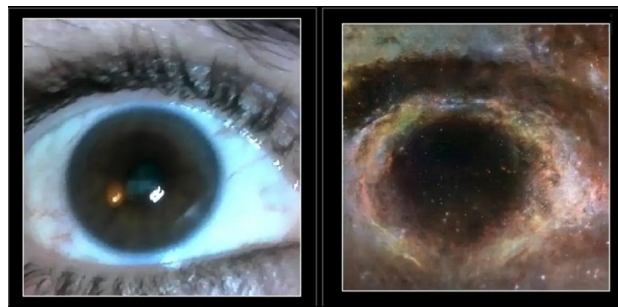


Fig. 7 Memo Akten, *Learning to See* [Video view C], 2017 (<https://www.memo.tv/works/learning-to-see>)

of what it's seen before" (Akten 2017). *Learning to See* is also a reflection on human vision, calling into question how humans "may construct meaning" (Akten 2019). Just like machine learning algorithms "which can only see through the filter of what they already know", human vision looks at things not as they are, but as the result of who "we are" (Akten 2017). In Akten's words, *Learning to See* uses machine learning technologies

to reflect on ourselves and how we make sense of the world. The picture we see in our conscious mind is not a mirror image of the outside world, but is a reconstruction based on our expectations and prior beliefs. (Akten 2017)

This means that not only the projected screens are the outcome of a technical imagination and a technical unconscious. It means that also the participant's 'direct' perception of the objects on the table should be understood as the result of a training process that is intrinsically technical.

As mentioned above, the fact that Akten decided to keep both screens (and not just the one with the neural network) can be read as a sign that his main aim was not to generate an effect of immersion, but rather to create a contrast between both images. This contrast is mirrored by the participant's experience who does not look at the objects on the table through his or her 'naked eyes', but rather through the mediation of two types of technical images. This can be read as a metaphor for the regular experience of perception: instead of a pure and naked relation to reality, perception is always mediated by the technical apparatuses that compose our relation to the world. This implies that the impossibility of an 'objective' relation to the world (free from inherited prejudices) applies not only for machine vision but also for human vision. Put differently, if we accept Akten's (2017) claim that *Learning to See* is a reflection "on how we make sense of the world", then it must be deduced that, like machine vision, human vision is intrinsically technical, traversed by a machinic faculty of imagination and a machinic unconscious.

At this point, Bernard Stiegler's theory of "originary technicity" (2010,2011,2016) becomes a useful framework to advance the analysis of *Learning to See*. According to Stiegler (2011: 60), technics (from the most primitive tool to the latest smartphone) function as a surface of inscription through which individuals inherit a specific culture and a specific relation to the world. Humans are, thus, constituted through a "process of exteriorization" in which "intergenerational support of memory [material culture] overdetermines learning and mnemonic activities" (Stiegler 2010: 9). This means that technics 'always already' precede the constitution of human perception and human memory. Stiegler uses Husserl's notions of primary and secondary retention in order to argue that human perception and human memory

are constantly shaped by these external surfaces of inscription, or tertiary retentions (2010: 8, 2011: 4, 2016: 480). Technics, as surfaces of inscription, modify "the relations between the psychic retentions of perception, which Husserl referred to as primary retentions, and the psychic retention of memory, which he called secondary retentions" (Stiegler 2016: 480). And since technics evolve over time, so does "the play between primary retention and secondary retention" (Stiegler 2016: 480). In other words, with the advent of each new technical surface of inscription comes a new organisation of our perception, our memory and our imagination. The history of media is both "the history of technics" and the history of the organisation of an individual's "interiority" (Stiegler 2011: 78).

Stiegler replaces Husserl's phenomenological analysis of an 'originary perception' for a study of the 'originary technicity' that constitutes our perception, memory and imagination. Following Jacques Derrida (1973: 45), Stiegler challenges the radical distinction ('heterogeneity') between perception and imagination in Husserl. The external surface of inscription ('the trace') works as an 'originary technicity' which, as Derrida (1978: 203) puts it, simultaneously "erases the myth of a present origin". Likewise, *Learning to See* erases the myth of a pure perception as the basis for objective judgement. In doing so, it challenges the opposition between natural and artificial perception. To rephrase Stiegler, it could be said that in *Learning to See* 'my vision' is always 'that of others', since it is never original or authentic, but rather inherited through external memory supports. This means that vision can never be detached from a non-lived past. Hence vision is always biased. From the perspective of Stiegler, there is no natural perception since perception is always already determined by the technical surfaces of inscription that constitute our relation to the world. As he puts it, if we accept that "lived reality is always a construct of the imagination and thus perceived only on condition of being fictional, irreducibly haunted by phantasms" then we are "forced to conclude that perception is subordinated to the imagination" (Stiegler 2011: 16). This means that there is "no perception outside imagination" and that perception is "imagination's projection screen" (Stiegler 2011: 16). Objective judgements are, thus, impossible, since individuals always look at the world through eyes that have been mediated by technical devices. The 'bracketing off' of inherited prejudice to perceive reality 'in itself' is an illusion that conceals the fact that technology permanently modifies our internal senses of perception and memory. Naked human vision too is always already machine vision. Human vision, like machinic vision, depends on the surfaces of inscription that function as an external faculty of imagination. In *Learning to See* the participant perceives his or her hands manipulating the objects on the table mediated by the two types of technical images that capture his or her gaze. These

technical images function as an external faculty of imagination that mediate his or her specific perception of reality. The soft-montage in *Learning to See* functions as perception's prosthesis: it opens up the black box of machine vision and renders visible the invisible process through which external surfaces of inscription constantly shape human vision. Since human vision is always technical, it is always already biased.

In relation to the specific case of machine learning algorithms, Stiegler (2016: 480) suggests that this technology is putting forth a systematic automation of the "faculty of discernment" (or faculty of imagination). Traditionally, the faculty of imagination is conceived as the human capacity to connect perception and memory. For Stiegler, instead, the imagination functions as a "post-production centre", a "control room" that assembles "the montage, the staging, the realisation and the direction, of the flow of primary, secondary, and tertiary retentions" (2011: 28). With the advent of machine learning algorithms, however, the montage of perception and memory is being delegated to an external faculty of imagination, just like the production of technical images was previously delegated to apparatuses (Flusser 2000). Furthermore, Stiegler adds that since tertiary retention constitute an external memory that precedes interiority, technical surfaces of inscription should be thought of as an unconscious, "if not *the* unconscious" (2010: 8, 2011: 17). Yet, Stiegler's use of the notion of the unconscious should not be understood metaphorically. Technical surfaces of inscription are, in fact, an (external) domain that defines our conscious capacities (interiority), hence halting the possibility of a non-biased, neutral perception. *Learning to See* highlights a profound connection between machinic vision, machinic imagination and machinic unconscious. For Stiegler (2011: 17), the unconscious needs to be understood more generally as a technical surface of inscription. This argument replicates that of Derrida (1978) which regards the unconscious as writing. Freud's 'psychic unconscious', Benjamin's 'optical unconscious', and Akten's 'algorithmic unconscious' are all result of a process of externalisation through which invisible structures become visible. Clint Burnham (2018: 9), for example, claims that since the internet "contains what we forget, or do not know that we know", it can tell us something about the unconscious in the contemporary world. This becomes clear in the case of Google's *Deep Dream*, in which Steyerl (2017: 57) has identified a concrete potential to visualise the unconscious of search engines and social media through a process of iteration. Likewise, it was mentioned that *Learning to See* unveils the algorithmic unconscious of neural networks by limiting its training dataset.

In all these examples, an external memory functions as the unconscious content that shapes individual perception and individual imagination. The apparatus's black box is not an externalisation of an internal capacity, but rather the constitutive element of any interiority. This implies

that human vision is always already technical and, as such, always already biased since it cannot bracket off the unconscious elements that structure it. The claim for algorithmic neutrality is a strong ideology that conceals the structuring character of technicity. Acknowledging this is a key step towards defining the politics of machine vision and hence developing, as Stiegler puts it, a "new political economy of consciousness" (2011: 39). According to Akten (2019), "in an increasingly polarised and divided society, we believe in the significance of trying to be sensitive to the idea that we ourselves may be as biased as one of these artificial neural networks, seeing the world through the filter of our own life experiences". *Learning to See* provides a first step towards this significant and necessary task.

**Funding** This article was supported by CONICYT (Chile) under the Grant Fondecyt No. 11170065.

## References

- Akten M (2015) Deepdream is blowing my mind. Medium. <https://medium.com/@memoakten/deepdream-is-blowing-my-mind-6a2c8669c698>. Accessed 1 July 2020
- Akten M (2017) Learning to see. <https://www.memo.tv/portfolio/learning-to-see/>. Accessed 1 July 2020
- Akten M (2019) Learning to see: you are what you see. ACM SIGGRAPH 2019 Art Gallery. [https://ualresearchonline.arts.ac.uk/15061/1/LearningToSee\\_AuthorVersion.pdf](https://ualresearchonline.arts.ac.uk/15061/1/LearningToSee_AuthorVersion.pdf). Accessed 1 July 2020
- Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete, wired. <https://www.wired.com/2008/06/pb-theory/>. Accessed 1 July 2020
- Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. ProPublica.org. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. Accessed 1 July 2020
- Benjamin W (2008) Little history of photography. In: Jennings M, Doherty B, Levin T (eds) The work of art in the age of its technological reproducibility and other writings on media. Harvard University Press, London, pp 274–298
- Boyer E (2015) Le Conflit des perceptions. éditions MF, Paris
- Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Paper presented at the 1st conference on fairness, accountability and transparency, New York. <https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>
- Burnham C (2018) Does the internet have an unconscious?. Bloomsbury Academic, New York
- Danks D, London AJ (2017) Algorithmic bias in autonomous systems. In: Paper presented at the twenty-sixth international joint conference on artificial intelligence, Melbourne. <https://www.cmu.edu/dietrich/philosophy/docs/london/IJCAI17-AlgorithmicBias-Distrib.pdf>
- Derrida J (1973) Speech and phenomena. Northwestern University Press, Evanston
- Derrida J (1978) Freud and the scene of writing, writing and difference. The University of Chicago Press, Chicago, pp 196–231
- Dreyfus H (1991) Being-in-the-world. The MIT Press, London

- Eubanks V (2018) Automating inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Press, New York
- Farocki H (2009) Cross influence/soft montage. In: Ehmann A, Eshun K (eds) Harun Farocki: Against what? Against whom?. Koenig Books, London, pp 69–74
- Flusser V (2000) Towards a philosophy of photography. Reaktion Books, London
- Fry H (2018) Hello world: how to be human in the age of the machine. W. W. Norton & Company, New York
- Garcia M (2016) Racist in the machine: the disturbing implications of algorithm bias. *World Policy J* 33(4):111–117
- Goodman B, Flaxman S (2016) EU regulations on algorithmic decision-making and a “right to explanation”. In: Paper presented at the ICML workshop on human interpretability in machine learning, New York. <https://arxiv.org/abs/1606.08813>
- Greenfield A (2017) Radical technologies: the design of everyday life. Verso, London
- Heidegger M (2008) Being and time. HarperCollins, New York
- Husserl E (1983) Ideas pertaining to a pure phenomenology and to a phenomenological philosophy. MartinusNijhoff Publishers, The Hague
- Marx K (1973) Grundrisse: foundations of the critique of political economy (Nicolaus M, Trans.). Random House; Penguin, New York
- Mordvintsev A, Tyka M (2015) Inceptionism: going deeper into neural networks. Google AI Blog, 2019. <https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>. Accessed 1 July 2020
- Paglen T (2016) A study of invisible images. <https://www.metropictures.com/exhibitions/trevor-paglen4/press-release>(press release). Accessed 1 July 2020
- Pasquale F (2015) The black box society. Harvard University Press, Cambridge
- Rorty R (1991) Solidarity or objectivity? Objectivity, relativism, and truth. Cambridge University Press, New York
- Steyerl H (2017) A sea of data: apophenia and pattern (mis)recognition. Duty free art: arte in the age of planetary civil war. Verso, London
- Stiegler B (2010) For a new critique of political economy. Wiley, Hoboken
- Stiegler B (2011) Technics and time, 3: cinematic time and the question of malaise. Stanford University Press, Stanford
- Stiegler B (2016) Ars and organological inventions in societies of hyper-control. *Leonardo* 49(5):480–484
- Virilio P (1994) The vision machine. Indiana University Press, Indianapolis
- Zarsky T (2011) Governmental data mining and its alternatives. *Penn State Law Rev* 116(2):285–330
- Zylinska J (2017) Nonhuman PHOTOGRAPHY. The MIT Press, Cambridge

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.





# Crossroads of seeing: about layers in painting and superimposition in Augmented Reality

Manuel van der Veen<sup>1</sup>

Received: 31 July 2019 / Accepted: 18 August 2020 / Published online: 26 September 2020  
© The Author(s) 2020

## Abstract

Augmented Reality (AR) is itself a technology in which two ways of seeing are crossed. Our field of vision is thereby superimposed with digital information and images. But before this, the real environment is already perceived by machine seeing, it is redoubled by a 3D-model, scanned, located and linked. In this brief investigation, I will face the way of seeing in AR with traditional procedures, like ‘trompe-l’œil’ and the so-called ‘velo’, to distinguish between what remains classic and what has changed. It is important to examine this as layering, because it is a very thin stack of techniques, technology, materials and media, we seek to watch through. Subsequently, I shall analyze a painting of the contemporary artist Laura Owens in which both ways are crossed, the traditional one and the one concerning AR.

**Keywords** Augmented Reality · (Digital) layering · Stack · Superimposition · Transparency · Trompe-l’œil · Velo (veil)

## 1 Introduction

The field of vision in Augmented Reality (AR) challenges our way of seeing by registering digital images and objects onto the real environment. Sometimes these images and objects emerge as registered and sometimes, they blur the boundaries between the digital and the real area. It is my approach to face AR with its cognates in art history to sort out the specific strategies and procedures of layering. It is not the goal to prove some continuous development from ancient illusion techniques to newer technologies. Rather, the new should be divorced from the already known to examine our current way of seeing in AR more profoundly. The view through the glasses of AR enables a new perspective onto the tradition. The technology merges various techniques and procedures; in particular, it crosses two ways of seeing: our view through the eyes and that of machine seeing. The latter processes the data, received by the sensors and cameras, within our field of vision to calculate a hybrid view. However, this leads into a double blindness, as each participant is blind to the other for a certain extent of the way.

In general, AR may be defined as an operation of superimposition. It overlays the real environment and one has to perceive them together. For this, there are various scenarios in earlier procedures. For example, Filippo Brunelleschi’s two experiments in front of the Baptistery San Giovanni and at the Piazza della Signoria, placing cut-out paintings surrounded by the movement of clouds or the living city. We must also think of the phantasmagoria, within which ghosts are projected into the real space and in real time (Elcott 2016). Furthermore, the schüfftan-process, perhaps less common, also establishes an interplay of real fragments with illusionistic complements. Finally, there are analogue panoramic boards that provide information about the location, at the location. All of these are relatives which are not the same, but are particularly suited to highlight differences.

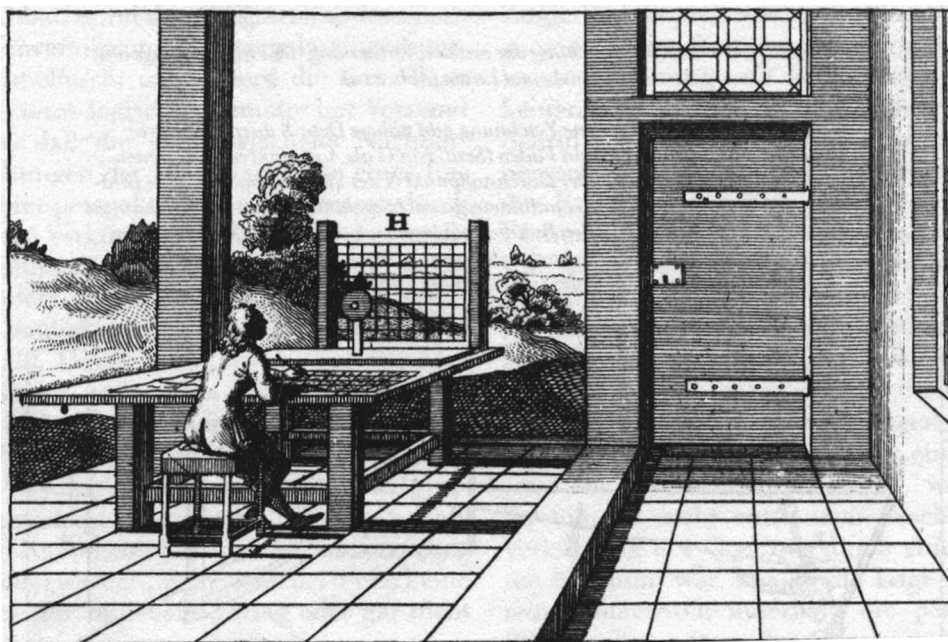
In this text, I would like to focus particularly on the layer of superimposition itself. It is placed between an imaginary and a real interface. To look at this at least semitransparent layer I suggest to investigate Leon Battista Alberti’s “how-to-do-it apparatus”, the so-called ‘velo’ (veil).

---

✉ Manuel van der Veen  
M.vanderVeen@me.com

<sup>1</sup> State Academy of Fine Arts Karlsruhe, Karlsruhe, Germany

**Fig. 1** Device for perspectival drawing of a landscape. Anonym, 1710. PD-Art/PD-old-100=/1810/?; PD-US



## 2 Metaphor and apparatus: about the ‘velo’ as a layering procedure

To analyze the ‘velo’ as an apparatus depends on the observation that the device itself establishes a way of seeing. As noted above, AR tends to blur the boundaries between digital images and the real surroundings. In contrast to the career of immersion into the image, which is generally associated with virtual reality, in AR immersive images are placed. Their status as images is covered to appear as a part of reality, which connects the current view to the traditional procedure of ‘trompe-l’œil’. However, it has not yet been decided whether AR will pursue an all over camouflage. My research project ‘Augmented Reality. Trompe l’œil and Relief as Technique and Theory’, of which the considerations here form an excerpt, suggests to describe ‘trompe-l’œil’ and sculptural relief as immersive and emersive<sup>1</sup> images. Since the ‘velo’ provides a layer to think the technological implications of AR and at the same time plays an important part in the interpretation of ‘trompe-l’œil’, it is particularly suitable for the following investigation.

In 1435 Alberti wrote in his treatise on painting and perspective ‘De pictura’ that the rectangular frame of paintings should be seen as an open window (‘aperta finestra’). This very well-known section from ‘De pictura’ also carries well-known difficulties. First, the solid and opaque surface of the canvas is denied. And second, what is seen through a window is the here and now; a painting of the fifteenth and

sixteenth centuries instead usually shows a somewhere and a sometime else of an ‘istoria’. The extensive discussion of the window metaphor cannot be pursued further here. For now, it seems more interesting to look up another passage from Alberti’s treatise in book two. Here, he does not only describe a metaphor, but also an actual apparatus, a device. This apparatus stretches out a semi-transparent cloth with a grid of threads as a layer between the artist and the motif (Fig. 1). It is precisely this cloth, which I proposed as a relative of the information and object layer, that we are dealing with in AR.

The technology circulates as a procedure, which places figures in the environment, as in the popular application ‘Pokémon Go’ (Niantic). But even before a figure, an image or an object occupies our field of vision, the technology perceives the environment, doubles or multiplies the pair of eyes through cameras and sensors to set up a (almost invisible) layer through which we perceive the environment. It might be helpful to quote Edmund Husserl’s ternary image theory, which consists of three layers of seeing: physical image—image object—image subject (Husserl 2005: 20). The physical image is for example the carrier, the physical support. The images of AR and of ‘trompe-l’œil’ are problematic as such because they seem to have no carrier.<sup>2</sup>

Due to that lost carrier, the layer of projection and the performance of the machine are suppressed. And with them, the place is blurred where the two ways of seeing, of

<sup>1</sup> Thanks to a productive conversation with Matthias Bruhn, about emersion and emersiv images, the parallel to the procedure of sculptural relief could be drawn.

<sup>2</sup> The missed physical image in Augmented Reality in relation to Husserl’s image theory was mentioned during a presentation by Stephan Günzel in Weimar, Germany (Das Diorama: Durch...Denken) called ‘Augmented Reality: Zur (In)Transparenz des Bildes’.

machine and body, are crossed. Everything that becomes visible through the glasses of AR is already processed. Husserl's image theory is based on the difference between image and environment, the difference that is at stake in AR and 'trompe-l'œil'. Thus, for Husserl, the image manifests itself in a conflict and this conflict is not caused by the realistic depiction (Husserl 2005: 51): "The appearance belonging to the image object is distinguished in one point from the normal perceptual appearance. This is an essential point that makes it impossible for us to view the appearance belonging to the image object as a normal perception: it bears within itself the characteristic of unreality, of conflict with the actual present. The perception of the surroundings, the perception in which the actual present becomes constituted for us, continues on through the frame and then signifies 'printed paper' or 'painted canvas.'" If the carrier is lost, the conflict with the actual present disappears, which in turn leads to an uncertain image perception in AR.

The motivation to look out for the carrier is not based on its status of being lost or hidden, which would end up in chasing a deception. Rather, it is due to the observation that the carrier is shifted instead. A 'trompe-l'œil' occupies a classical support, but only to be its alienation, because the procedure stacks other depicted supports onto the image support (e.g. planks of wood, papers, canvases, etc.). With each depiction, the physical support is pushed forward piece by piece (Fig. 2). Therefore, we could record that the carrier is suppressed by both, its multiplication and by its disappearance.

In AR a carrier also exists as a (semi-transparent) display in the optical-see-through technique, which directs the image to the eye and as a touchable screen in the video-see-through technique of handheld devices. To call it see-through thus ties the technique directly with that of perspective as a seeing through. But the carrier is already slipped. On one hand, the video-see-through technique shows both, the actual surroundings behind the screen and the superimposition on the screen. In the optical-see-through, on the other hand, the carrier is placed directly in front of the eyes to appear imaginarily over there, in the middle of the surroundings. The carrier is, therefore, no longer a background, but rather shifted forward. In 'trompe-l'œil' and AR, the three levels of image perception collapse: the environment seems to become the carrier and finally, determines image object and image subject.

In summary, the physicality of the carrier, as well as the status as an image, become instable. However, the fact that the carrier is shifted, multiplied or transparent changes the way of seeing beyond a mere deception. In the following, I would like to introduce the 'velo' as a layer to think the lost carrier as a literal interface. Not as a background—but as a layer between the observer and the surroundings, which also requires a shift in perception. In AR we do not just look at



**Fig. 2** Cornelis Norbertus Gysbrechts, *Quodlibet or "Vanitas-Stilleben"*, 1675. oil on canvas. 41 × 34,5 cm. Wallraf-Richartz-Museum & Fondation Corboud, Köln, Inv.-Nr. WRM 2828. Copyright: Rheinisches Bildarchiv Köln, Rolf Zimmermann, rba\_c011283. <https://www.kulturelles-erbe-koeln.de/documents/obj/05011135>

an image, we also look at the world through a layer. A layer which organizes the complex of world and image and thus our way of seeing. With the 'velo', different ways of this organization are to be worked out. The layer of the 'velo' is first constitutive as a translation function, then it is perceived with its own materiality, only to finally give up its materiality again and become an operation of structuring.

## 2.1 The 'velo': a layer for translation

The 'velo' is a semi-transparent cloth with a gridded surface through which one can see into depths. Decisive for the change in mind is the transition from a metaphor of an open window to a real studio tool. The framework continues inside the 'velo' and leads to the crucial difference to the metaphor, which is highlighted by Anne Friedberg: "but while Alberti suggested the rectangular frame and planar surface of a metaphoric 'window'; as a device for geometric calculation, his 'velo' did not require the calculation of orthogonals and vanishing points. It was, instead, dependent solely on its frame and its inset quadrants as a device to 'map'; the three-dimensional world onto a two-dimensional plane." (Friedberg 2006: 38).

This mapping procedure is a translation and not a construction. If one looks through the ‘velo’, the image is already in the frame or better, within the many small frames—it just needs to be transferred onto the paper. Round bodies and their relief are already present on the surface, likewise to a projection. While turning the gaze back and forth, the artist translates what he sees, frame by frame onto the similar grid on the sheet of paper, watching the outside—not to orientate in the landscape but rather on the drawing. Emmanuel Alloa describes this translation, in reference to the ‘velo’, as a taming of the mobile. Contrary to this, he emphasizes the greater mobility of the individual elements in the grid, which causes the subdivision of the objects (Alloa 2011: 156). This description includes an important difference to AR, which does not translate an image onto a sheet of paper. It translates the actual view in real time, by superimposing data, also in real time—with that, it is rather an unleashing of the mobile. A popular effect of AR is to translate fixed images into moving ones. In doing so, these images are superimposed by themselves, but in motion.

## 2.2 The ‘velo’: a layer with its own materiality

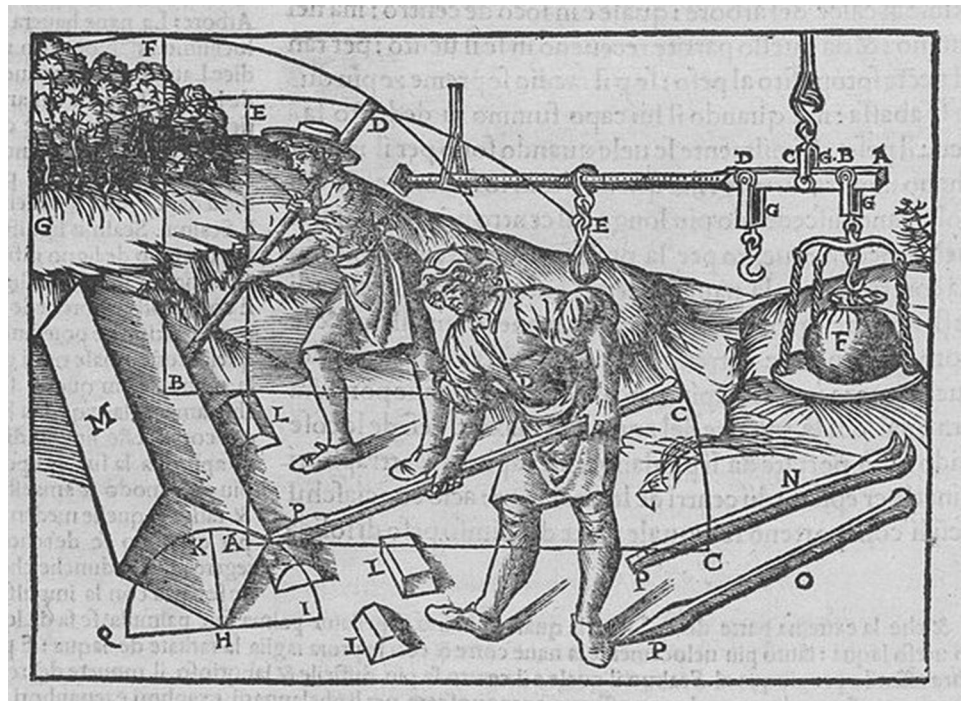
Regarding the translation function, it must be concluded that through the ‘velo’ only what is in situ can be perceived, what actually is placed behind the frame. Hence, behind that surface a fiction is impossible—but on the surface it can be reintroduced. ‘Trompe-l’œil’ as a procedure is related to the ‘velo’ in making its materiality visible and with this, a further function can be assigned to the layer. A ‘trompe-l’œil’ usually starts with a redoubling of the image carrier. Thus, it frees itself from the Albertian window and begins to approach the object, to the opacity of the canvas itself. Sybille Ebert-Schifferer shows this approach in her text ‘Der Durchblick und sein Gegenteil’. For her, the ‘velo’ is a projection surface and this surface becomes a membrane which, although transparent, is a material separation between the space of the viewer and that of the picture (Ebert-Schifferer 2016: 16). The ‘velo’ refers to both the classical representation and to the object status of the painting, thus it creates a hybrid view. For ‘trompe-l’œil’, it is important to appear as an object at first, not as a picture. Therefore, it uses different techniques to make its own materiality credible. Hence, the space of ‘trompe-l’œil’ is extremely flat. According to Ebert-Schifferer, in ‘trompe-l’œil’, the membrane of the ‘velo’ stretches slightly forwards and backwards. She imagines how artists have attached notes onto the ‘velo’ or that a fly came to rest on it. If the ‘velo’ is understood as a material layer, then new possibilities show up: first, one may use the space in front of the ‘velo’ and second, the motif behind it can be superimposed. It is precisely this space in front of the layer, that makes ‘trompe-l’œils’ as objects so believable and which connects the traditional procedure with AR. As

a material layer, it becomes an object of use—to pin something into it or to write on it.

## 2.3 The ‘velo’: a layer as a way of seeing

Above, we have noted that the grid of the ‘velo’ causes the subdivision of the objects. Reading between the lines, one could say the grid provides an organization. The individual quadrants are elements of a relationship—a relationship that can be changed. From a translation function to the visualization and use of its own materiality, the path of the ‘velo’ branches out even further. It slowly leaves the place of the studio to expand into the everyday perception. To follow this path, it is illuminating to look at a few didactic illustrations, which are designed close to the actual field of vision. A trace for this transfer is found in the books mentioned by Samuel Y. Edgerton. He examines technical and scientific treatises, which have been printed since 1520 with numerous illustrations. In these books, word and image build a unity as never seen before, which Edgerton attributes to the imaginary grid celebrating its career at that time in cartography. He emphasizes that it was the ‘velo’ that educated artists to *see* the underlying geometry in nature. This didactics spread, therefore, all educated people were able to think this invisible, but indispensable grid that underlies every picture (Edgerton 2004: 181). Due to the technical, didactic and practical advantages, a unique image form developed, which Edgerton calls an ‘incongruent sign convention’. This is described as a superimposition of otherwise illusionistic scenes by flat, abstract geometric diagrams placed directly above them to explain the underlying mathematical principle. As shown in Fig. 3, the superimposition marks a perfect workflow. And if we look closely, it is recognizable that there is more than one layer. The second angle is place between the two figures, the layer is multiplied and shifted into space. The practical component of the grid was discovered from the autodidactic craftsmen-technicians. A hybrid image was the result of that incongruent sign convention—as if one depicts something in depth through the grid, to draw afterwards on the ‘velo’ itself, which means preserving the ‘velo’ and registering it into the image. The gridded layer as interface was used to add constructively specific information. These technique books invented increasingly incongruent drawing conventions that move further and further away from the mere illusionistic representation of perspective: they duplicate objects several times in one picture, mix perspectives and explode assembly drawings. To see the underlying geometry of nature meant, being able to depict an object unnaturally from different perspectives within one image and without any logical separations. The objects are dismantled, labeled and didactically prepared, but placed in a natural landscape. To illustrate this, Figs. 3 and 4 show how the superimposition in AR is live instead, but the similarity

**Fig. 3** Woodcut from Cesare Cesarianos edition of Vitruvius, 1521. Collection Metropolitan Museum of Art. Creative Commons CC0 1.0 Universal Public Domain Dedication



**Fig. 4** Application based animation by RE'FLECT for engineering, repair concept [https://www.re-flekt.com/hubfs/REFLEKTONE\\_RepairConcept\\_1920x1080.jpg?hsLang=de](https://www.re-flekt.com/hubfs/REFLEKTONE_RepairConcept_1920x1080.jpg?hsLang=de)

is still quite recognizable. An engineer's field of vision (and with this live guidance, certain competencies become more irrelevant) is overlaid with a workflow that both presents an order and anticipates an action. What we see is directly translated into an understandable view, a program to follow, for more efficiency movements and for learning by doing.

The surroundings are visually redoubled by an animation and this animation can be cut up and rearranged. Furthermore, the superimposition by textual marking itself already generates a structure for the perceived.

In short, the 'velo' changed a way of seeing. Its material semitransparency is expressed in the theoretical and

practical ambivalence of the apparatus. The ambivalence of perspective painting (to have a flat surface, but depict depth)—is not concealed by the ‘velo’, but rather exposed. Represent depth and draw on the surface. Ultimately, it is not about switching between the two views: both are to be kept in the same field of vision. Starting as a translation function in the studio, its own material structure becomes more and more visible, as if the artists could not ignore the field of vision and the materiality of this tool in the working space. In the end, the interface of the ‘velo’ is transferred onto thinking. One begins to take the invisible layer as a structuring operation. What is seen through the layer is made more understandable on the layer.

For AR, the translation function, the materiality of the layer, as well as the operation of structuring are decisive. In the first place, there is always a translation function. For the current AR applications on handheld devices, the environment is translated into a 2D video image which is directly superimposed. ‘Head-Mounted Displays’ (HMDs), like the ‘HoloLens 2’, recognize the real environment by spatial mapping and translate it to a live 3D model. What also happens in the optical-see-through technique occurs explicitly in the handheld device. The translation function, the reduction to a flat plane, is augmented to an operation of structuring. By superimposing, the objects and the data are compressed into a flat unit of meaning. Before we look at different opportunities how this unit of meaning could be organized, a further difference should be marked.

In AR the translation of the space is not done by the artist, but by machine vision. Viewers are blind to this process. For example, the world is perceived by the cameras and sensors via spatial mapping, so that objects can be placed credibly in the surroundings. In some applications a grid is animated that spreads over the environment, following its ups and downs, which in turn is only a representation of machine vision, since there is no direct communication between the code and the perception. Within the machine vision works a program, which stipulates our point of view, what is the meaning of program (in Greek: ‘pro-graphein’). Katja Glaser and Jens Schröter point out that augmentation describes a program of efficiency, functionalization and optimization. And with AR, this program also inscribes itself into its practices and its field of vision (Glaser and Schröter 2013: 44). Without a carrier the images do not appear as programmed, what one sees is just the representation of the computed surroundings.

Ultimately, both ways of seeing are blind to each other for a certain extent of the way—our field of vision is programmed, but our perception also includes aspects that are beyond the reach of the sensors. The glasses of AR can also provide what is seen and with that the viewers are able to inscribe themselves into the world to program it.

### 3 Bundle, loose stack, and heap as models of layering

In AR, the material carrier is a semi-transparent surface which is slightly darkened. Images projected onto it appear as if they were on site by adjusting their size to the depth of the space. It is as if the real surface of projection itself is projected and extended into the room, exactly this layer itself is sometimes depicted in the field of vision—at least as a pinboard or interface like those in the application ‘Spatial’ by ‘Hololens 2’.<sup>3</sup> This makes it possible to place something in front and behind this layer. The imaginary surface can also be multiplied, thereby the individual layers overlap each other and suggest space. Ultimately, this layer is superimposed with information, pictures and objects relating to what can be seen through the semi-transparent surface.

Translation, materiality and structuring operations are inscribed as meta-levels into the grid of the ‘velo’. Therefore, I would like to describe the layering process as a stack of these different functions and operations. A stack is characterized by the fact that different levels can be gathered in one place as well as it shares the hybrid status between theoretical and practical characteristics. For this, a stack works also transformatively—it gathers individual, mostly flat elements (for example sheets of paper as in ‘trompe-l’œil’), brings them into a common relationship and generates space. In theory, a stack assembles different levels of autonomous functions, but in superpositions, it is a passageway through all of these functions. The view through the glasses of AR is a view through a stack of layers, both literally and metaphorically. A thin stack of different technologies, techniques, media, materials, functions and operations. With the ‘velo’, the layers of these stack could be bundled between the operation of translation and the function of structuring.

Additionally, a stack contains an intensive aesthetic potential, which presents a unique way of seeing. A stack oscillates between horizontal and vertical. It determines the space of ‘trompe-l’œil’, the computer desktop and AR. The use of a stack shifts from the desk to the desktop, as Friedberg notes: “The user would manipulate from a position as if in front and also above [...] ‘desktops’ that defy gravity and transform the horizontal desk into a vertical surface with an array of possible documents and applications: ‘icons’ that represent objects or, more exactly, object-oriented tasks.” (Friedberg 2006: 226). This is crucial in AR—due to the

<sup>3</sup> To see the animation adequate, please watch the demonstration video at 9:00 min. <https://www.youtube.com/watch?v=uIHPPtPBgHk>.

better legibility the text is usually set up parallel to the viewer's own field of vision.<sup>4</sup>

In 'trompe-l'œil' papers are stacked to leave the surface minimally behind, thereby different layers are visible at the same time. What we see is a stack of sense-fragments, of text-quotations and picture examples which come out towards the viewer (Fig. 2). With that, a stack piles up flat units into something three-dimensional transforming the work of art into an everyday object at the same time. A transformation that encompasses the core of the 'trompe-l'œil'. Every etching, drawing or text bundled with a ribbon in 'trompe-l'œil', turns into something to use instead of something to look at. Wade Guyton organized an exhibition at the Aspen Art Museum in 2017. He stacked his paintings on the wall, which can be seen as a typical studio situation. Isabelle Graw mentioned they would become a sculpture. The consequence—they cannot longer be experienced aesthetically, instead they have been transformed into a product that can be packaged, exchanged, and traded (Graw 2017: 238). Usually, a stack is bound to gravity, but on screen and in AR the individual layers can be vertically aligned and may appear semi-transparent. In this way, they create a linked image together with the background. In AR, these layers additionally refer to what is visible in the surroundings. Above, I drew a few parallels between the 'velo', 'trompe-l'œil' and AR. Within the following, I present two different ways in which the stack of layers transform our seeing in the named procedures, beginning with 'trompe-l'œil' to switch to AR. Both procedures model an interplay of different layers, because they are visible simultaneously. Subsequently, I will propose a third possibility, a thought experiment—the concept of the heap, to confront order with chaos.

### 3.1 The stack as bundle

A painterly reflection as well as the current technological one might be examined through the layers extending into depth. The surface of 'trompe-l'œil' can be indicated by cracks and fissures. They refer to an aging process and to a fragility of the specific materials as well as to a deeper level underneath. This allows the viewer to see different layers at the same time. However, the cracks are not placed by chance. In a text about the broken glass in 'trompe-l'œils', Monika Wagner shifts the focus of attention to the materiality, actually depicted through cracks and their structural function as a comment. She also emphasizes the significance of the 'velo' as a medium of flatness. As the 'velo' helps to translate the

space into the flatness of the picture, the broken and thus visible glass ties the illusory space to the surface (Wagner 2010: 41). What is far apart in reality can thus be connected in the flatness. In other words, a reference is created, because two different things seem to be on the same layer. Using glass, different patterns beside a squaring are possible. This example is about the transparent materiality of glass, the very glass that is the carrier for the projection in AR.<sup>5</sup> The glass as a carrier of information and as a transparent layer between the observers and the image. The broken glass in the painting of Laurent Dabos about 1808 becomes visible (Fig. 5). It thus provides both, protection (of the underlying layer) and visibility (of the content below). In case of a self-aware positioning of these cracks, it is possible to organize the image through the cracks. Highlighting specific areas on it and making other less clear. Thanks to the transparency of the glass it is possible to superimpose the annotated image with the visual comment, without erasing the image (Wagner 2010: 45). Both share the same field of vision and yet not the same depicted layer. However, it only achieves this through its material-specific properties. The 'trompe-l'œil' of Laurent Dabos shows a stack as well as a broken glass. The origin of the cracks is the text, from which different lines link the individual pictures and figures. The figures' view, highlighted, because it is free of glass, is as sharp as the edges.

The 'trompe-l'œil' thus found an extremely specific layer structure: a layer structure in which the individual layers simultaneously remain in the area of the visible. On the solid, wooden surface are various images and texts arranged, superimposed with a broken glass. These three functionally different layers are bundled into one fixed unit. The unity of that bundle is constituted on one hand by the everyday object of the picture in the frame and on the other hand by the fact that all layers are structurally related to each other.<sup>6</sup> Finally, the layer of the glass, if it is perceived, decides how we have to read the lower levels.

### 3.2 The stack as a loose stack

The possibilities of AR are based on similar strategies as 'trompe-l'œil', which allow to illuminate and expand each other. The things that in reality are far apart can thus be compressed to the same layer with its superimposition. It is the challenge to bundle the information superimposed

<sup>4</sup> In his book *Cultural Techniques* Bernhard Siegert examines the 'trompe-l'œil' as a conflict between two cultural techniques, gazing and reading (Siegert 2015: 164–191). Which also refers to the verticality and the horizontality of a picture.

<sup>5</sup> Referring to precisely these fractures and cracks, AR often animates a wall breakthrough to allow fictitious elements to break into the real space. This emphasizes the materiality of the wall.

<sup>6</sup> Thanks to a discussion with Carolin Meister, the concept of the bundle could be worked out as a fixed stack in which the two outer layers hold an in-between together.



**Fig. 5** Laurent Dabos, trompe-l'œil with print of tsar Alexander I of Russia, together with other prints and drawings behind a broken pane of glass, ca. 1808. oil on panel 63,5 x 50,5 cm. PD-Art/PD-old-100=/1835/France; PD-US

on the glass with the related object as tightly as possible. To make this clear, although the arrangement in ‘trompe-l'œil’ and in AR seems similar, AR does not present a fixed bundle. Therefore, it could be described as a loose stack in which every layer is able to change its positions immediately, without a reasonable cause. With reference to a museum related device developed by the ArLab Weimar, Oliver Fahle describes the possibilities of the technology and how it changes the concept of the image. Instead of the common term of immersion, usually referring to virtual reality, Fahle explains the technique of AR as participation. It creates a view on another visual layer of the same image (Fahle 2006: 93). He applies this to the specific constructions of Arlab, in which earlier stages of the same painting are projected onto its present layer, thereby the chronological succession is less visible. Decisive for the argumentation is the view onto other layers of the same picture. According to Fahle, the picture is thus augmented by a visual halo, which, however, does occupy the picture itself. The final layer is confronted with earlier stages and information, therefore the pictorial event intervenes in the one unchangeable work and mediates between the one and the many (Fahle 2006: 95). The previously invisible pre-stages now participate

with the visible original. The chronological order can be reversed and restacked. The stability of the final work is weakened, without actually being transformed. It shows up as a layer-network that constantly creates new references and thus evokes a shift in mind. The work of art itself is a shift in mind, but the participation of different layers allows to think about layering, temporarily as well as spatially, and to reconsider the final work.

### 3.3 The stack as a heap

I would like to stress a third possibility of a stack to undermine the impression that it is always about a comprehensible system of layering. What I have in mind is a stack fallen down from the desk, which is scattered all over the ground. This kind of layering is a heap, a random arrangement of data and images. The paradox of the heap is that its structure is not recognizable and furthermore, the heap itself cannot be determined. One cannot define how many elements comprise a heap and yet you know that it will remain a heap if you remove some elements. AR is, therefore, also capable of superimposing the field of vision with layers in such a way that it appears to be filled up. The superimposition is that rich in number, that one cannot see what is superimposed: a data heap which collects information from all around to tear the field of vision into pieces.

The possibility of accumulation deprives the stack of its stability in several aspects, firstly, because it introduces disorder and secondly, because it shows that each augmentation includes a reduction—any superimposition, no matter how transparent, carries the possibility of a concealment. With the concept of the heap, the question of the limits of superimposition comes into view. AR aims to filter the diversity of reality for more efficient use by linking specific information to objects. For now, as one opportunity, the infinite diversity of the world and the huge data heap of possible links collide. A stack offers the promise of an understandable arranging, even if it reorders chronological and spatial relations. In ‘trompe-l'œil’, a mess is always a calculated one, in AR, the reference to the place protects from randomness. If an object and an information appear connected, we imply a logic—however, we imply a unit of meaning that does not necessarily belong to it, just because it is visually bundled.

Furthermore, just imagine that anyone could leave messages with AR or that all available information about a place would be visible simultaneously—the place would be covered by comments. The commented would drown by its comments. This is only a thought experiment—hence, every superimposition includes the possibility of filling up the field of vision, which opposite would be the uncovering. The model of “heaping” is an extreme case of the superimpositioning, which reminds us, that knowledge is not only an accumulation. Nevertheless, there is something constructive



about this extreme case. Assume that our perception is always occupied: a grid of knowledge through which we interpret the environment (Serres 2010: 74). Hence, AR can make us aware that our way of seeing is superimposed by the already known—to work on its uncovering.

#### 4 About the reactivation of ‘trompe-l’œil’ in the age of digital layering

Layering is a constitutive procedure, which redistributes the way of seeing. If tied to specific techniques and technologies, the sequence of layers begins to shift. The ‘velo’ as an apparatus, which at first has stretched out a layer in front of the eyes, mediates, theoretically as well as practically, between the ‘trompe-l’œil’ and AR. Our gaze, confronted with this current field of vision, is forced to consider a new way of thinking and seeing. The image structure of ‘trompe-l’œil’ and the construction of the ‘velo’ are useful to analyze newer procedures of layering and vice versa the newer procedures allow a more precise description of the traditional ones. To conclude this line of thought, I wish to focus on a work of art in which both ways, tradition and innovation, are crossed.

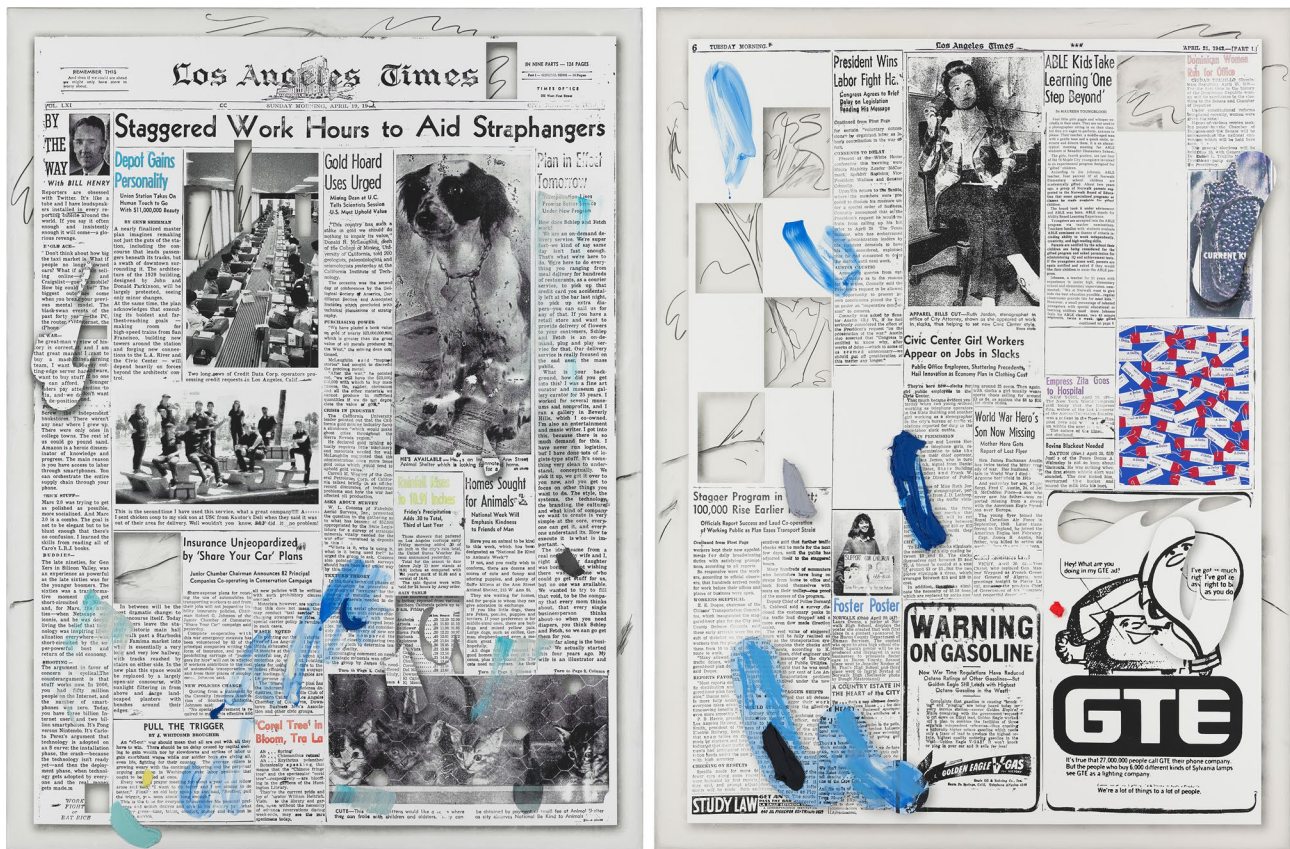
For this essay, I would like to end up with a brief analysis of Laura Owens’ untitled diptych from 2015, which is exhibited in the Museum Brandhorst in Munich, Germany (Fig. 6). Laura Owens applies different techniques known from ‘trompe-l’œil’, which she elegantly transfers into a thinking of the digital. The thesis is that Owens reactivates the ‘trompe-l’œil’ as a traditional procedure, because it is an adequate analytical tool for digital image culture. Digital images are not welded together with a carrier. Hence, they show a floating weightlessness. One can allocate a carrier to digital images, but in fine arts the invariance, the unchangeable and necessary mutual conditionality of image and carrier is decisive. For example, a painting has a fixed size and this specific size is necessary for its appearance. Very strictly formulated by Henri Matisse, who, therefore, could not even make a sketch of a smaller format than the original: “If I take a sheet of paper of a given size, my drawing will have a necessary relationship to its format. I would not repeat this drawing on another sheet of different proportions” (Flam 1995: 38). ‘Trompe-l’œil’ is characterized by the fact that it unsettles the alliance between image and carrier. The carrier of ‘trompe-l’œil’ pretends to be a part of the real environment (a wooden board, a pin board, etc.) instead of being part of the picture. Moreover, ‘trompe-l’œil’ passes off the figures on the carrier as carriers themselves (a sheet of paper). AR also disguises the carrier, to make the image float and assert it as a true part of the environment. Owens’ image production may be described as one that makes use

of the congruence and difference of these two procedures to work onto the alliance of image and carrier.

In the diptych, different techniques structure the various levels of the picture plane. Oil, acrylic paint, Flashe Vinyl Paint, charcoal and gesso assemble on the canvas. I will start with the core layer, the newspaper that fills several ‘trompe-l’œils’ and provides a career in cubism. Owens uses original silkscreen plates from the 1942 Los Angeles Times, which she found in her studio. It thus already begins with an anachronism, which is pursued even further. The technique of screen printing is applied, but then digitally manipulated and blown-up to the size of the canvas (350.5 × 264 cm). The blow-up shows a variance of the original as well as it is credible, since the digital newspaper does not have a strict format. The canvases look like two big screens. I call the newspaper a core layer, because there are further levels both in front of and behind it—a membrane stretched on both sides. To speak of a core layer already rises the suspicion that the carrier has been shifted. The picture is not about foreground and background, it is about different layers, with different functions.

The lowest layer in this work is a drawing of thin, grey strokes, which form a landscape on the canvas. They appear as wallpaper that seems to be placed independent of the layers above. As if the drawing has already occupied the background, which is now challenged by the newspaper. However, this conflict is calmed down by the fact that Owens has digitally perforated the newspaper. These remind us of the broken glass. It is not the physical materiality, rather it is its digital surrogate which is cut. Small holes that allow to look through them. Those small holes which structure the ‘velo’ to organize the space behind. Due to the frontality of the writing, the newspaper marks a solid layer which is impregnated, while the shadows, especially in the cut-out parts, expose the layer as being above the background. The newspaper itself is superimposed by apparently gestural brushstrokes, as well as by cut-outs from the newspaper. Some brushstrokes and cut-outs also cast shadows and thus float on another level above the newspaper. Single strokes of color, such as the striking black in the lower left half of the right-hand picture emerge almost haptic and stretch the membrane forward towards the viewer.

I tried to sort three layers in this painting, but this sorting is deceptive, as Owens interweaves the different levels. The superimpositions and procedures have references to each other. Therefore, in the lower right corner of the left picture, single fragments are cut out of the wire netting and depict a pair of eyes next to those of the cats. From the photograph, an elongated shape runs upwards, which continues the digital cut-out above. This cut-out of the newspaper again is behind the newspaper to add a further layer. Moreover, the color gestures, which are highlighted by an artificial shadow, are definitely no longer gestures.



**Fig. 6** Laura Owens, untitled, 2015. Oil, acrylic paint, Flashe Vinyl Paint, charcoal and gesso assemblage on the canvas. 350.5×264 cm. Collection Museum Brandhorst Munich Copyright: Laura Owens, bpk, Bayerische Staatsgemäldesammlungen, Haydar Koyupinar

They are the result of a planned approach. On the other hand, the impasto applied oil bulges build up a materiality which was just negated in the floating constellation. Finally, Owens introduces blanks into the newspaper and replaces some articles from 1942 with recent or perhaps invented ones. Although there is still much to say about this painting that cannot be fully elaborated here, instead of a layer structure I would like to name three meta-levels.

#### 4.1 Layering of different production techniques

Occasionally, reference is made to skeuomorphism in relation to the paintings of Owens (the strategy in which a traditional process is digitally imitated without retaining its function or materiality). The familiar perception makes it easier to handle the new objects. This could mean both the artificial shadow and an artificial impregnation on the screen. In the painting discussed, the layers generate a transfer of various production processes, as well as the transition from imitated to physical materiality. Owens combines digital techniques with traditional ones in one field of vision. What represents information without a

carrier in the digital world can appear materially captivated, and what is traditionally associated with a carrier, begins to float on the surface. Different production techniques are displayed and refer to the craftsmanship which ultimately culminate in a representation of these operations. Print, photography, color, drawing, writing, all of them are individual layers and media that mutate into a cipher, each oscillating between the traditional and the digital.

#### 4.2 The layering of different spatial levels

The plane of the newspaper draws an inner frame, which can then be crossed by a pasty mass of paint. This overstepping of the inner frame is supported by shadows. Like a staple, the turquoise color mass at the bottom left connects the newspaper to the carrier. Further techniques and layers creep in and remind us of nailing and cutting in traditional 'trompe-l'œil'. Furthermore, there is type, which is traditionally entangled with a carrier and thus supports the materiality of the core layer. Next to the type, however, there are images that burn holes into the solid plane, as

the digital cuts do. A membrane stretched on both sides without creating an illusion of a physical object. The free drawing behind the newspaper and the brushstrokes that seem to be dancing in the air prevent a comprehensible order. The carrier is pushed forward by the newspaper and simultaneously calmed down by the haptic reality of the color mass.

### 4.3 The layering of different temporal levels

Both the integrated newspaper articles and the different spatial levels allow an anachronistic sequence that is constantly interrupted to continue at another location. At least there is a carryover from the tangle of different layers to the production process, which cannot be clearly traced back. Hence, the modern paradigm of painting to show a transparency of the made as made, is negated, suggesting interchangeability of both, the arranging of layers and of production sequences. Viewing the different layers as being apparently at the same height makes it possible to create new references. Owens' painting can be described as a stack, in which it is never clear which side is at front or which element was placed at the beginning. However, this allows different levels and techniques to be linked. Owens' arrangement is a layer-network of different relations that can be re-articulated and re-contextualized over and over again, a stack of layers in which each side seems to be connected to each other. As if each layer is represented by a pane of glass and thus has its own background. Each of them is transparent and stands out from the layer below at a real distance, but at the same time, these layers are constantly being penetrated anew. For the view, units of meaning are created when different layers seem to be close to each other, however, with the movement of the eyes the layer structures change and the units of meaning are restacked with them.

The thesis that Owens reactivates the 'trompe-l'œil' as an adequate analytical tool for digital image culture is based on the following overlaps. The 'trompe-l'œil' works into the association between image and carrier in order to undermine its alliance. There is no figure in front of a background, rather every possible figure camouflages itself as a further carrier which appears bundled together by a representation of operations, such as stapling, nailing or gluing. The representation in 'trompe-l'œil' is now linked to the indirectness of digital operations: skeuomorphism, representation of object-oriented tasks, or finally the well-known representation of touch. In Owens' picture, all these indirect operations are linked. This process, she extends to art-historical operations, like the physical gesture (which absolutely requires a direct physical application of paint). To interrupt directness, a layer, more precisely, a shadow layer, is slid in between. The shadow is, therefore, present before the application of paint. The 'trompe-l'œil' unsettles

the carrier, which was never a binding one in the digital. Since 'trompe-l'œil' stacks carrier on carrier, there is no background anymore. A circumstance which Owens transfers to the digital image culture: the newspaper is not the background, instead it appears as a core level from which it can act forward and backwards. The invariance, i.e. the fixed alliance of image and carrier, is decomposed together with a strict sequence of readability, both spatially and temporarily. AR tends to merge technology and reality, as 'trompe-l'œil' tends to embed an image into reality—Owens combines both tendencies, to work at the border between technology, image and reality. The 'trompe-l'œil' in awareness of digital image culture leads Owens to combine and cross information that is traditionally associated with a carrier and information that is not. Both AR and 'trompe-l'œil' are not isolated instances, they want to infiltrate everyday life and everyday perception with the potential to reflect on it.

## 5 Summary

Due to the carrierless appearances that AR throws into space, the 'velo' was questioned as a semi-transparent layer in which the space behind remains visible together with the structure of the grid. The 'velo' as interface is an apparatus of translation, whose materiality and structure, as crucial differences to the window metaphor, become constitutive for the field of vision in 'trompe-l'œil' and in described illustrations. As in AR, the semitransparent layer is used to map something onto it that relates to what is displayed in depth—a mode of thinking and seeing that establishes a layer between the viewer and the visible, commenting on what is seen. This technique can also be found in 'trompe-l'œil', although the layers are pressed very tightly together so that they unalterably bundle a fixed order. While digital images do not require a specific carrier, AR reintroduces the invariance of placement through the back door. They are not tied to one place and yet they are bound to it by a reference. The 'velo' was built as a translation of depth into flatness, this transfer overlaps the field of vision of AR and thus guarantees the proximity of superimposed information with the object—as if they were connected on one layer. It can superimpose objects—with previous or future versions of this object or with references to other images or objects. A flat object can also be superimposed with a deepening or heightening that transforms spatiality. Also fixed images can be superimposed by motion. Strict orders, whether chronological or in spatial depth, thus become loose and can be restacked. Hence, even these new orders are not fixed, which means that the projection can overlay an object without erasing it as well as it can be removed again without leaving any residue. Since the relationship is not a fixed one, this demonstrates a great potential for open, flexible and variable

commitment. The information is then also not definite, because the space of information is constantly shifted. Ultimately, classical categories such as background and surface are no longer stable, as they have become interchangeable through digital superimposition. The way of machine seeing in AR recognizes the real environment. To superimpose the real environment means to transform the perception of this environment. The biggest challenge for this new way of seeing is to define the limits and differences of the new field of vision, as these are constantly stretching and become blurred in the process. “Crossroads of seeing” ultimately means to pause at this crossroad, not only to look at the intersection, but also to see where the ways divide.

**Acknowledgements** Prof. Dr. Carolin Meister, Prof. Dr. Stephan Günzel, Prof. Dr. Matthias Bruhn, Moritz Queisner MA.

**Funding** Open Access funding enabled and organized by Projekt DEAL. The research project ist funded by FAZIT-STIFTUNG Gemeinnützige Verlagsgesellschaft mbH, Frankfurt.

**Availability of data and material** Not applicable, separate permissions are required.

## Compliance with ethical standards

**Conflict of interest** Not applicable.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Alloa E (2011) *Das Durchscheinende Bild. Konturen einer medialen Phänomenologie*. Diaphanes, Zurich
- Ebert-Schiffner S (2016) Der Durchblick und sein Gegenteil. Malerei als Täuschung. In: Hedinger B, Boehm G (eds) *Täuschend echt. Illusion und Wirklichkeit in der Kunst*. Hirmer, Munich, pp 16–23
- Edgerton S (2004) *Giotto und die Erfindung der dritten Dimension. Malerei und Geometrie am Vorabend der wissenschaftlichen Revolution*. Wilhelm Fink, Munich
- Elcott N (2016) The phantasmagoric dispositif. An assembly of bodies and images in real time and space. *Grey Room* 62, Winter, pp. 42–71. <https://www.columbia.edu/cu/arthistory/faculty/Elcott/Phantasmagoric-Dispositif.pdf>. Accessed 15 Sep 2020
- Fahle O (2006) *Augmented Reality. Das partizipierende Auge*. In: Neitzel B, Nohr R (eds) *Das Spiel mit dem Medium Partizipation—Immersion—Interaktion*. Schüren, Marbourg, pp 91–103
- Flam J (1995) *Matisse on art*. University of California Press, Berkeley
- Friedberg A (2006) *The virtual window. From Alberti to Microsoft*. The MIT Press, Cambridge
- Glaser K, Schröter J (2013) ‘Tag that wall’. *Augmented Reality-Apps am Beispiel der Street Art zwischen Skripten und Praktiken. Sprache und Literatur* 44(1):30–48. <https://doi.org/10.1163/25890859-044-01-90000004>
- Graw I (2017) *The love of painting. Genealogy of a success medium*. Sternberg Press, Berlin
- Husserl E (2005) *Phantasy, image consciousness, and memory (1898–1925)* (trans: John Brough). *Collected Works Vol. XI* Springer, Dordrecht, Netherlands. <https://doi.org/10.1007/1-4020-2642-0>
- Serres M (2010) *Malféasance. Appropriation through pollution?* Stanford University Press, Stanford California
- Siegert B (2015) *Cultural techniques. Grids, filters, doors and other articulations of the real* (trans: Geoffrey Winthrop-Young). Fordham University Press, New York
- Wagner M (2010) *Das zerbrochene Glas. Opake Kommentare auf einem transparenten Medium*. In: Hedinger B, Boehm G (eds) *Täuschend echt—Illusion und Wirklichkeit in der Kunst*. Wilhelm Fink Verlag, Munich, pp 40–47

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Artificial intelligence and institutional critique 2.0: unexpected ways of seeing with computer vision

Gabriel Pereira<sup>1</sup> · Bruno Moreschi<sup>2</sup>

Received: 25 July 2019 / Accepted: 18 August 2020 / Published online: 14 September 2020  
© Springer-Verlag London Ltd., part of Springer Nature 2020

## Abstract

During 2018, as part of a research project funded by the Deviant Practice Grant, artist Bruno Moreschi and digital media researcher Gabriel Pereira worked with the Van Abbemuseum collection (Eindhoven, NL), reading their artworks through commercial image-recognition (computer vision) artificial intelligences from leading tech companies. The main takeaways were: somewhat as expected, AI is constructed through a capitalist and product-focused reading of the world (values that are embedded in this sociotechnical system); and that this process of using AI is an innovative way for doing institutional critique, as AI offers an untrained eye that reveals the inner workings of the art system through its glitches. This paper aims to regard these glitches as potentially revealing of the art system, and even poetic at times. We also look at them as a way of revealing the inherent fallibility of the commercial use of AI and machine learning to catalogue the world: it cannot comprehend other ways of knowing about the world, outside the logic of the algorithm. But, at the same time, due to their “glitchy” capacity to level and reimagine, these faulty readings can also serve as a new way of reading art; a new way for thinking critically about the art image in a moment when visual culture has changed form to hybrids of human–machine cognition and “machine-to-machine seeing”.

**Keywords** Institutional critique · Computer vision · Error · Image analysis · Contemporary art

## 1 Introduction: Duchamp's Fountain is a urinal

*MORESCHI: On 23 October 2017, Gabriel sent me an email. In it, there were two images—a painting of Christ and Duchamp's Fountain. The email went on with a series of graphics, percentages and keywords analyzing these two images. At no point were they interpreted as art. Duchamp's Fountain was described as a plumbing fixture, product design and as... a urinal. Behind this reading was Google's state-of-the-art AI: Google Cloud Vision (Fig. 1).*

The image of Duchamp's *Fountain* is especially relevant as a starting point. With this artwork, as Calvin Tomkins (1998) details in his biography of Duchamp, the French artist aimed to test how democratic the New York Society of

Independent Artists was in the selection process for their salon. Those who work with art may be familiar with this controversial story: after having lunch with Walter Arensberg and Joseph Stella, Duchamp invited them to accompany him to J. L. Mott Iron Works, a store in New York that specializes in sanitary equipment. There, he bought a porcelain Bedfordshire urinal. Later, in his studio, Duchamp flipped the urinal upside down and signed the bottom left side with the name R. Mutt and the year (1917). He then submitted the piece to the exhibition, without a return address. According to Tomkins, when the package was opened by the salon's jury, the juror George Bellows cried out: “It's indecent! It's indecent! We cannot show this. This thing is nothing more than what it is.”<sup>1</sup>

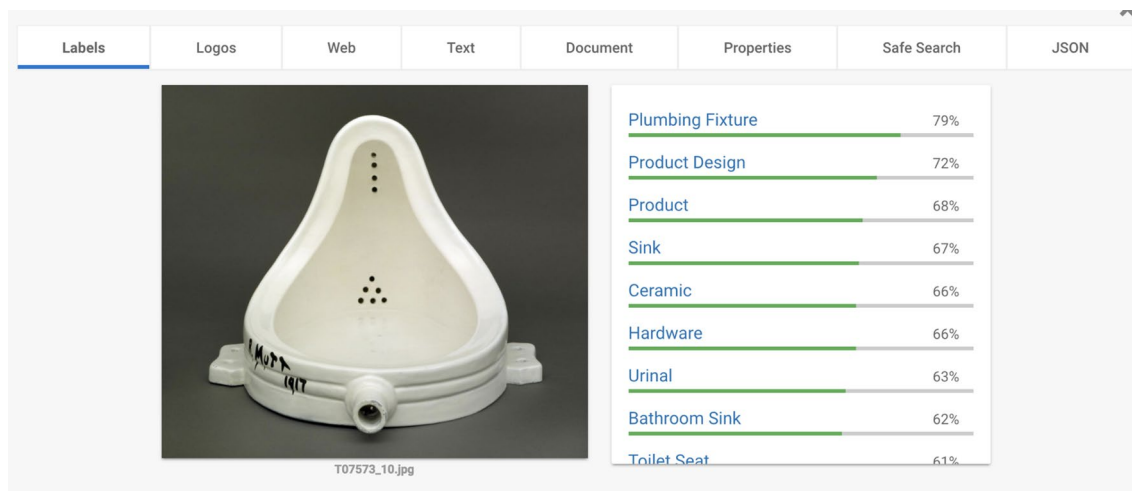
Today, Duchamp's *Fountain* is considered one of the most influential artworks of the twentieth century, an iconic image that is widely known in the popular imaginary. The act of transforming the urinal into a fountain by placing it

✉ Gabriel Pereira  
gabrielopereira@gmail.com

<sup>1</sup> Department of Digital Design and Information Studies, Aarhus University, Aarhus, Denmark

<sup>2</sup> Faculty of Architecture and Urbanism, University of São Paulo, São Paulo, Brazil

<sup>1</sup> The story of the *Fountain* has been under dispute in recent years. Research by historian Irene Gammel indicates there is evidence that the piece was actually created by dada artist Baroness Elsa, although Duchamp was the one that ultimately proposed it to the jury.



**Fig. 1** Print from the e-mail that Moreschi received from Pereira, with the reading of Duchamp’s *Fountain* by the AI Google Cloud Vision. None of the results considered the image a photograph of an artwork

in the artistic space is a crucial ontological shift proposed by contemporary and conceptual art. But 100 years later, when Google’s Cloud Vision looked at the image of this artwork, its “eyes” didn’t see anything different than Bellows’: “ceramic, product, urinal.”

Google Cloud Vision is the result of much human and machine labor. It is based on Google’s recent expansion into the new and promising field of computer vision: using algorithms, machine learning, and a lot of data to train “smart” machines to see and understand the world around us. Fei-Fei Li, a Professor at Stanford and also former Chief Scientist at Google, is one of the most prominent voices of the computer vision research field. In her widely watched Ted Talk *How We’re Teaching Computers to Understand Pictures*, she explains what this means: “Just like to hear is not the same as to listen, to take pictures is not the same as to see, and by seeing, we really mean understanding. (...) Vision begins with the eyes, but it truly takes place in the brain.” (Li 2015).

This difference between taking pictures, seeing, and understanding is intriguing, especially in a moment when the utopian computer vision discourse (as above) claims that it is possible “to teach the machines to see just like we do: naming objects, identifying people, inferring 3D geometry of things, understanding relations, emotions, actions and intentions” (Li 2015). But, as we first experimented at that moment, commercially available computer vision algorithms from leading tech companies are not trained to “understand” artworks—they do not understand the context, the subtext, the emotion. When they do interpret it correctly, from the point of view of the human observer, they read them in their superficiality: the “thing as nothing more than what it is.”

Going back to the case of Duchamp’s *Fountain*, the jury of the salon finally decided not to exhibit the piece. When the artist got it back, he took it to the famous photographer

Alfred Stieglitz, to be photographed using the same method as a sculpture. The *Fountain* itself probably had the same destination of many other of Duchamp’s ready-mades (the trash bin)—replicas of the work now abound in many prominent museums. But the image of the artwork by Stieglitz remains, and is now on the cover of books, magazines and widely available on Google Image Search. This story reveals the way in which art and its history are constructed and experienced through images (rather than through the works’ materiality). Now, using artificial intelligence, like the one by Google, it is also clear that these images are not embedded with what these artworks mean, their context, history, and subtext. But this is not necessarily a problem in itself, as there is a rich possibility in all of this: because commercial AIs are not at all familiar with works of art, we have interpretations of their images that are almost always devoid of a more “subjective” sense of art and its context—which were so vital to Duchamp and indeed other artists.

*PEREIRA: I’ve just watched the first episode of “Ways of Seeing,” the influential BBC TV show by art critic John Berger. I’m particularly impressed by the scene where he asks kids to describe a painting of Christ by Caravaggio. The kids (very adorably) speak incessantly, conjecturing what the subjects are doing (maybe stealing the food or about to kiss?), and who they are (male or female? Jesus?). Berger points out how these kids “demystify” the artistic work by looking at it “very directly,” from their own experiences, ignoring the context in which the images exist (Fig. 2).*

The AIs that power computer vision can only understand the world based on their own “experiences” when making assessments about the world. AI requires data to be trained on, from which it generates a model; for example, pointy ears means cat, floppy ears means dog. This model is most often not visible or interpretable by humans, and frequently



**Fig. 2** Frame of episode 1 from *Ways of Seeing*, TV show written and presented by British art critic John Berger in 1972 (BBC Two) and adapted from a book of the same title

involves patterns that are not visible to the human observer (Olah et al 2018). Could AI then be used to bring a fresh, denaturalized set of “eyes” to art, expanding and levelling what artworks are (or could be)? But also, if art is such a new thing to the “eyes” of computers, could it also reveal what its underlying experiences are, much like the kids use their experience to talk about Caravaggio’s painting? It was with these questions that we began this research. How revealing are these “eyes” that have never entered a museum, and how can we make use of them to see for ourselves? How could we use these non-human perspectives to “illuminate our understanding of the world,” thus unsettling “the relations between what we see and what we know in new ways”? (Cox 2017: 14).

We were invited to work with the Van Abbemuseum (NL) collection and proceeded to reading their images using commercial image-recognition (computer vision) artificial intelligences from leading tech companies. The museum’s collection, consisting of conceptual and contemporary art, as well as some older and lesser known works, was particularly suitable for our proposed inquiry.

In the first part of this article, we describe our methodology and how we approached the creation of the platform for reading images through AI, how we analyzed the results, and what categories came out of it. In the second part, we look at the question: what happens to art (i.e. images of artworks) when it is read through commercial AIs? This means putting art through a system that is not specifically made for it, where it is not protected by the elements of the artistic apparatus. Drawing from institutional critique, we investigate how this may work to expand and level the meaning of artworks. In the third part, we follow with: how does commercial AI react to art, a content that is different from what it is used to? And, what does this procedure reveal about the underlying values and epistemologies of popular

AI tools? Here we borrow from literature on critical AI/algorithm studies and other academic research on technology/data, where algorithms are considered in their sociotechnical complexity. We conclude by stating the contributions we offer, as well as our considerations for future research and practice.

## 2 Methodology: “Honor thy error as a hidden intention”

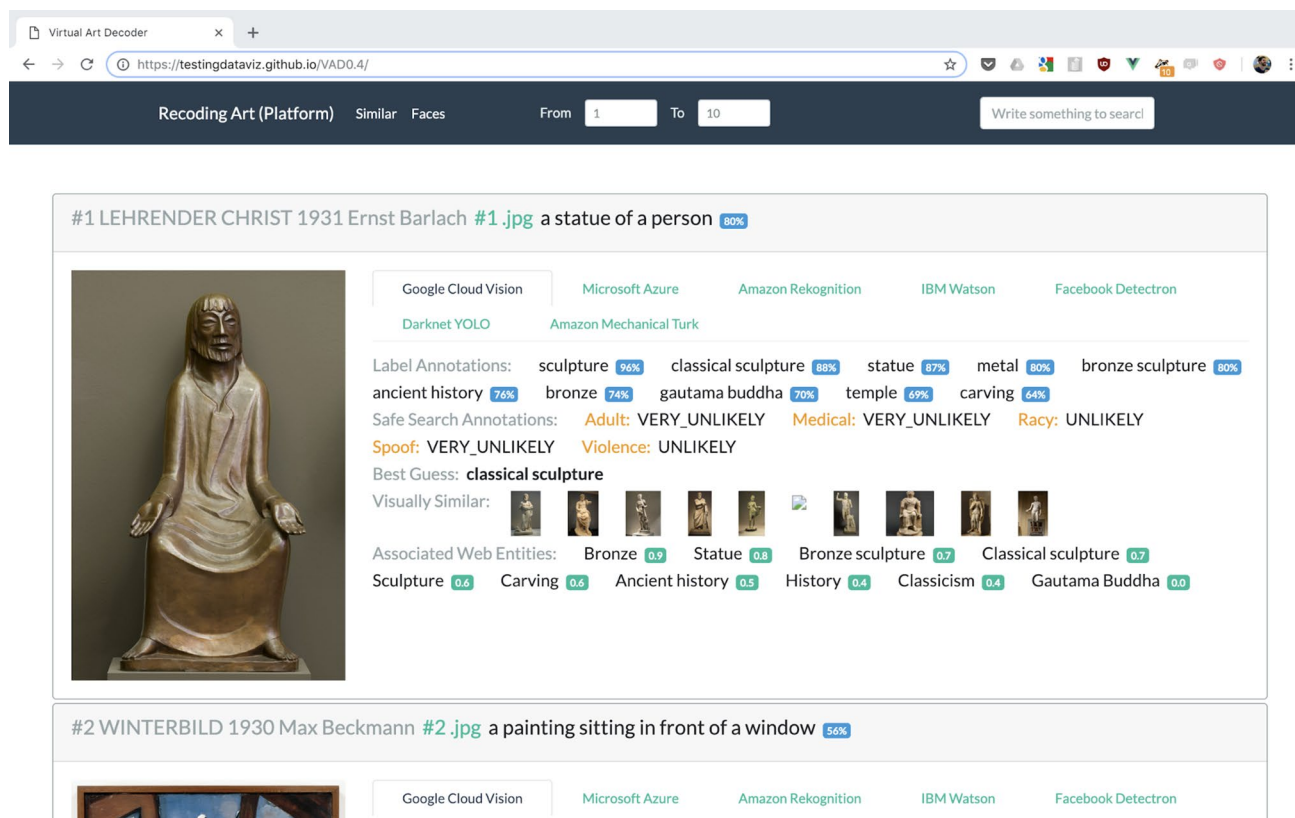
The Van Abbemuseum collection is generally made up of contemporary and conceptual art. We received all the images of the collection (about 2500 high-resolution photographs of artworks), but decided to focus on two of the permanent exhibitions, *The Making of Modern Art* and *The Way Beyond Art*. This resulted in working with a total of 654 images.<sup>2</sup>

To create a new way of interpreting this set of images, Pereira created a script to send the images of the artworks to six of the most commonly used commercial AI services from Google, Microsoft, Amazon, IBM, Facebook, and the widely used open-source YOLO library. The results obtained for each artwork are shown through a custom web interface (Fig. 3), which is accessible and open-source (enabling other readings and analyses) through this link (<https://testngdataviz.github.io/VAD0.4/>).

This centralized AI results interface was created prior to the official start of Moreschi’s residency at Van Abbemuseum. This meant that our first contact with the works in the collection was digitally mediated. Following this logic of physical detachment, even though he stayed only a couple of blocks away from the museum and its collection, the first two weeks of Moreschi’s stay in Eindhoven did not focus on the museum itself (and its physical works). Instead, he dedicated all his time to the analysis of the approximately 55,000 results obtained from the analyses of 654 works (available on the Recoding Art interface), and to the construction of a method capable of organizing the results through identified patterns.

“Honor thy error as a hidden intention.” This card, from the set of cards *Oblique Strategies* created by Peter Schmidt and Brian Eno to aid in the artistic process, epitomizes our methodological approach. This advice was valuable in a selection process involving interpretations that at first seemed like blatant misunderstandings by dumb machines. We decided to steer away from a feeling of superiority, in relation to technological systems. Minimal attention was put

<sup>2</sup> They were chosen because the first exhibition deals directly with the changes in the status of art in modernity, especially its reproducibility, and the second is dedicated almost exclusively to contemporary art, much of which dissociates what is seen from its signification.



**Fig. 3** Recoding art, an open-source website with the Van Abbemuseum’s art works reading by AIs

on results that were “true” or “correct.” On the contrary: we decided to value the unexpected outcomes.

Although the struggle for algorithmic auditing, accountability, and ethics is very important, given the amount of problems AI is already causing and how these errors affect people (especially underserved minorities and marginalized communities), focusing exclusively on “solving bias” may serve as a diversion from critically interrogating these systems and understanding them in their complexity (Powles and Nissenbaum 2018). Here, we turn to commercial Computer Vision systems’ failures as a way of critically and imaginatively speculating on the machinations of the systems of both AI and Art.<sup>3</sup> It is about critiquing the operating logic of algorithms and showing that they are neither a “given,” nor “certain,” and thus complicating the “mythical, objective omnipotence” (ibid.) that they so often evoke.

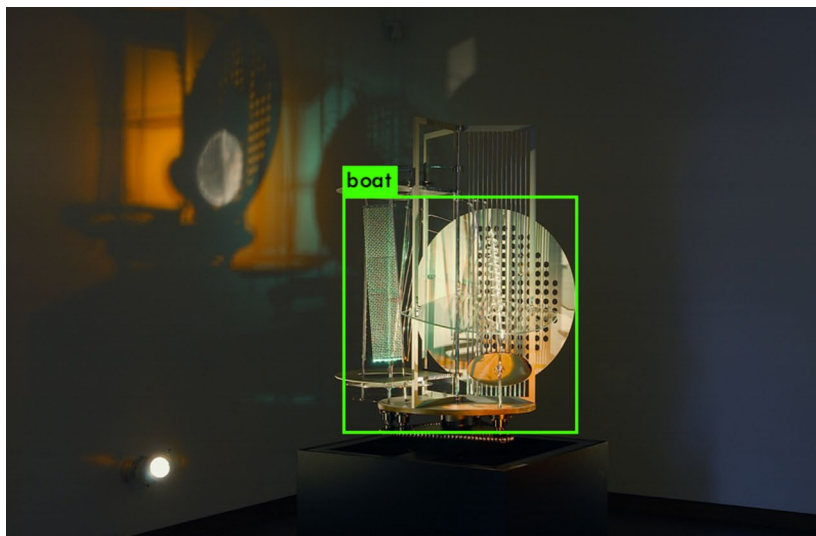
During the analysis process, Moreschi classified works by results with similar characteristics. The same work could be included in more than one of the groups, which are discussed

in the next section. It was only after coding and categorizing the collection’s images that Moreschi visited the museum’s two exhibits in person. During the third week of work, this approach to the artworks happened in the most traditional way: walking through the exhibition space like any other visitor. However, during the fourth and final week of research, the mediation with the works was again denatured during the filming of *Recoding Art*, a short film which integrates the outcomes of this research.

During the last week of work, Pereira and Moreschi worked together in person. Quite familiar with the new collection of artworks that emerged from the AI analyses, we decided to interact with Amazon Mechanical Turkers, as a way to better understand the human layers of AI and avoid the oversimplified idea that AI is completely automatic. These workers are responsible for doing tasks that are still impossible for computers, such as classifying images inside of predefined categories, thus creating the training data for AIs. We surveyed a random sample of Turkers, asking them for descriptions of some of the artworks from the collection, and if they considered these so-called artworks to be art.

<sup>3</sup> It is worth noting that the subject matter of this research (artworks from a collection) is particularly suitable for this approach, since, unlike predictive policing and other egregious algorithmic systems, these errors do not directly cause harm.





#23 LICHT-RAUM MODULATOR (1922-1930) replica 1970 László Moholy-Nagy #23 .jpg a lamp that is lit up at night 49%


Google Cloud Vision Microsoft Azure Amazon Rekognition IBM Watson Facebook Detectron Darknet YOLO Amazon Mechanical Turk

Label Annotations: lighting 77% light fixture 76% lamp 66% still life photography 52% lighting accessory 51%

Safe Search Annotations: Adult: VERY\_UNLIKELY Medical: VERY\_UNLIKELY Racy: VERY\_UNLIKELY Spoof: VERY\_UNLIKELY

Violence: VERY\_UNLIKELY

Best Guess: laszlo moholy nagy light space modulator

Visually Similar: 

Associated Web Entities: Light-Space Modulator 1.0 Light 0.9 Bauhaus 0.6 Art 0.5 Modulation 0.5 Photonics 0.5 Light art 0.5

Artist 0.4 Constructivism 0.4 Design 0.4

**Fig. 4** *Licht-raum Modulator* (1922–1930, replica 1970), by László Moholy-Nagy, described by Microsoft’s AI as “a lamp that is lit up at night,” a similar result to Google’s AI (“lighting,” “lamp,” “light fixtures”). Darknet YOLO (open-source AI) goes further and sees the

work as a possible “boat,” which helps us to construct an interesting speculative hypothesis: that, from the interpretive logic of AIs, the circular reflection on the wall can be a full moon in the high sea

### 3 Denaturalizing art through AI: a possible institutional critique 2.0

The glitches and mistakes of AI help us to denaturalize the art system and its functioning. The art system is highly codified and embedded with power/value systems. When we start considering the AI results that do not necessarily follow the structures of specialized meanings of art, we are unmasking much of what specialized discourses attempts to disguise. In this sense, many of the results obtained from the AIs invited us to think about important issues regarding the art system, which are not always apparent. In addition to that, many of these results can help mediate these works to non-specialized audiences, initiating a more accessible relationship with these objects. Among the results are:

#### 3.1 Art as everyday objects

Interpretations such as these show that artworks are, beyond their discourse, made of materials that are also found outside of the museum context, in everyday life. As

is the case when Duchamp’s *Fountain* is read as an actual urinal, these readings invite us to see works of art in a way that is disconnected from the idea of authorship. To analyze these results is to think about the process of symbolic transformation of artworks, one of the processes that underpin contemporary art. These results, much like the results across the next pages, help to remove the so-called aura from the art object and transform a very important art collection into an assortment of easily recognizable objects (Figs. 4, 5, 6, 7).

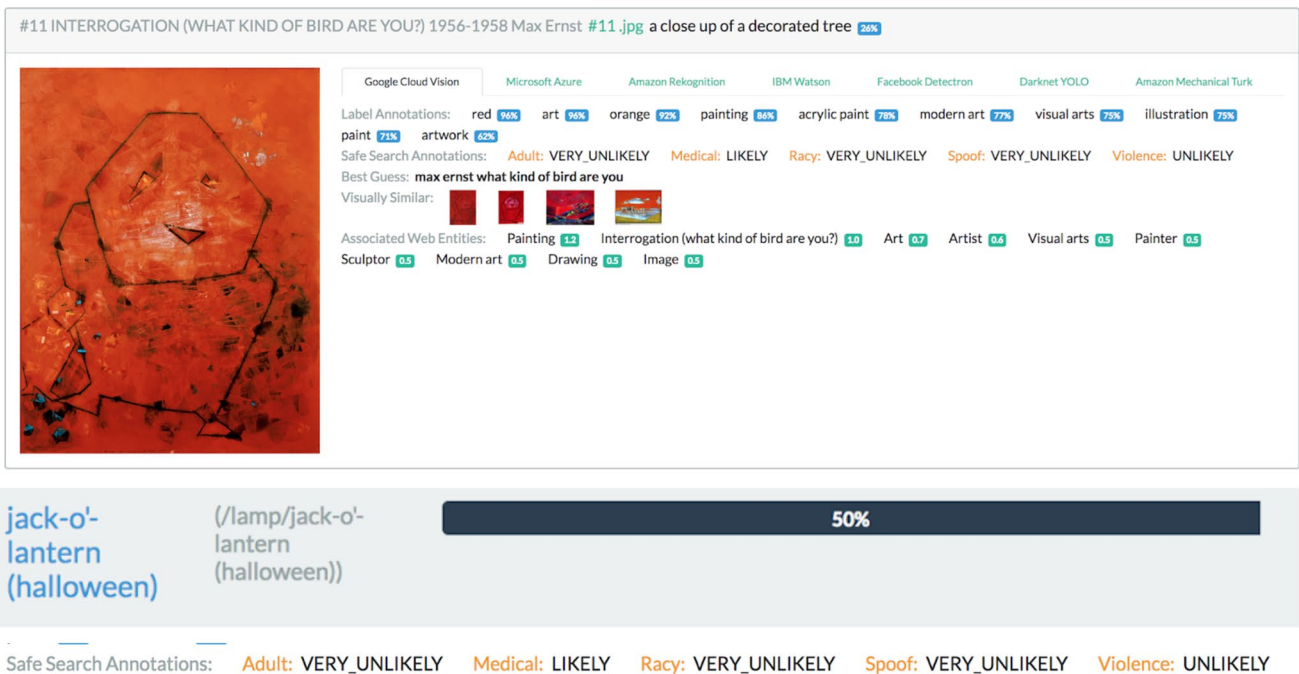
#### 3.2 IKEA shopping cart

The vast majority of the works (almost 90%) were read, in at least one of their results, as consumer products that are easily found in department stores. Such results are valuable in critical art studies for reinforcing the fact that works of art are essentially commodities—even if much more expensive than curtains—and placing our current understanding of what art is within the context of capitalism and consumer society (Figs. 8, 9, 10, 11, 12).



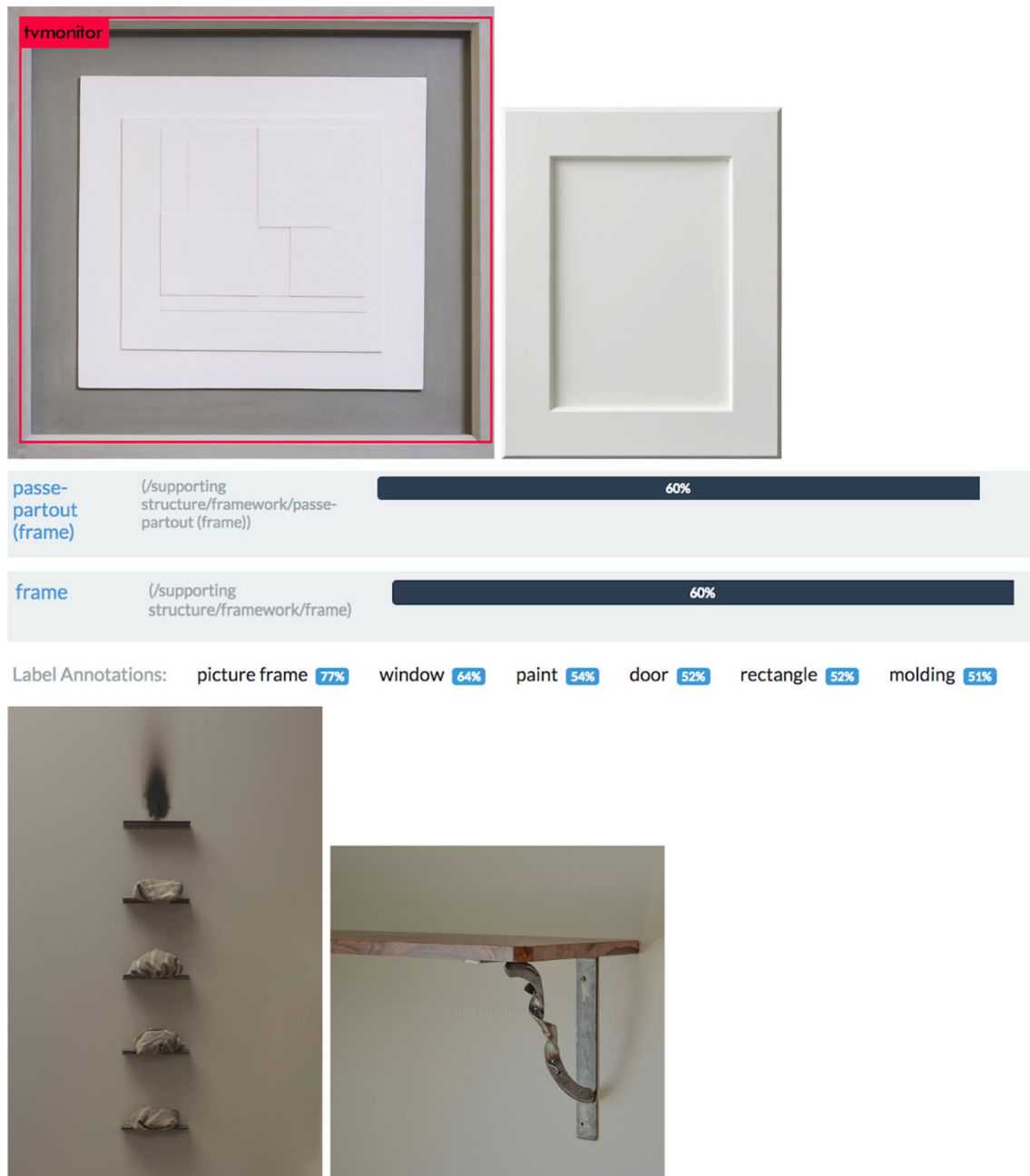
**Fig. 5** *Aux Abords De La Grande Cité* (1960), by Corneille, read as “ejection seat” and “a close up of an old computer.” A considerable part of the works analyzed by Microsoft Azure’s AI is understood as approximations of something. Since the AI’s gaze does not operate

from the human logic of physical distance between the observer and the observed, the concepts of approach and depth radically change here—anything not recognizable at first may indeed be the detail of an everyday object



**Fig. 6** *Interrogation (What Kind of Bird Are You?)* (1956–1958), by Max Ernst, read by IBM’s AI as a “jack-o’-lantern,” the pumpkin that is traditionally carved on Halloween in the United States—an example that shows how the interpretations by AIs are constructed from a U.S. ethnocentric logic. This example also shows that the fact that

works by well-known artists in the collection were read as art does not prevent them from being interpreted as things that are unrelated to the artistic context. It is curious to see the multifaceted ability of AIs to offer, within the same set of results, both the legitimated layer of the image as well as its pre-artistic state



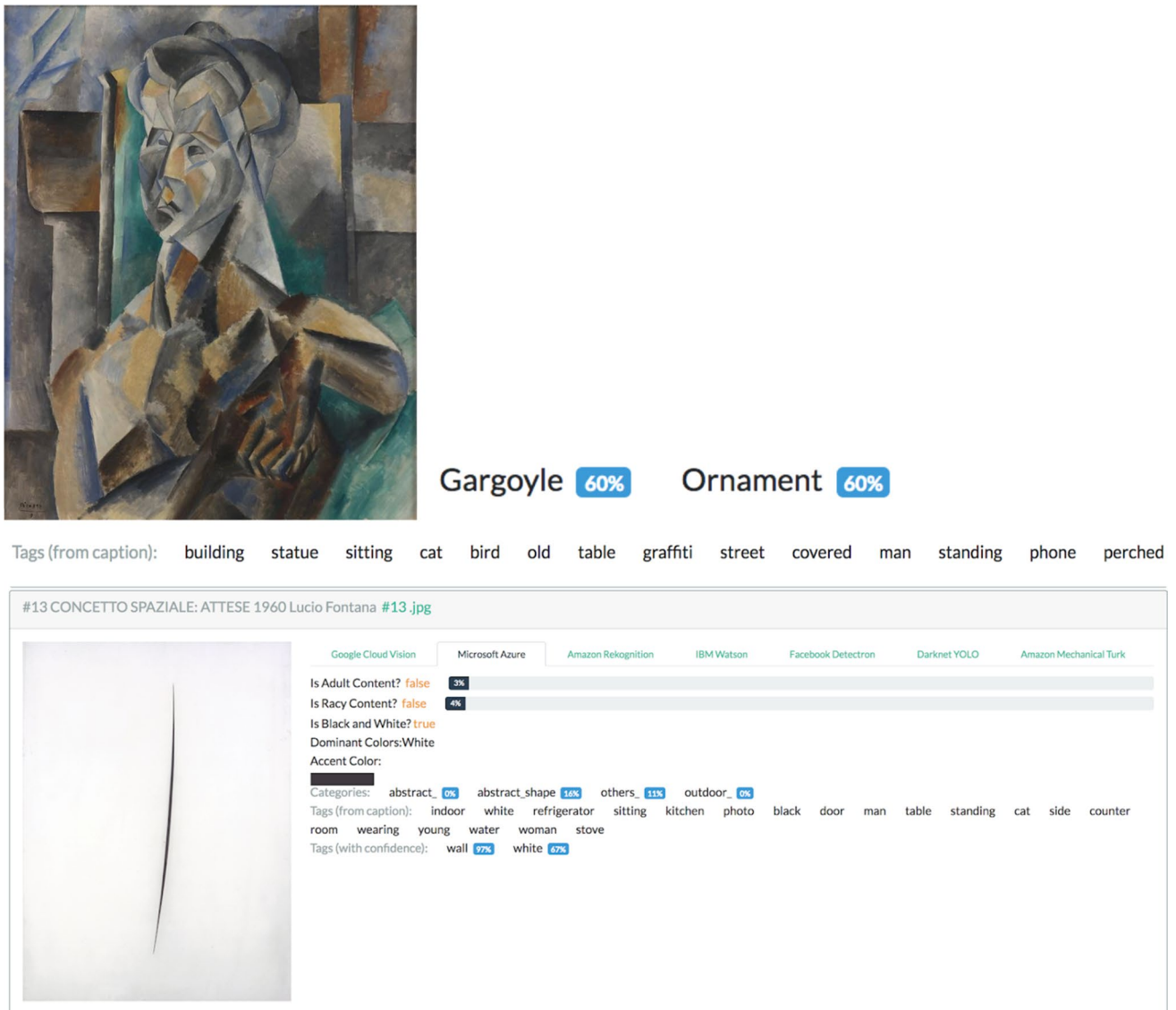
**Fig. 7** Some images received interpretations that prioritize the physical structures that protect or support the works (“picture frame,” “framework,” “supporting structure”) rather than their legitimized artistic contents. This occurred with *White Relief* (1936), by Ben Nicholson, often understood by the AIs as merely a frame, and *Untitled* (1980), by Jannis Kounellis, which, for Google, has to do with the image of a shelf, which in fact is something necessary for

exhibiting the work. Results such as these “de-structure” the hierarchy between layers of the art object that are considered to be artistic and non-artistic, and invite us to view envelopes and bases as part of the artistic structure that is often indispensable in the legitimation of what is art. Framed works are also often read as television monitors, which lead us to the second group of results

### 3.3 Self-promotion

In figurative paintings, AI tends to read people as posing for the camera, which poetically shows how art is a space for human exhibitionism—including selfies and people

practicing sports. These results invite us to think of art (and its contemplation) as an essentially social and egoic practice by human beings, a process of constant self-affirmation (Fig. 13).



**Fig. 8** *Femme en vert* (1909), by Pablo Picasso, as a “gargoyle,” “ornament” and “phone.” And *Concetto Spaziale: Attese* (1960), by Lucio Fontana, as a “refrigerator,” “stove” and “kitchen.”

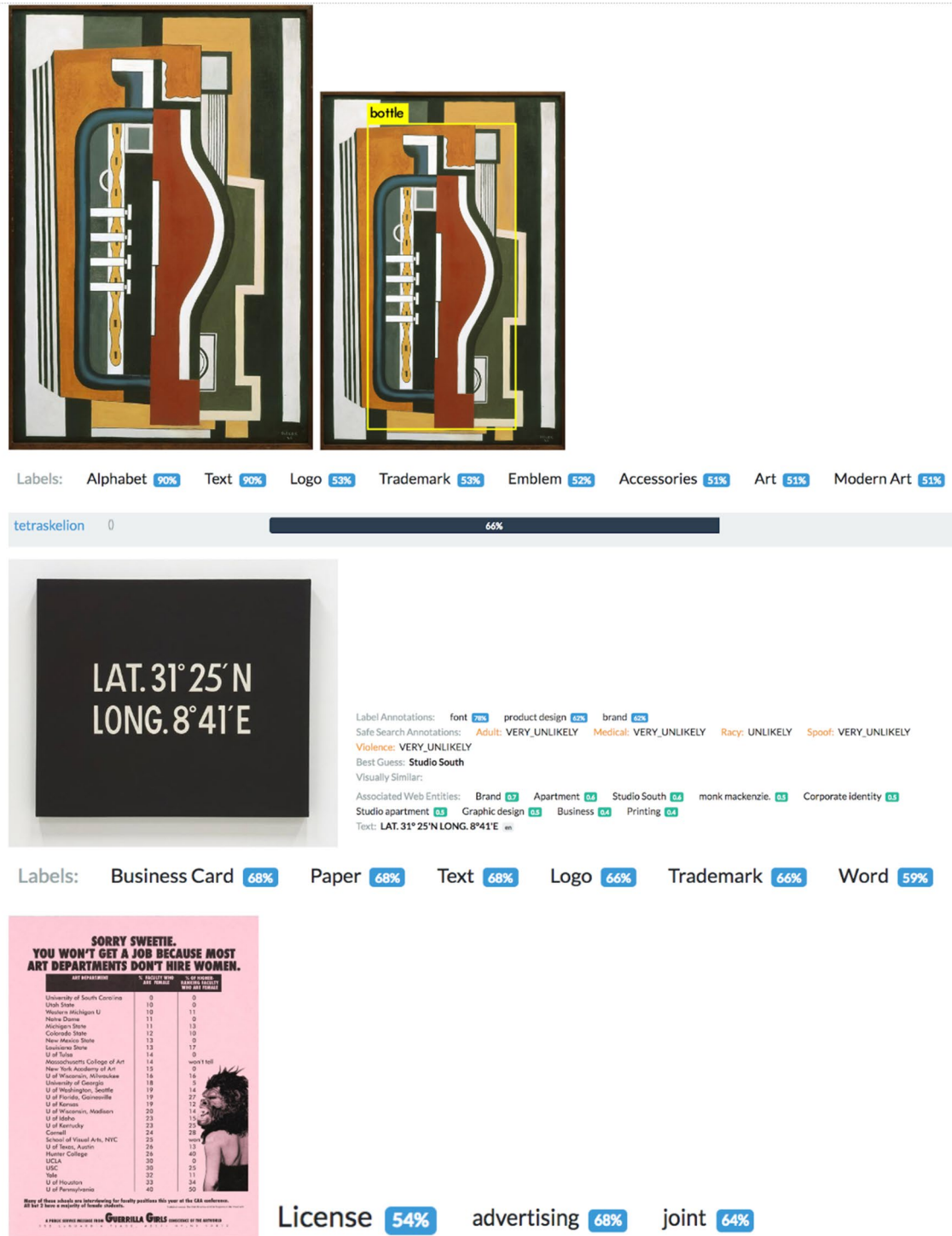
### 3.4 New titles

Microsoft Azure Computer Vision is an AI service that describes images in short sentences. During our experiment at Van Abbemuseum, we performed an exercise in detachment with regard to the artist and their intentions: we began to use these descriptions as new titles for works in the collection. Procedures like this help to demystify the authorship and origin of art objects, creating less fetishized paths of comprehension. Because they are almost always funny, phrases such as these can be valuable material for art classes for non-specialists and young students. Also regarding textual results: Google’s AI sometimes

identifies texts where there actually are none, thus creating curious descriptions (Fig. 14).

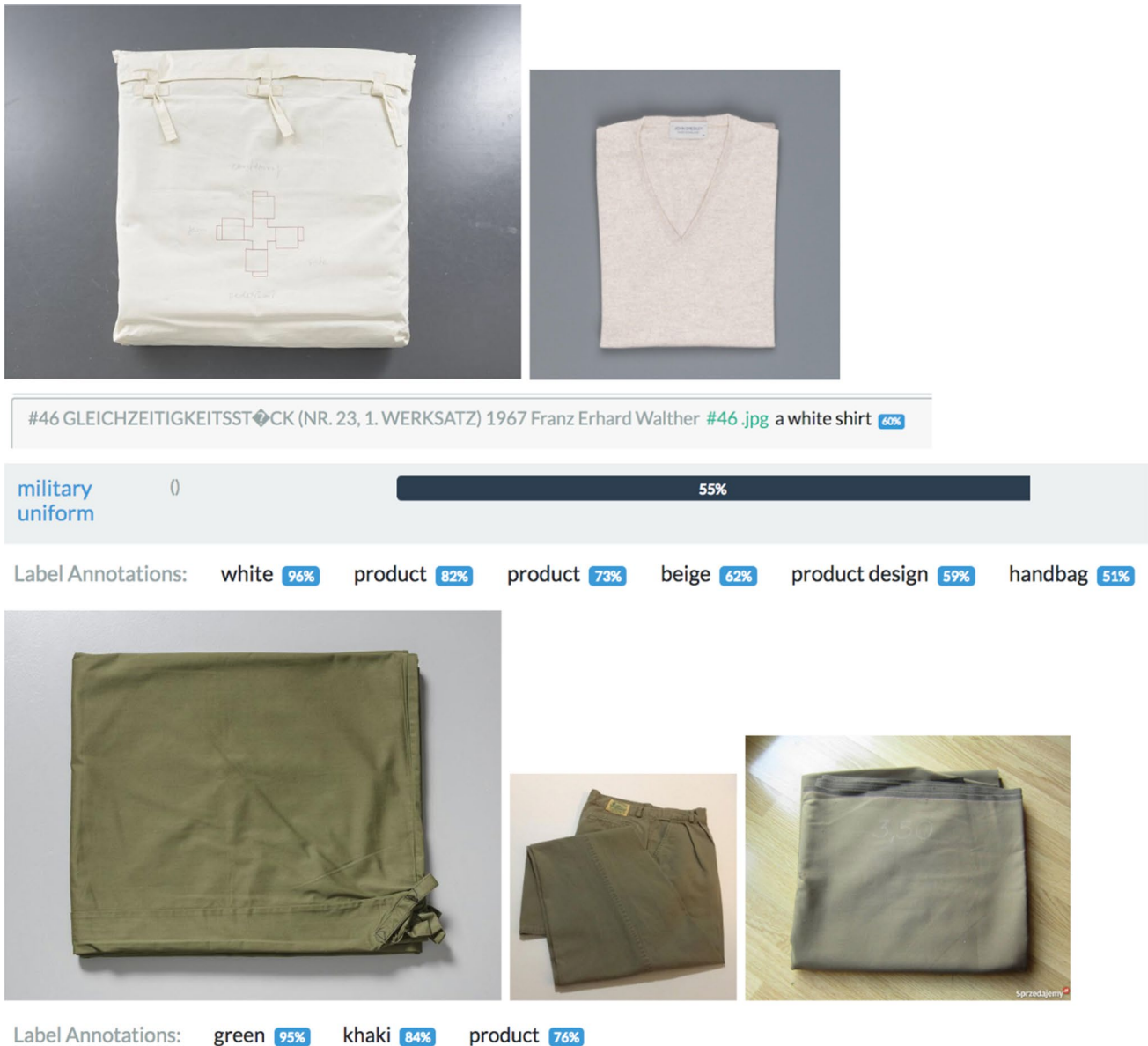
### 3.5 Passages: windows, doors and (why not?) some tables

Poetically, this shows that the space contained by the frame of an artwork creates a space that follows different rules than the space outside of the frame, and that goes on beyond the wall where the artwork is placed—a microsystem that has values and significations of its own. Almost every time there was an interpretation of a “window,” there was also a “TV monitor.” Although it is a typical case



**Fig. 9** Cubist works and those with textual content tend to be related not only to marketable objects (“product design,” “bottle”), but also to specific companies or more general ideas of the business world. This is the case with *L'accordéon* (1926), by Fernand Léger, associated with Tetraskelion Softwares, a company in Jaipur (IN) that offers technological solutions for travel agencies. The same is true for *LAT. 31° 25' N, LONG. 8° 41' E* (1965), by On Kawara, (“brand,”

“business,” “corporate identity”) and the poster *Sorry, Sweetie, Way To Go, Dude!* (1994), by Guerrilla Girls, (“license,” “advertising,” “joint”). This demonstrates that the capitalist logic in AI readings is broader than just interpreting images as products—it also includes notions and practices from consumer society that are not necessarily material



**Fig. 10** *Gleichzeitigkeitsstück (Nr. 23, 1. Werksatz)* (1967), by Franz Erhard Walther as “a white shirt,” “military uniform,” “handbag” and a lot of t-shirt images as visually similar. *Balance (Nr. 26, 1. Werksatz)* (1967), by Franz Erhard Walther as “a bag of luggage,” “clothes” and a lot of trouser images. *Politisch (Nr. 36, 1. Werksatz)* (1967), by Franz Erhard Walther, as “fabric.” Performance fabrics

are almost always read as fashion clothes or accessories by the AIs, which makes some sense, since many of them were worn by artists and/or the public. Here, we have an interesting moment where the AIs actually agree with contemporary art, since most artists, curators, and art critics do not consider these fabrics actual works of art either, but documentary remnants of a previous artistic experience

within the first and second groups of this list (art as objects and IKEA shopping cart), identifying monitors also suggests a depth expansion of the exhibited work. The works read as tables certainly had these results because a framed painting may visually look like a table when seen from above. This recurring result can be seen as an invitation to view paintings from other perspectives, not only face to face or at eye level (Fig. 15).

### 3.6 New temporalities

When AIs do not understand the historical context of an artwork, it allows us to look at art as another kind of object—stripping it away from authorship and historicity. Readings such as these can help in the construction of new narratives of art history, helping to build new associations between societies from different regions and/or periods. Moreover,



Fig. 10 (continued)

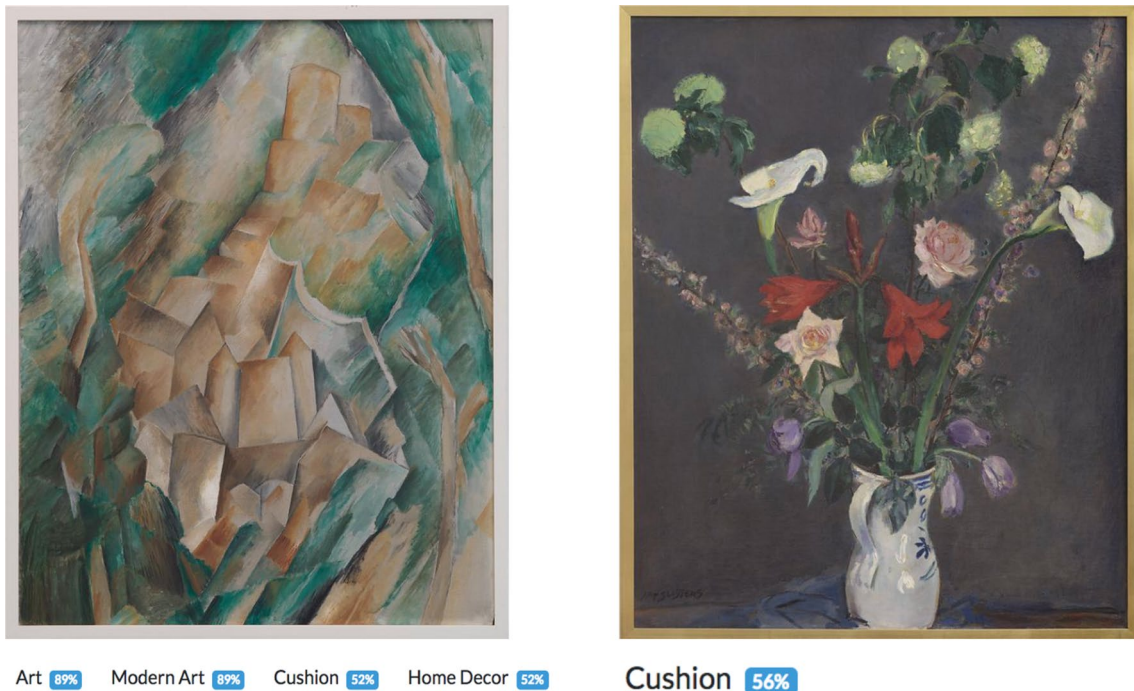
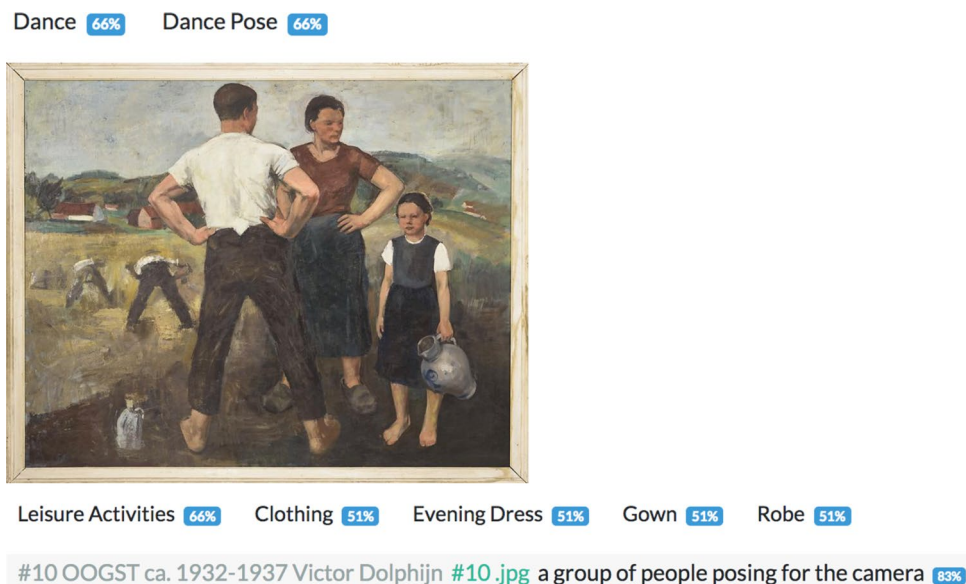


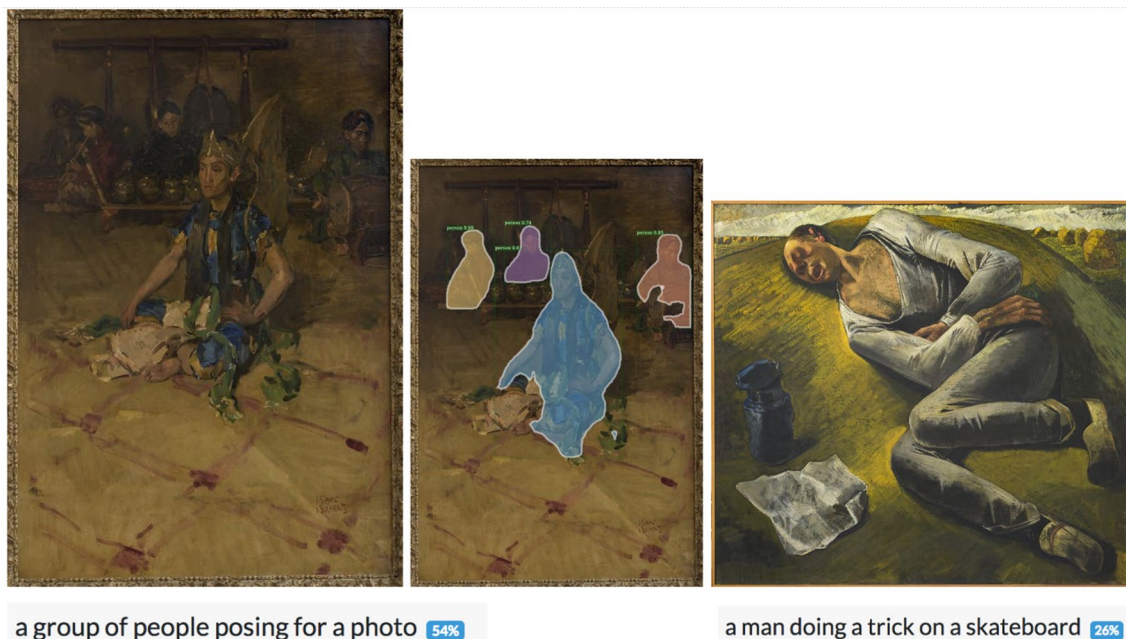
Fig. 11 As with *La Roche-guyon* (1909), by Georges Braque, and *Vaas met Bloemen* (1929), by Jan Sluijters, colorful paintings tend to be read as cushions, which remind us how the visual content of works of art can expand beyond the museum and fit into more popu-

lar, household products. Results such as these also relate to museum shops and their practices of transforming images of artworks into souvenirs



**Fig. 12** As with *Oogst* (ca. 1932–1933), by Victor Dolphijn, images containing people are interpreted based on the objects that appear within them. In virtually every case with human representations, there were results related to their clothing and other personal objects—including moments in which only those objects were identified and not the humans holding them. Results such as these are a

reminder of how part of building an individual’s identity in capitalist society is formed with the help of the objects they possess, and the properties of such objects. The same painting was also described as “a group of people posing for the camera” and as a possible “dance pose,” which brings us to the idea of displaying these products, and to the third category



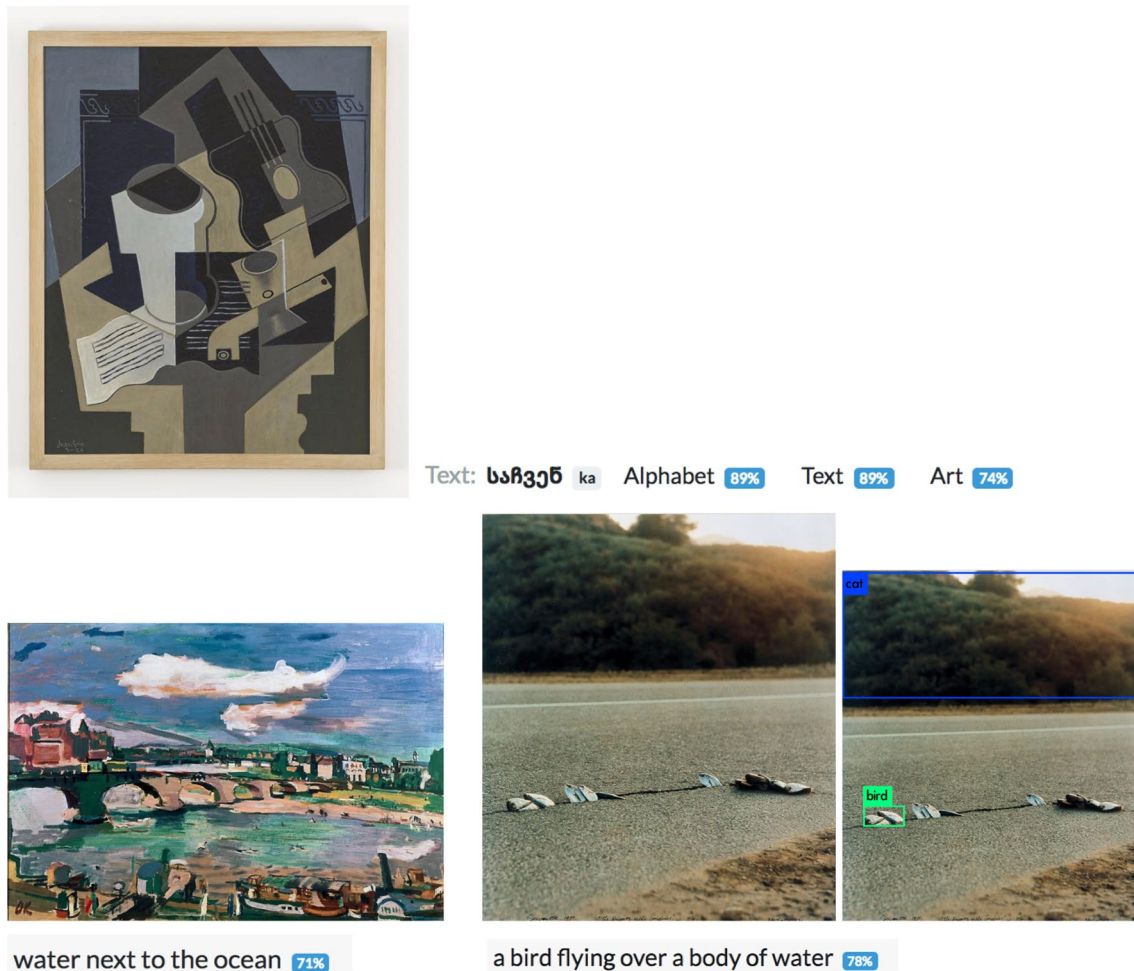
**Fig. 13** *Javaanse Danser* (ca. 1921–1922), by Isaac Israëls, described as “a group of people posing for a photo.” *Slapende Boer* (1936), by Hendrik Chabot, as a skater doing tricks

some of the new temporalities offered by the AIs do not transform the image into something of a different time but suggest more recent or later moments of what is represented there (Figs. 16, 17).

### 3.7 Personification processes

Often images of artworks were read as flesh and blood people, or as performing human tasks. Images read as people





**Fig. 14** According to Google Cloud Vision, the painting *Nature Morte* (1920), by Juan Gris, contains the Georgian word “სსფგბ” which translated into English by Google Translate becomes “display.” *Augustusbrücke Dresden* (1923), by Oskar Kokoschka, was summed

up by Microsoft’s AI as “water next to the ocean,” adding more poetry to the scene. The *Discovery of the Sardines* (1971), by Ger van Elk, is described as “a bird flying over a body of water,” completely reversing the image’s idea of aridity

show how the AI’s understanding system does not differentiate between the concepts of representation and presence. There were also many cases of sculptures (not necessarily human bodies) that were read as people, which emphasize the physical strength of large works. It was also interesting to note the human attributions related to some works, such as a sit-down painting—a typical process of prosopopoeia (Fig. 18).

### 3.8 Visual similarities, new and more democratic possibilities

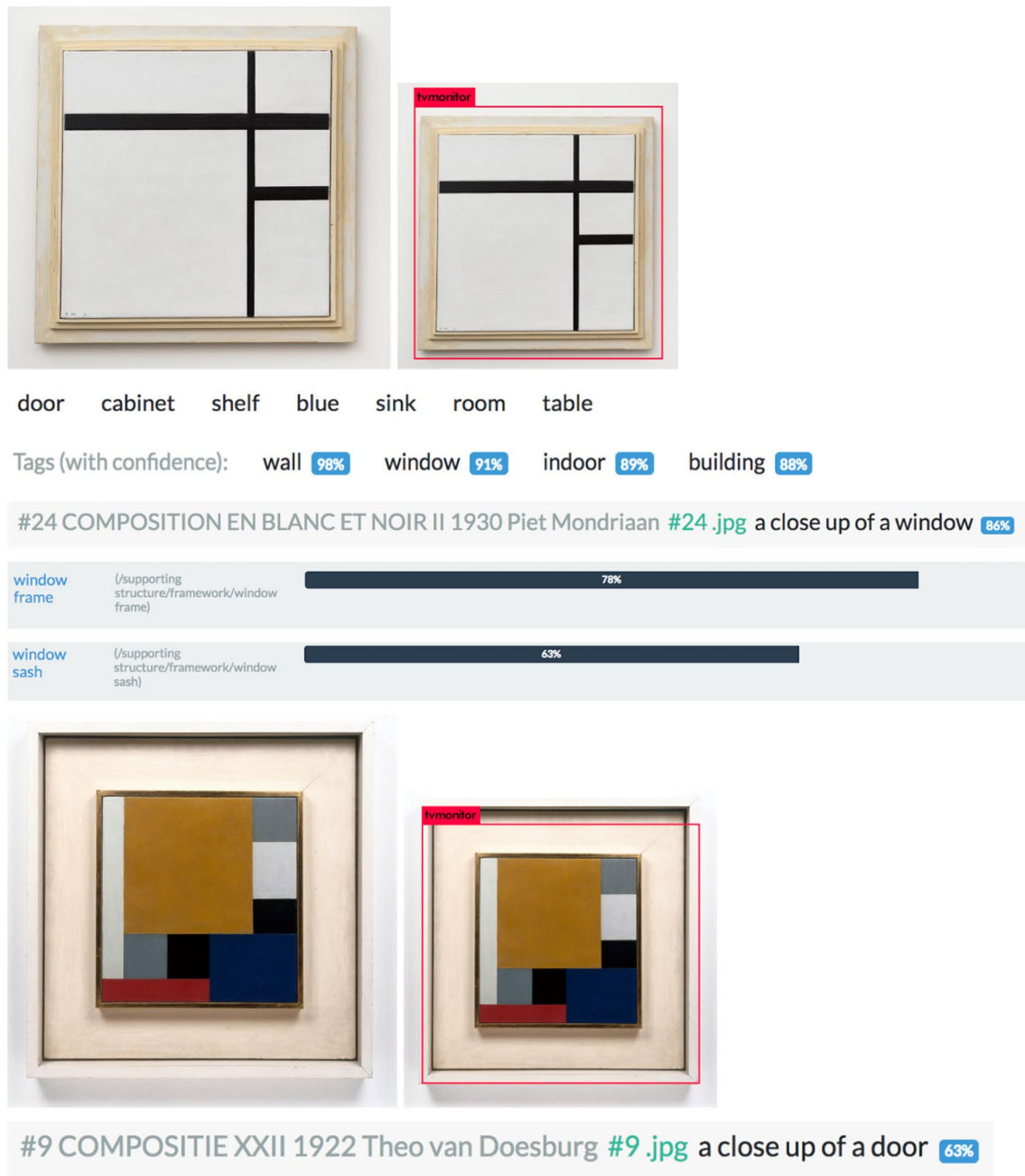
The fact that AIs associate museum artworks with other images of similar visual forms in their databases results in a maximized mode of experiments that have long characterized the study of artistic images. Many of the associative processes of these “intelligences” have to do with practices developed by historians such as Aby Warburg (2010) and his

*Mnemosyne Atlas*. Due to this, considering these results may be important for expanding this field (Fig. 19).

### 3.9 Incomprehensible results, that are very poetic (and that we really like)

Many of the results of the AIs were not fully categorizable into homogeneous groups of results, as is always the case with some works in any museum collection. This shows that art and AI have in common a high load of unpredictability. These results also suggest a possible use of the AI readings in the expansion of the poetic layers of art, contrary to the productivist and efficiency-focused logic of those who argue that AIs must necessarily provide precise results (Fig. 20).

As made evident by the above examples, our experience in using AIs to interpret images of artworks can be seen as new mode of a practice known as “institutional critique.” The term is related to a series of procedures that seek to

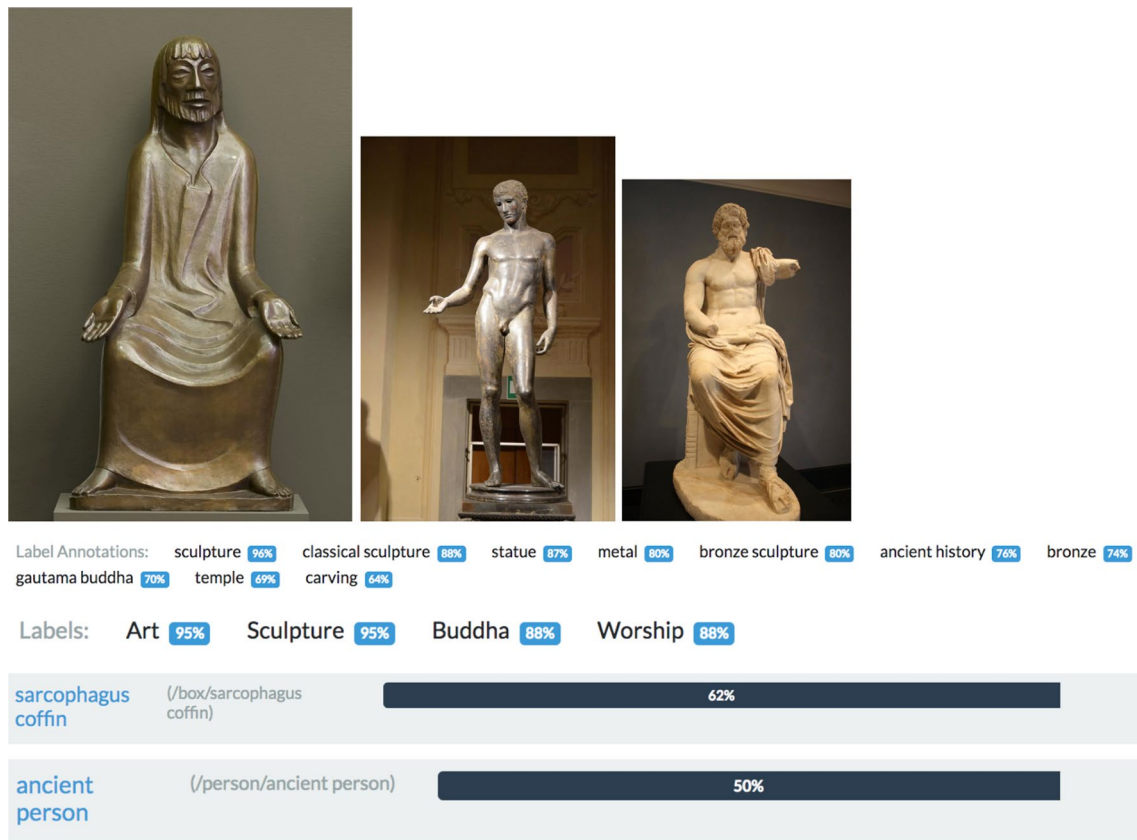


**Fig. 15** *Composition En Blanc Et Noir II* (1930), by Piet Mondriaan, was read as “a close up of a window,” “window frame,” “window sash” and “table.” *Compositie XXII* (1922), by Theo van Doesburg, was read as “a close up of a door.”

reveal the structures that make the art system function. Historically, the institutional critique practice operates from the critical repertoire of conceptual art and conceptualisms of the 1960s and 1970s, especially in the expanded concept of art (Freire 2006).

According to Andrea Fraser (2005), this mode of analytical approximation of art and its elements follows the premise of considering the social context as intrinsic to art—to her, art is never the object of art, but rather a network that

is interconnected with this object of socially constructed elements. Our AI experiments were successful in revealing elements of this construction: when we took photographs of works away from the context of a museum and into that of computer vision algorithms, art seemed to lose its support, and the results obtained were almost never related to the art system. The highly specialized and elitist codes that permeate the artistic field—well protected and validated by powerful actors such as art institutions, curators, gallery owners,



**Fig. 16** *Lehrender Christ* (1931), by Ernst Barlach, read as “Buddha,” “sarcophagus coffin,” and associated to images of Ancient Greek sculptures

specialized critics and even artists—were visibly ignored by computer vision, which instead offered different paths for understanding the artworks.

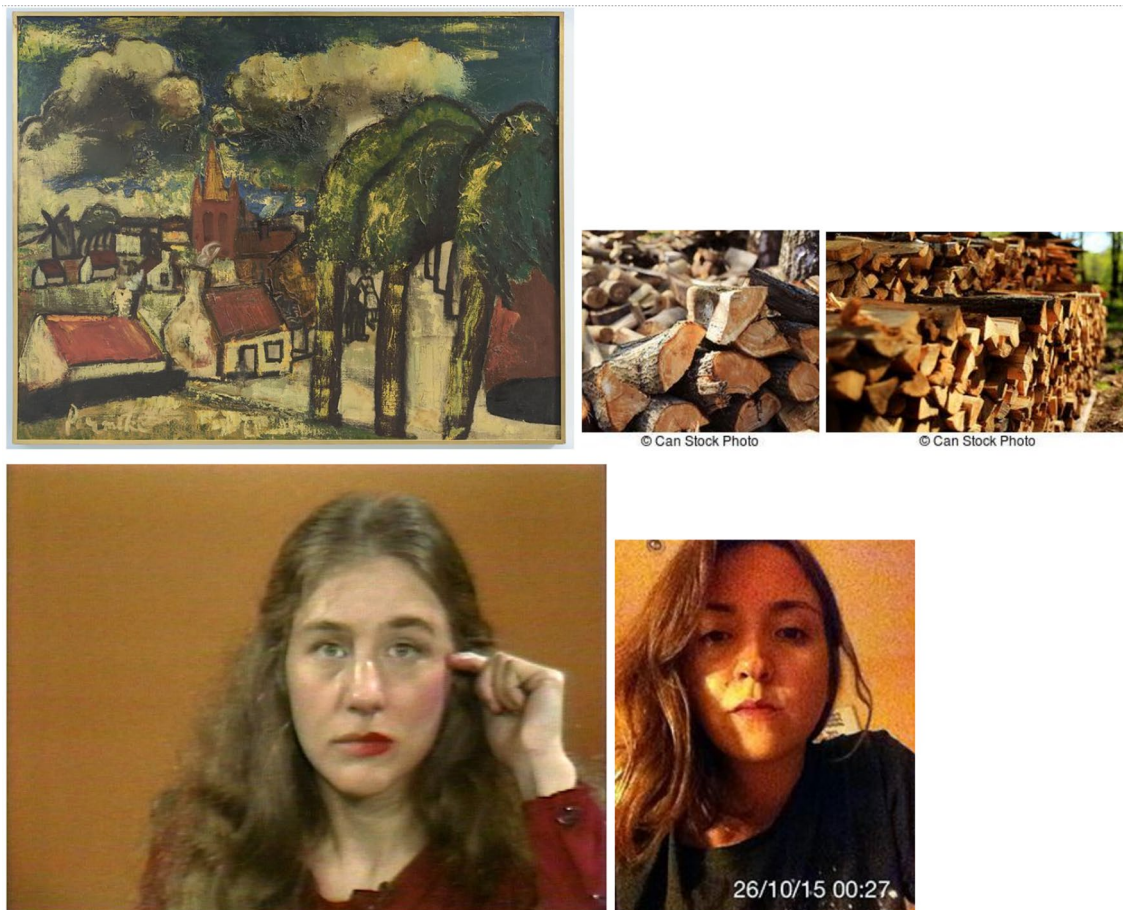
In other words, when looked at without prejudices and with an open mind, seeing art through the glitchy results of computer vision allowed us to distance ourselves from specialized meanings and create relevant materials for the critical study of artistic works and the system they are inserted in. Bringing art and AI together in a critical way serves not only to reveal the latent power structures of the artistic field, but also to democratize it, opening its meaning and significations to the people who engage with art (as espoused by institutional critique). This way of seeing and its potential should be understood and appreciated in its partiality, though, as it intentionally doesn’t engage with the wide context of art history and its specialized discourses.

The unexpected ways computer vision sees art, both by levelling and expanding the potential meanings of artworks, is also particularly innovative and relevant in a moment when visual culture has changed form to hybrids of human–machine cognition and “machine-to-machine seeing” (Paglen 2016), with a plethora of limitations and problems which we address in the following section.

#### 4 Denaturalizing AI through art: looking critically at algorithms

Another possible course of action is to use all the glitches we have just seen to denaturalize AI’s gaze. The results: a list of analyses of every single image of the art collection, when looked at carefully, can work like a reverse engineering of these systems, exposing some of how computer vision “sees” the world. Beyond pointing inefficacies, we can interrogate AI “not only as modes of adjudicating in the world, but also and in their very essence, modes of knowing about the world” (Elish and Boyd 2018: 74). They help question AI’s positioning as a magic “view from nowhere” (Haraway 1988), and the power of their epistemologies and ontologies of understanding the world. This critical reading of AI understands it as one of the many ways of understanding the world that privileges certain values and renders other things invisible.<sup>4</sup>

<sup>4</sup> For more scholarship critically exploring the limitations of computer vision’s ways of seeing see, e.g.: Mintz et al (2019), Buolamwini and Gebru (2018), Crawford and Paglen (2019), and other articles in this special issue.



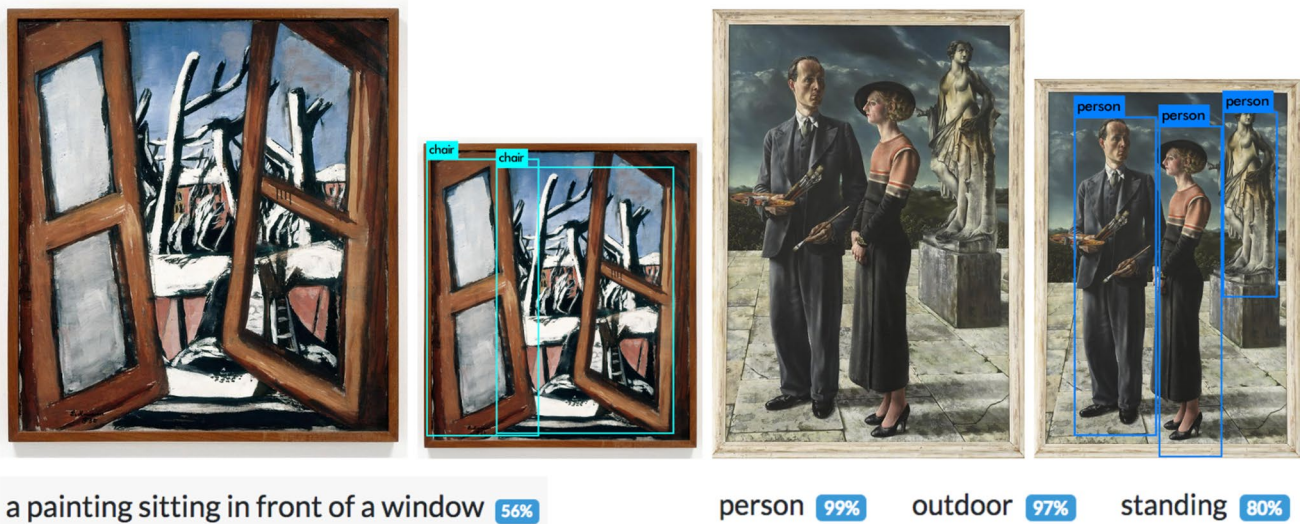
**Fig. 17** *Dorp in de lente* (1936), by Constant Permeke, was associated by Google’s AI to images of firewood, which suggests a later moment for the trees represented in this bucolic painting. The same AI related the frame of the video *Martha Rosler Reads Vogue* (1982),

by Martha Rosler, with the image of a younger woman—it could be a younger Rosler, but in fact is another artist, the Spaniard Cristina Garrido

The commercial AI systems we used, as any other algorithm, are “designed to work without human intervention, they are deliberately obfuscated, and they work with information on a scale that is hard to comprehend” (Gillespie 2014: 192; see also Gillespie 2016). The AIs did not need to stare at an image for seconds, minutes, or even days to assess what it means. Instead, they offer multiple results (often conflicting), alongside “confidence ratios”: a percentage of how much the prediction can be trusted. Moreover, they do not expose how they actually work under the rig, being presented as inscrutable black boxes: their processes are not directly interpretable to the user. Although some of them, i.e. Facebook Detectron and Darknet YOLO, are open-source, they still operate in a highly specialized way that is not inviting to a deeper understanding of the system. And so, as we looked at the results and tried to make sense of them, a few questions kept coming up (some of which were considered in the previous section):

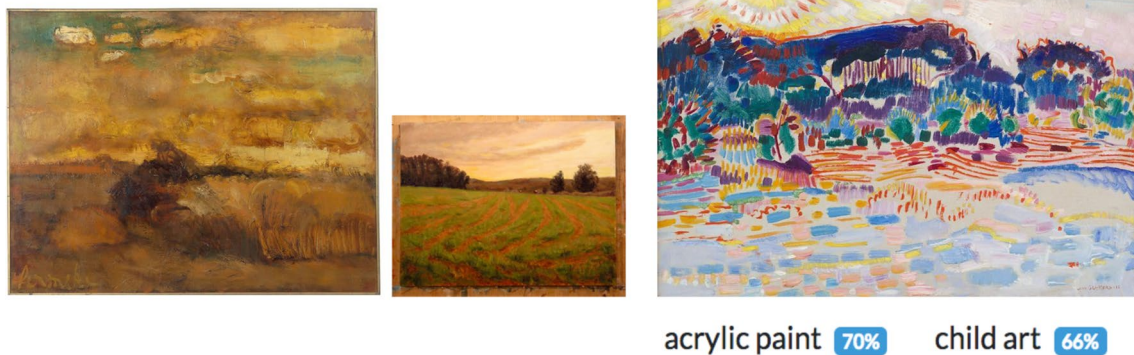
*MORESCHI and PEREIRA: Why so many windows? Why so many tables? Why so many cushions? Why so many close ups? Why so many elephants? Why so many cats? Why so many things related to skate? Why so many computers? Why so many umbrellas? Why so many “Sky plc—company tv cables”?*

These questions become interesting as they expose the issues that underlie contemporary machine learning training datasets. The basis of machine learning, and what makes it different from traditional AI, is the idea that algorithms can, with enough data input, build themselves by making use of large-scale data. In the case of image classification, which is our focus here, “a dataset is used to train a typical machine learning device, a neural net, and the neural net classifies subsequent images probabilistically” (Mackenzie 2017: 4). A “reading” by the AI must be understood as a prediction, which necessarily “depends on classification, and classification itself presumes the existences of classes, and attributes that define membership of classes” (Mackenzie



**Fig. 18** In the painting *Winterbild* (1930), by Max Beckmann, the readings of different AIs complement each other. For Microsoft Azure Computer Vision, the work is “a painting sitting in front of a window.” Seated where, exactly? Probably in one of the two chairs

read by Darknet YOLO’s AI. Similarly, the sculpture in the background of the painting *Schilder Met Zijn Vrouw* (1934), by Carel Willink, was read as a person, standing alongside the couple



**Fig. 19** Some associations of images created approximations of consecrated works with artistic manifestations that are not considered “museum art.” This is the case of *Zomer* (1932), by Constant Permeke, compared by Google’s AI to an amateur painting by an unknown artist. *Landschap* (1910), by Jan Sluijters, was interpreted

as a possible painting by a child, corroborating with the idea that modern art was interested in abstract and unconscious experiences, as opposed to the academic realism of the late nineteenth and early twentieth centuries

2015: 433). The classes within the training datasets, along with the images that compose these classes, are responsible then for defining what the AI can “see.” What’s interesting is how two very disparate things (or “classes”) may become approximated with each other, whether they look like each other or not for our human eyes. For example, a 1936 painting of a man sleeping by Hendrik Chabot (see Fig. 13) is read by Microsoft Azure as “a man doing a trick on a skateboard,” which is not what the image depicts—there is, however, some similarity because of the body’s position, which becomes visible after engaging with the AI’s interpretation.

The prototypical construction of a dataset to enable computer vision occurred through Fei-Fei Li’s *ImageNet* initiative.<sup>5</sup> This project was responsible for gathering a huge number of images (3.2 million images in total), which were originally organized into categories: 12 subtrees with 5247 synsets (Deng et al 2009). To define what these categories

<sup>5</sup> To be clear, not all of the commercially available AIs we used are based on *ImageNet*, but the project was responsible for triggering a spark. By providing plenty of data about objects and their properties, and creating multiple competitions around it, the field became legiti-



**Fig. 20** In the painting, *Moeder en Kind* (1922), by Gus de Smet, an elephant (marked in blue) is read in the room by Facebook’s AI. This was also one of the beautiful cases in which a work was read as a “mirror,” referring to the idea that the understanding of an artwork is a reflection, a consequence of the way of thinking of those who look at it. Microsoft’s AI went beyond the idea of object and added

in the conceptual work *B 12,000,030 = 25 = 16X17 = NOIR BLANC BLEU* (1975), by André Cadere, the information “air”—the true context of art and all other things of this world. But, of course, since not everything is poetry in the AIs, Google has associated this conceptual work to the image of a lamp

would be, the project made use of a previously existing structure called *WordNet*. Created in 1985 by Stanford psychology professor George Armitage Miller, with funding from the military and DARPA, *WordNet* was devised to work like a dictionary, but one in which words existed in relation to others (not in alphabetical order). This index of words in a machine-readable logic would become the categories used for all images: dogs, pudding, tracks, excavation. Behind *ImageNet* is a desire to have “more data” for training machine learning algorithms, thus allowing them to recognize more objects in images. In the dataset, for example, there are 1289 images of skateboards: “A board with wheels that is ridden in a standing or crouching position and propelled by foot.” (ImageNet 2019)

*ImageNet* took images from the photo-sharing website Flickr, where users upload their personal photos and often choose to keep them copyright-free, allowing others to use them. It is interesting to think about the origin of dataset’s images and how they make direct use of user-generated data, crowd-produced by all of us. As Cheney-Lippold (2011: 178) warns, the “algorithm ultimately exercises control over us by harnessing these forces through the creation of relationships between real-world surveillance data [Flickr,

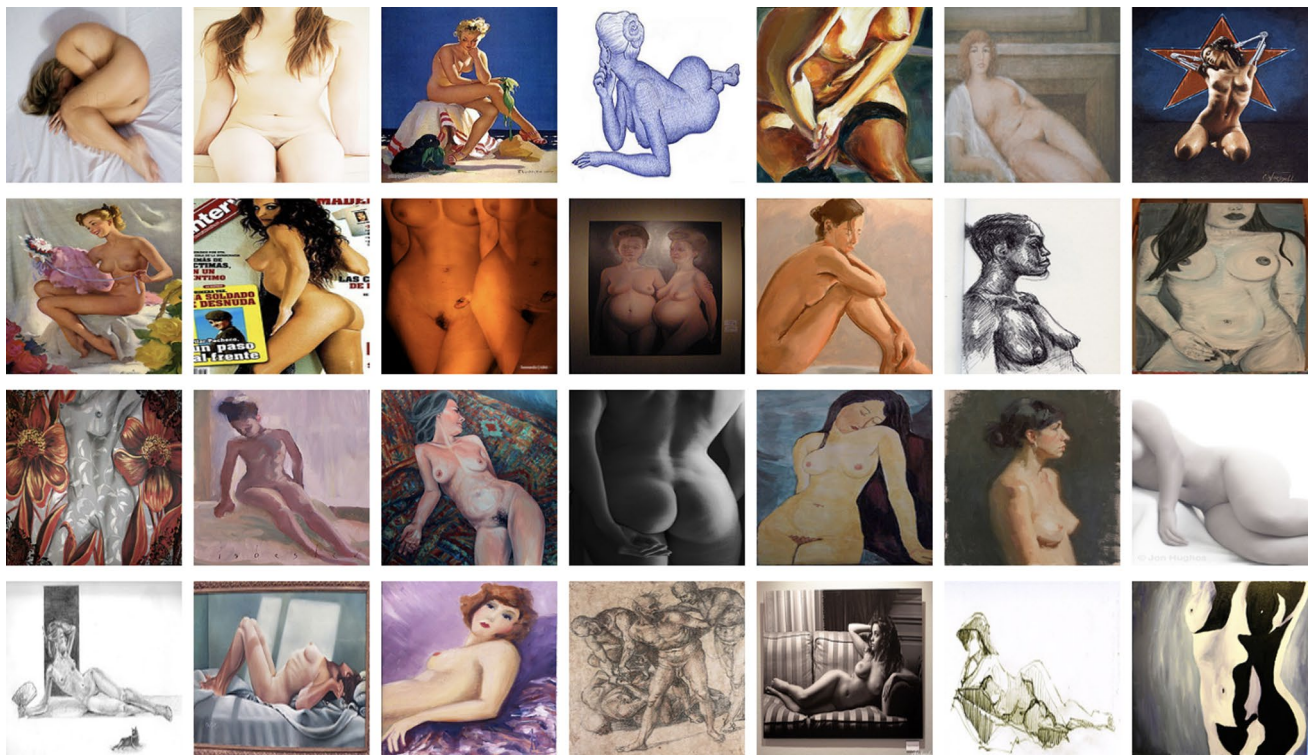
in this case] and machines capable of making statistically relevant inferences about what that data can mean.” In the end, the ways of seeing of these computer vision AIs are directly tied to the origin of their datasets.

As can be noted in our results, art is not a “material” that commercial AIs are extensively trained on. Images of art and artistic works are only a tiny fraction of what users upload to Flickr, which mainly consists of imagery of contemporary life and social practices, mostly from the United States, where Flickr is most popular (estimated over 25% of its content). We have seen previously how AIs read an immense number of artworks as products, especially department store or home decor products. These are highly accessible, but diverse in value (which is slightly ironic when talking about artworks): tables, shelves, curtains, refrigerator, furniture, cushions, clothing, mobile phones, computer/laptop, TV monitors, etc. These products are part of the wide catalogue of stores such as IKEA, but also present in the modern home imagery. Besides products, the high frequency of sports, cats, dogs, selfies, mirrors, are other indications of the origin of these images in contemporary day-to-day life. The fact that artworks are read as that (although they seldom represent these things) attests to the high frequency of these things in the original training databases.

These readings point not only to the origin of images, but to the way images become useful for the clients of commercial AIs. Google, Amazon, Facebook, and others build their AIs to identify, categorize, and see the world

Footnote 5 (continued)

mated and useful for the industry. If an AI does not use it, it is certainly made in connection to it.



**Fig. 21** Screenshot of some of the images that are under the synset “Nude, nude painting” (“A painting of a naked human figure”) in ImageNet. There are a total of 1229 images, most of which represent a sexualized, naked, thin, white woman

as commodities.<sup>6</sup> Other frequent results, such as business cards, advertisements, and billboards are not products per se, but are also related to contemporary capitalist life. All of this reasserts the way these AI systems embed values that are ideologically capitalist and focused on value-production for the companies that use it.<sup>7</sup> A client featured on Google Cloud Vision’s home page, Urban Outfitters, an American multinational lifestyle retail corporation (i.e. clothes shop), uses the AI system to “automate the product attribution process by recognizing nuanced product characteristics like patterns and neckline styles” (Google 2019). No wonder the performance fabrics seen before (Fig. 10) are almost always read as fashion clothes or accessories, metaphorically transformed into Urban Outfitter’s hipster turtleneck long sleeve t-shirts.

<sup>6</sup> And to support the military, but this arguably happens through other systems based on the commercially available ones; or through military grants, which also underlie the whole system.

<sup>7</sup> A possible consequence of this is: why are museums using these same AIs, in so many projects with Google Arts & Culture, for example?

AIs also often identify “raciness” in abstract sculptures, and where there are images of women, undressed or not.<sup>8</sup> Although these images come from a context that is considered different from a pornographic image, they nonetheless fall under the same category and classification. The consequence of this is felt when these same machine learning systems are used for content moderation, often without subsequent human analysis, thus frequently determining artworks as pornographic (see Gillespie 2018). When we look at *ImageNet* as a prototypical training database, the nude female body appears in a sexualized way, an object of the male gaze (see Fig. 21), not too differently than throughout much of the history of art (see Parker and Pollock 2013). This understanding of the female body, among other problematic categories/classes, trickles down to the systems that are used for image recognition in social media and other platforms, through the training data they use (see Crawford and Paglen 2019). This is just one way that AI systems flatten images of artworks by ignoring their context, or what their images mean in a broader sense.

<sup>8</sup> Images with nude women, as in the painting *Liggend Naakt* (1931), by Jan Sluijters, or even dressed, as in *Moeder en Kind* (1922), by Gust de Smet, and *Boerderij* (1919), by Heinrich Campendonk. The same has also happened with images of more abstract sculptures, perhaps because of possibly phallic shapes, such as in *My neck, my back curve silently* (1930), by Karin Arink.

Taking this even further, we argue that the obsession of these AIs with nakedness through categories such as “racy” (from Google Cloud Vision) and “pornographic” conceals their lack of contextual comprehension through a numerical result, thus normalizing problematic categories. Bowker and Star (2000: 35) in their classic book *Sorting Things Out*, discuss classification as a way of seeing the world and naming it. They are interested in “how basic categories and standards are formed, and how they are formed as ordinary,” to examine the power these systems embody. They state:

The advantaged are those whose place in a set of classification systems is a powerful one and for whom powerful sets of classifications of knowledge appear natural. For these people, the infrastructures that together support and construct their identities operate particularly smoothly (though never fully so). For others, the fitting process of being able to use the infrastructures takes a terrible toll. To “act naturally,” they have to reclassify and be reclassified socially (2000: 225).

AI’s categories have a lot of power, as they become embedded in our everyday world, as previously discussed by many other scholars (see Noble 2018; Eubanks 2018; O’Neill 2016; D’Ignazio and Klein 2020; Crawford 2018). This certainly happens as AI systems become carelessly incorporated into public services, such as policing (Brayne 2017), but also more subtly through image recognition in our day-to-day life. It is imperative to recognize that categories such as racy, pornographic, etc. do not engage with any context or meaning behind images, and thus can ever only be of partial use to really engaging with visual culture. But even when the machine is not sure, it generates results; leaving no possibility of not knowing, contrary to art. As discussed by Amoore (2019), by transforming doubt into weighted probabilities, a number between 0 and 1, “the single output of the machine learning algorithm is rendered as a decision placed beyond doubt; a risk score or target that is to be actioned.” This epistemology confounds correlation with causation: formal similarities are the way to understanding, and if something looks like something else, it must be that. Jesus Christ being read as a man (or even a woman) shows exactly how the surface-level reading does not engage with a more embodied, cultural, and contextual reading of reality. This is a crucial way of showing that AI represents only one of the many ways of knowing and working through the world.

## 5 AIs and the exploitation of labor in technological capitalism

*MORESCHI: On the last night of work, I asked Gabriel an essential question: What, in this process, was not machine, but human? That’s when he told me about Turkers. People*

*who are paid meager cents of a dollar to categorize images in systems such as ImageNet. We wanted to talk to them to understand how they made their choices. We requested a survey of the collection’s images, sent to them without any additional information. We asked them to describe the images and whether or not they considered those images to be art (Fig. 22).*

In our experiment with these “lower levels” of the AI stack—the Turkers—we were able to get a glimpse of how datasets such as *ImageNet* are built. Although AI systems are branded as an external, non-human, objective, “view from nowhere” (Haraway 1988), as its mathematical and statistical analysis claims its way of understanding the world as the only possible way (rationalization), they have their origins in these workers, spread out throughout the world (75% in the US, 16% in India, and many other countries in a lesser degree; see Difallah et al. 2018), who are being paid meager salaries. It is imperative to make clear that AI’s backbone is constituted by human thought, labor and clicks on a screen: these systems are built on “workers’ invisibility” (Irani and Silberman 2013). Platforms such as Amazon Mechanical Turk “commercialize the thesis that humans are important cogs in computational machines” (Finn 2017) and treat them as such.

More than invisible, Gray and Suri (2019) define the work of Turkers as “ghost work” in their book-long review of on-demand digital work performed by this vast, invisible human labor force. This term highlights the central irony of how on-demand work is prevalent today, hidden in the shadows of so-called artificial intelligence. It is this “ghost” aspect that makes it so that we often cannot see what is behind the scenes, hiding AI’s materiality (see Crawford and Joler 2018). As Gray and Suri explain, “Mturk workers are the AI revolution’s unsung heroes. (...) Humans trained an AI only to have the AI ultimately take over the task entirely” (2019: 8).

When asked to describe images of artworks from the collection, Turkers presented quick, direct, and not particularly analytical readings of the visual material. As they attempt to complete the HIT (Human Intelligence Task) in the shortest time possible, which, in this case, ended up being just over a couple of minutes, they engage with the image with a particular distance, not looking for or relating it to any context that is not offered, or even from their own perspective. The descriptions attempt to maintain a detachment and objectivity. Their answer to what they think of the artwork is also quick and direct, exposing the very brief relation with the image: it is “bland”, “sexual”, “ugly”, “classy”, or something else.

The “artificial artificial intelligence,” as Turkers are described by Amazon (Finn 2017), give many of the same responses as we have seen with the AIs that were previously analyzed: they seem to treat things based on what they look





<span>Google Cloud Vision</span> <span>Microsoft Azure</span> <span>Amazon Rekognition</span> <span>IBM Watson</span> <span>Facebook Detectron</span> <span>Darknet YOLO</span> <span>Amazon Mechanical Turk</span>						
What do you think	Description	Is this Art?	Have you been paid fairly?	Completion time (seconds)	Reward	
I think it is a classy nude painting	<ul style="list-style-type: none"> <li>nude woman laying on bed with back facing viewer</li> <li>woman in the nude on a bed with back facing viewer</li> <li>painting of a nude woman on a bed with back facing viewer</li> </ul>	yes	yes	115	\$0.40	
It is very well composed and has a great palate of color	<ul style="list-style-type: none"> <li>woman lies prone in bed</li> <li>woman resting, soundly</li> <li>a woman at rest</li> </ul>	yes	yes	132	\$0.40	
I think that the woman is ugly.	<ul style="list-style-type: none"> <li>There is a woman naked, facing the bed, laying down.</li> <li>Laying down on white sheets is a naked woman on a bed.</li> <li>The woman without clothing is holding her arm over a pillow, while naked.</li> </ul>	yes	no	159	\$0.40	
Very sexual Sleeping	<ul style="list-style-type: none"> <li>Sleeping on the bed.</li> <li>Very Nude showing.</li> <li>Without dress sleeping on the bed.</li> </ul>	yes	yes	161	\$0.40	
It's bland.	<ul style="list-style-type: none"> <li>Woman sleeping on couch</li> <li>Naked woman sleeping on couch</li> <li>Nude form from behind</li> </ul>	yes	no	124	\$0.40	

**Fig. 22** Painting *Liggend Naakt* (1931), by Jan Sluijters, as described by five different Turkers. One of them thinks it is “very sexual,” while another says “the woman is ugly.” All of them take just over 2 min to complete their task. They all agree: the image is Art

**Fig. 23** A scene from *Recording Art* (a short film based on this research) wherein Turkers analyze works of art. The film premiered at IDFA Competition for Short Documentary 2019, Amsterdam



like, a “first-impression,” made in the haste of their race for generating enough income. What is essential in these responses, though, is that Turkers “at the heart of this system not only take on the challenge of endless micro-tasks managing ambiguity—they also take on the affective work of acting as a human element inside of a computational application” (Finn 2017). By this, it is meant that HITs expect from Turkers “a response that is both mechanically reliable and reliably human... constantly negotiating between computational and cultural regimes of meaning.” In the end, these descriptions and opinions about artworks may sound like they are coming from the AIs, as they are made manifest through a direct, objective, distant format that fits into the

expected computational tone, but they are generated by a real human seeing, their human mind, culture, idiosyncrasies, and context.

“Artificial artificial intelligence” is only valued because it is intrinsically different from artificial intelligence. Unlike AI, Turkers are humans: they have feelings, opinions, agency, families, and bills to pay. Although directly participating in a precarious, neoliberal, pro-employer marketplace, where they have very little say or protection, and are identified as an alphanumeric code, their responses must be understood as human: they come from a different position of power than that of AI, a position which must be understood critically. When Irani and Silberman (2013)

asked Turkers what their major concerns were with the platform, workers pointed to Amazon and how it does not really care about them, exposing a clear power imbalance where the needs of workers are not prioritized. As put by one of their Turker respondents, “I don’t care about the penny I didn’t earn for knowing the difference between an apple and a giraffe, but I’m angry that MT will take requester’s money but not manage, oversee, or mediate the problems and injustices on their site.” (2013: 615) The difference between an apple and a giraffe, between an artwork or something else, when analyzed by Turkers, and later by the AIs that are built from their labor, must all be understood as part of this larger system of worker invisibility, low pay, lack of rights, instability, and Amazon’s political economy of monopoly.

*PEREIRA: Thinking about non-specialists in AI, we found it important to show the human labor behind these machines, as a way of raising awareness of what AI actually is. In the short video we made (Fig. 23), we tried to use an accessible language to show this infrastructure, explain the role of Turkers, and use their own voices to read descriptions they would give to images of artworks, thus foregrounding their contribution to AI systems. At the same time, we denaturalize artificial intelligences as both AI and human Turker readings are shown to have similarities (as well as differences).*

This process not only demonstrated the importance of bringing these discussions of critical infrastructure to society in general, but also the importance of creating interdisciplinary research teams that can operate from different backgrounds and perspectives. In this moment where image recognition algorithms become embedded across society through social media platforms, as well as through the use of technology by museums and archives, we must be attentive as to how these algorithms operate, what they obfuscate, and which kinds of invisible labor they rely on. And, more radically, we must be concerned with the increasing pervasiveness of the logic of the algorithm, the “if–then” causal understanding of categories that confuse seeing with understanding, correlation and causation, nudity and pornography. As described by Simanowski (2016: 55), “If reason is reduced to formal logic, then any discussion becomes unnecessary because if–then relations do not leave any room for the ‘but’ or the ‘nevertheless,’ nor for ambivalence, irony, or skepticism; algorithms are indifferent to context, delegating decisions to predetermined principles.”

Our experiment with denaturalizing AI through art points exactly to this intrinsic instability of the “artificial” in “artificial intelligence”: the lack of seeing more than what things are. The artistic space, with its openness to different logics and “ways of seeing,” serves as fertile ground for exposing AI’s binary, capitalist, and value-laden gaze. At the same time, however, it also provokes us to reconsider, to be more

creative and explorative. What are we really seeing in an artwork? And how could we ever be so sure?

## 6 Conclusion

The process of seeing paintings, or seeing anything else, is less spontaneous and natural than we tend to believe. A large part of seeing depends upon habit and convention (Berger 2008).

We have explained in this article how commercially available AIs can work, through their glitches, to level and reimagine artworks, giving us a fresh set of eyes for understanding art. At the same time, these very glitches can serve as a peek down the stack of ever-present algorithmic image recognition systems, helping us speculate on their inner workings and critique their limited perspectives.

Art is historically formed through internal deconstructions, some of which become paradigmatic like Duchamp’s *Fountain*. These processes of self-critique and analysis are essential for the field and its relation to the world, as notably proposed by institutional critique. The intermingling of art and AI described in this article continues this self-critical artistic practice in a time of “ways of machine seeing.” Likewise, this research also aims at incorporating the creative practices from the artistic world into the AI field, provoking it to be more (self-)critical and experimental.

We present our research in this article, alongside different artistic outputs, as a way of experimenting with research methodologies and interdisciplinary positions. This research was also presented at the museum whose archive we analyzed, creating a crack for reflexivity within an elitist, codified space. This can be seen as a contribution to art education in museums: we propose that people, as exhibition visitors, be invited to experiment with the distanced look of AI as a way of critically thinking about the art system. We hope such practices help to create new relationships, openings, and connections for those who are non-specialists to explore art critically. Instead of museums using commercially available AI from Big Tech in an uncritical way, why not make more radical and creative uses of technology?

As AI continues to expand, change and “improve,” we understand these results in their limitations: they are a snapshot of how they worked when we tested them. As more and more data are produced in our everyday lives and interactions, and as companies continually train their models, it also continually changes how art is read. In our experience, throughout our research, we have seen both minor and major changes, which we think point to the simultaneously productive and critical instability of AI and art.

**Acknowledgements** Portions of this article appeared in a different version in the Van Abbemuseum’s “Deviant Practice Research Programme 2018-19” electronic publication (CC BY-NC-ND 4.0). We’d like to thank the special issue editors, reviewers, and others that have contributed and supported this research project. Special thanks to Giselle Beiguelman, the staff of the Van Abbemuseum (especially Nick Aikens, Evelien Scheltinga and Christiane Berndes), and the Center for Arts, Design and Social Research.

**Funding** This research has received funding from the Deviant Practice Research Programme at the Van Abbemuseum (Netherlands), and the Center for Arts, Design and Social Research.

## References

- Amoore L (2019) Doubt and the algorithm: on the partial accounts of machine learning. *Theory Cult Soc* 36(6):147–169. <https://doi.org/10.1177/0263276419851846>
- Berger J (2008) *Ways of seeing*. Penguin, London
- Bowker GC, Star SL (2000) *Sorting things out: classification and its consequences*. MIT Press, Cambridge
- Brayne S (2017) Big data surveillance: the case of policing. *Am Sociol Rev* 82(5):977–1008. <https://doi.org/10.1177/0003122417725865>
- Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. 1st conference on fairness, accountability and transparency. PMLR 81:77–91
- Cheney-Lippold J (2011) A new algorithmic identity. *Theory Cult Soc* 28(6):164–181. <https://doi.org/10.1177/0263276411424420>
- Cox G (2017) Ways of machine seeing: an introduction. *Peer Rev J About* 6(1). <https://www.aprja.net/ways-of-machine-seeing-an-introduction/>
- Crawford K (2018) AI Now: social and political questions for artificial intelligence. Distinguished lecture presented at the Tech Policy Lab/University of Washington, Seattle. <https://youtu.be/a2IT7gWBfaE>. Accessed 19 Feb 2019
- Crawford K, Joler V (2018) Anatomy of an AI system: the amazon Echo as an anatomical map of human labor, data and planetary resources. AI Now Institute and Share Lab, (September 7, 2018) <https://anatomyof.ai>
- Crawford K, Paglen T (2019) Excavating AI: the politics of training sets for machine learning. <https://www.excavating.ai/>. Accessed 1 Jun 2020
- D’Ignazio C, Klein LF (2020) *Data feminism*. MIT Press, Cambridge
- Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. <https://doi.org/10.1109/CVPR.2009.5206848>
- Difallah D, Filatova E, Ipeirotis P (2018) Demographics and dynamics of mechanical turk workers. In: Proceedings from proceedings of the eleventh ACM international conference on web search and data mining—WSDM ‘18, New York
- Elish MC, Boyd D (2018) Situating methods in the magic of big data and AI. *Commun Monogr* 85(1):57–80. <https://doi.org/10.1080/03637751.2017.1375130>
- Eubanks V (2018) *Automating inequality: how high-tech tools profile, police, and punish the poor*. St Martin’s Press, NY
- Finn E (2017) *What algorithms want*. MIT Press, Cambridge
- Fraser A (2005) Was ist institutionskritik? *Texte Zur Kunst* 59
- Freire C (2006) *Arte conceitual*. Jorge Zahar Editora, Rio de Janeiro
- Gillespie T (2014) The relevance of algorithms. In: Gillespie T, Boczkowski PJ, Foot KA (eds) *Media technologies: Essays on communication, materiality, and society*. MIT Press, Cambridge
- Gillespie T (2016) Algorithm. In: Peters B (ed) *Digital keywords: a vocabulary of information society and culture*. Princeton University Press, NJ, pp 18–30
- Gillespie T (2018) *Custodians of the internet*. Yale University Press
- Gray ML, Suri S (2019) *Ghost work: how to stop silicon valley from building a new global underclass*. Houghton Mifflin Harcourt
- Google (2019) AI & machine learning products: cloud vision. <https://cloud.google.com/vision/> Accessed 18 Feb 2019
- Haraway D (1988) Situated knowledges: the science question in feminism and the privilege of partial perspective. *Fem Stud* 14(3):575. <https://doi.org/10.2307/3178066>
- ImageNet (2019) Skateboard. <https://imagenet.stanford.edu/synset?wnid=n04225987>. Accessed 7 Mar 2019
- Irani LC, Silberman MS (2013) Turkopticon. In: Proceedings from proceedings of the SIGCHI conference on human factors in computing systems—CHI ‘13, New York
- Li FF (2015) How we’re teaching computers to understand pictures. Retrieved 2020-08-07 from [https://www.ted.com/talks/fei\\_fei\\_li\\_how\\_we\\_re\\_teaching\\_computers\\_to\\_understand\\_pictures?language=en](https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures?language=en)
- Mackenzie A (2015) The production of prediction: What does machine learning want. *Euro J Cult Stud* 18(4–5):429–445. <https://doi.org/10.1177/1367549415577384>
- Mackenzie A (2017) *Machine learners: archaeology of a data practice*. MIT Press, Cambridge
- Mintz A, Silva T, Gobbo B, Pilipets E, Azhar H, Takamitsu H, Omena J, Oliveira T (2019) Interrogating vision APIs. SMART data sprint: beyond visible engagement. <https://smart.inovamedia.org/smart-2019/project-reports/interrogating-vision-apis/>. Accessed 1 Jun 2020
- Noble SU (2018) *Algorithms of oppression: how search engines reinforce racism*. NYU Press, NY
- O’Neil C (2016) *Weapons of math destruction*. Crown Books, Largo
- Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, Mordevintsev A (2018) The building blocks of interpretability. *Distill*. <https://doi.org/10.23915/distill.00010>
- Paglen T (2016) *Invisible Images (Your Pictures Are Looking at You)*. The New Inquiry. <https://thenewinquiry.com/invisible-image-s-your-pictures-are-looking-at-you/>. Accessed 1 Jun 2020
- Parker R, Pollock G (2013) *Old mistresses: women, art, and ideology*. I.B Tauris, London
- Powles J, Nissenbaum H (2018) The seductive diversion of ‘solving’ bias in artificial intelligence. <https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53>. Accessed 1 Jun 2020
- Simanowski R (2016) *Data love: the seduction and betrayal of digital technologies*. Columbia University Press, NY
- Tomkins C (1998) *Duchamp: a biography*. Holt Paperbacks, NY
- Warburg A (2010) *Atlas Mnemosyne*. Akal, Madrid

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Causality, poetics, and grammatology: the role of computation in machine seeing

Iain Emsley<sup>1</sup>

Received: 29 July 2019 / Accepted: 18 August 2020 / Published online: 9 September 2020  
© The Author(s) 2020

## Abstract

Digitised collections and born digital items, such as photos or video, exist beyond the scale of human viewing. New methods are required to read, understand and work with the data, resulting in computation becoming increasingly central to both creation of a cultural reality and as the interpretative tool and practice. If artists' look, then how might a machine see as a critical tool? Developing work on computational culture and the Next Rembrandt project as unstable digital object, this paper considers how the medium affects computational critical practice. Drawing on Heidegger's view of causality and Derrida's grammatology, this paper explores how the medium acts a locus between the human and machine readings and the remediations that occur within the reading. This is developed as through a reading of how the interface translates the signs and symbols and how this affects the reading. By reconsidering the critical assemblage and using it to think with, the human and the machine are seen as critical partners. Attending to the materialities of the reading through a playful approach that decentres potential meaning encourages us to glimpse beneath the surface and gestures towards a critical practice as understanding both computation and its materiality.

**Keywords** Materiality · Causality · Digital humanities · Computation · Interface

## 1 Introduction

Digitised collections and born digital items, such as photos or video, exist beyond the scale of human viewing. New methods are required to read, understand and work with the data, resulting in computation becoming increasingly central to both creation of a cultural reality and as the interpretative tool and practice. If artists look, then how might a machine see as a critical tool? Developing work on computational culture and the unstable digital object, my aim in this paper is to consider how the role of the medium affects computational critical practice. Situated in a critical Digital Humanities (Dobson 2018; Berry and Fagerjord 2016) perspective, this paper considers form and how this affects materiality. Heidegger's (2012) consideration of causality to consider what computational criticism reveals about itself. Considering the computer as the first metamedium (Manovich 2013)

allows us to question the fourth causality, that of the maker, and how this affects cognitive practice through interfaces.

This paper develops a previous consideration of the computational object (Emsley 2019) as a digital pharmakon, an object that is both poison and medicine. This is particularly useful when thinking of the Next Rembrandt (2016) and the superficiality of the image's surface, hiding its constructed nature in plain sight. The project is a machine-generated image based on a reading of digitised Rembrandt portraits from museums and collections and is created by TU Delft, J. Walter Thompson (JWT) Amsterdam, Mauritshuis, ING and Microsoft. The project was intended to raise questions about the power of data for Microsoft's customers and as a playful way of demonstrating ING's support for Dutch cultural institutions, working in tandem with other Dutch institutions. Versions of the image are available digitally on the project's website and it was printed using a 3-dimensional printer. The latter material form echoes painting but is manufactured from ink and a computational object.

Its surface might provide an appealing form, either aesthetically or epistemically, whilst simultaneously hiding issues and ideologies in its structures. Code can embed an argument deep within its structures and through its manner

---

✉ Iain Emsley  
I.Emsley@sussex.ac.uk

<sup>1</sup> University of Sussex, Brighton, UK

of assembly, one that requires software and code studies skills to begin to read and interpret. This assumes that the code or the training data are available to be read. The image output may require a supporting human reading, such as gender and class, and a computational one. Yet the surface is perhaps too perfect and requires readings of its substrate and a material understanding of the medium of the creators and their role. When studying an image using digital methods, we use a machine to see and remediate it from bits to models and grammars. These structures both create, and are created by, the interface effect (Galloway 2012) resulting in a grammarology (Derrida 2016a). I want to think about how the *techne* is created through a series of patterns that generate patterns, moving from an abstract concept to a concrete epistemic object.

By reconsidering the critical assemblage and using it to think with, we see the human and the machine as critical partners. Attending to the materialities of the reading through a playful approach that decentres potential meaning encourages us to glimpse beneath the surface. The reading is considered within a wider economy but to begin recognising those constraints and that theory might be embedded in tools that are deployed as critical practice. In reflective moment, we might consider our own positions as consumers but also as producers and to think about how we can be critical of critical tools and consider that they may have their own positions built into their assemblages. From this, I contend that the condition of the post-digital underpins the analytical process and reading computational culture that considers both human and machines in the process.

### 1.1 Situating the computational in culture

Situating potential readings as both computational and human reveals a potential issue. The tools for machine seeing are embedded in the same social and cultural economy as the object under study. These issues present a case for learning how to engage with and understand how the discourses merge and represent a cultural reality.

A consideration of the way the cultural object is created presents challenges in reading them. The consideration that the work of art is created within a series of codes (Greenblatt 1984) can be useful extended to consider the code itself as both containing and presenting code of culture. In learning how to read them, Anderson and Pold's (2014) considerations of writing being dangerous are echoed as they may contain hidden purposes from their creation. Viewing the Next Rembrandt as a piece of art provokes not only questions about the artistic work but also the purposes in creating it. How might machines engage with Rembrandt's own art to create a new image? Given that this image uses computational material, how might it be read and with what tools? Manovich's conception that computers are the first

metamedium, that the medium can be used to build other mediums and tools, raises a potential issue in the notion in the act of inscribing the data which I will attend to through Derrida's grammarology. It provokes questions to what is being written and how is it being translated whilst making the object.

Having created this epistemic object, we need to reflect on what the new ways of seeing might bring to culture and computational culture more widely. While the object cannot be removed from the conditions that creates the object, we can take a critical view of these condition to understand its effects on culture (Liu 2012). Acknowledging the role of the computation goes to the making the role of criticism less incomplete and to begin reconsidering the way that computation creates a series of objects as a grammarology. Having identified this gap, a change of tack is required to begin a closer reading of the digital object.

## 2 Creating a digital object

The Next Rembrandt and critical approaches to it are both *pharmakon* and designed objects. What I want to consider is this portrait as digital object derived from a distant reading of the portraits and, in particular, the agential workflows that enable it to exist. The painting's very name suggests a new Rembrandt, created 350 years after his death. It may be derived from Rembrandt's work but clearly cannot be by him or his school. Its real creators are not identifiable from the name but requires some reading of the secondary material. The name's provocation suggests that viewers may be wary of the object's superficiality.

The image's existence as a digital object suggests a functional transformation from images collected from different museums to becoming a digitised collection in storage. This representation is enabled through the function of digitisation and the workflows that it enables, suggesting that this data is the final form that emanates from a series of software processes and workflows. Each of these workflows suggest organisational and technological influences and patterns to make the work possible.

This digitisation of a collection of physical artworks into a digital object suggests a notion of art or culture as a potential computational form. A painting becomes a mathematical matrix where the colours and their positions create an alternate materiality. This transposition leads to a loss of information and a forgetfulness (Ellul 1965) as they convert the painting into a numeric form and language. The parts that cannot be digitised are discarded. The images require digitisation so that the final portrait could be created. It can only read the data in a way that the thing to be read supports. A critically meaningful response may come from an intentional, playful challenge but this still requires

an understanding of the material (Berry 2014). To create a reading, one must develop a tool to create the machine representation and one that is written in the medium being read, so requiring us to consider how and what is being written and read with a machine.

The final image was created using facial recreation algorithms to determine the patterns that bring the image together. The data was derived from the digital reading using facial recognition algorithms. Using a model, the cultural images are converted into a digital format as paint becomes a pixel, representing a colour using a standard and a position. The mapping of a human epistemic object to a technical epistemic object requires a translation. Not only is one conceptual pattern being translated from a human concern, but the form is also translated into a technical one; one that can see within the generated numeric symbols. A human reading of the eyes in a portrait may be computationally read as a measurement of the varying colours in the pixels of an eye, its circumference and the distance between the two eye features. Within the context of the training data given to it, the reading may understand that the eyes are typically placed either side of the nose feature and has a mathematical relationship to other salient features. The salient features might be derived from the same training set or assumptions made. The computational process views the picture through technical relationships that are reliant on the underlying digital object. The Next Rembrandt team used 6000 points to classify the features (Dutch Digital Design 2018) to create a typical Rembrandt face from the data, having analysed specific features such as ears, nose and mouth. In part, this moves from the concrete form of the face to an abstract mathematical model of one. Converted into a model, the idea of face can be read at scale.

The abstraction allows the machine to create the pattern of the face. At this moment, the use of these features is a form of inference. The facial creation draws on the patterns identified facial recognition to guess at how the face might look before the proportions, also machine calculated, were applied. This concrete pattern for the machine is used to create an abstract concept of a face that is represented as a picture. In the act of the computational becoming a maker, the created symbols become signifiers and are interpreted and re-presented. The face exists as a numeric model in a space made by the pixels but it only becomes viewable once the pixels are converted from numbers to colours in a position. The viewable image is represented through a file which can be printed out digitally or as a 3-dimensional printed image. Rather than leave the image in 2-dimensions, a height model was created to simulate the paint layers for the printed object. The act of printing raises questions about the remediation of the digital image into a painted one and the embedded interpretations that might be interpreted by viewers. This version of the digital object is not one that

enables a potential feedback loop into the underlying data but imposes itself. The paint simulation echoes the original medium but cannot be of it.

Our ability to read the pattern is affected by the use of a human set of experts who helped determine the final image and the group who guided the final form. A human team determined that the image would be a male within demographics. Whilst the machine is able to determine light and shape as patterns and associate these with features, even testing them by writing unexpected wrinkles into the eyes, it is reading the computational object. While it has its uses, the pattern has limitations and are challenging to learn to read. One way of reading patterns is that they identify ways of writing software or interaction design as a learned way of working with the machine. As machines learn from an abstract pattern, it then perhaps considers its own pattern to create and write. These patterns then create a new sense of interaction but one at a cognitive practice. By recognising the difference between the makers behind the provided responses, we might begin to see that the sense of self in this work is made by both humans and machines. Unpicking this process provides a space for reflection and interpretation to engage with the object.

## 2.1 Reading with the machine

Instead of being fixed in the manner of dried paint or a printed text, data points and models can be manipulated and queried. This form of data enables the grammatology of the digital model through its existence, but it also erases itself and its construction in the representation. We might begin to reveal this form through using software to read the image in and then translate it between formats, to query it or to write in errors and glitch it. These processes use patterns of concepts to approach the forgotten materiality though, in some cases, the operation may be quite easy. An image might have a simple file conversion from a JPG to a PNG format than can be achieved either through a script or a desktop programme, such as Preview on an Apple machine. These conversions abstract the protocol and material transformation models from the user to change the form. The very ability to be able to alter the underlying structure with little apparent changes to the final form suggests that there is a remediation of these forms to show the image. Although these changes hint at the existence of an object, it still does not allow it to be read. It suggests a superficial model that is already hiding the software processes that brought it into existence. The image is partially revealed as a shell. Before turning to its creation, I briefly consider critical approaches through glitch studies.

Glitching (Menkman 2011) can be seen as a critical material approach to deconstructing form. If the glitch is deliberate, rather than exploiting an unforeseen situation, then it

requires an understanding of the potential for the symbol of the interface and might be seen as a pattern. In its playfulness, the application of the glitch reveals the remediating model though it contains its own understandings of the symbols and what they might signify. It is a critical reading of the computational surface that deconstructs the model and its remediations.

The Rembrandt image relies on a close reading of 346 paintings to create the units of distance as a ‘form of knowledge’ (Moretti 2007). This scale of reading may be beyond the human ability to read these in a scholarly fashion. Where this project created one final image as their output, Manovich (2012) cultural analytics project on Manga scanlation pages points to different ways of reading. Using scans of Manga pages, tagged by a community, the project used digital image processing to create a computational reading. Using the numeric forms of the colours to extract features based on them, they were able to take a quantitative approach to the pages. The resulting data is shown on a 2-dimensional graph, such as a scatterplot or image plot, to show the new form of relationship. Where Next Rembrandt focuses on human features to extract through computational patterns, the Manga project uses the computational data to explore relationships between pages using the numerical representation of colour and how the machine creates it in the file format.

Using visualisation as a descriptive system (Manovich 2012), the project uses the medium as metamedium to communicate what the critical tool is discovering. Computation is being used as a tool to translate a question into a pattern search and then to build its own response from the same medium. Although visualisation is not limited to machines, these of it here makes the computational reading sensible to humans. The type of visualisation affects this through its layout model and how the data points are constructed. Where Next Rembrandt falls more into an art historical recreation of cultural objects, the Manga project presents an analysis of the project’s underlying data. Yet without knowledge of the underlying data and processes, it is hard to read the scatterplot and appreciate the arguments involved in it. One has to learn the language of graphs and visualisation to begin questioning and reading them.

Echoing the relationship between recognition and creation, the question about the maker—machine or human—is raised. The computational tools work with existing media, such as painting and print, once they are digitised. Menkman’s glitching is another expression of this where the computational materiality enables tools to work with it and to alter it, but it is still the artist that creates the glitch. Within the other two projects, the computational tool is more apparent as a joint maker with humans. The assemblage becomes the site of theorising, either through testing the materiality or using it to read data. Using the patterns within the computational object in different ways to an intended purpose,

the tools write an epistemic object that it is created as a visualisation to make it sensible for a human. The machine enables the reading to take place through the application of a mathematical model as the data is not really human readable and it can provide a reading. This must then be interpreted to consider what and how the response is created. It points towards a material issue with computational techniques where the reading is constructed from the same medium as the subject under test. It also demands a computer literacy to either code or to understand the response. As the computational object is approached, it hides itself behind the projected form as an entity derived from the form that is requested either through the purposes, such as a projection of economic or technical power through culture, or the critical question posed about it.

## 2.2 Patterns as forms

Before considering the issue of reading and writing, a consideration of patterns as forms that both mould and are moulded from the materiality of the digital is useful. I want to suggest here that the pattern is a way of thinking as well as affecting the material, demanding a consideration of the materiality of the pharmakon to begin to read it. The word pattern needs a consideration here, as both a thing that is found as well as something that is desired to be found. I think that it is useful to consider these in the light of the mix of functionality and representation that Anderson and Pold (2014) argue for but to consider it in the light of using machine driven tools. Patterns can be seen as a generative act (Derrida 2016a) or, in the sense of dark patterns, to create a misreading (Dieter 2015). Using a pattern, such as facial recognition, a machine learning algorithm uses a model against the data structures. In many ways, the machine can read the patterns more closely than a human to a generate a new model, or a sign. Knowing of these models and how they are constructed sets up a series of links that tie in the sets of models. These links might be considered as ways of generating meaning and as a way to divert attention from questions of purpose. By trying to identify the patterns used, we can begin to approach the concept, the form, of the thing that is to be made. As the abstract concept of the form, such as a theory or pattern, combined with materiality towards the presented object. A consequence of the transposition through models is the creation of a digital double of the original object.

In digitisation, the location of memory and understanding is moved from the human to the computational as the reading is only made possible through the translation of the cultural into a technical form. Rembrandt’s portraits had to be scanned and digitised for the algorithms to complete their viewings, yet the digital is not aware of the portrait’s life before the scanning process. It is only when paint is

transposed into pixels that their computational use is made possible and that a new critical reading is required to understand the technical and social means of production of the final form. As the material is fundamentally changed, the digital must make a copy of the image rather than translate it. A successful reading of the doubling understands this materiality and the languages and logics that are required to interact with it (Stiegler 2019). These logics are not only computational to understand the new material form but also cultural to understand the remediated object.

When a new version of the reading is begun, existing results and data may be irrecoverably lost. It becomes a “redoubling” (2019), where Stiegler points to the forgetfulness or lack of history in the reading. Unless a pattern is used to act as a form of storage, either of the results or the provided parametric model, the reading loses its own history and context. This redoubling continues the move to the technical through allowing it to create its own present and historical realities. The tradition of comparing readings of the same object or artist requires access to these readings, not only in the final paper form but to also understand the forms used in the reading. Although this points to the open knowledge and data discourses, this paper’s concern is the way that the post-digital reading creates a grammarology within the digital object. Each reading becomes a micro epochal moment, an interruption that becomes a new moment in time, through the presentation of itself but forces a recognition of this use of time as model. Such forgetful re-presentations may alter the perceived grammarology and patterns used in the reading.

Critical computational methods are created to understand this computational double. Having explored the surface understanding of the translations between human and machine form to create the conditions for a reading, I want to turn inwards to the computational to consider the next challenge. The form that is being considered is remediated through the interface, a meeting point of two components, where the individual parts translate the model or data given to them. Processed in order, the concept is continually remediated through a series of models, either as data structures or implemented algorithms, to create a pattern matching machine. These patterns can either read a pattern as it is or statistically infer it.

As such, we might read the Next Rembrandt as a series of statistical inferences based on existing historical data that reveals the hidden computational judgements. The inferences are judgements based on existing knowledge and readings. Unlike a human critic who may raise some of these in the final output or leave enough traces for the reader to infer the judgement, the machine uses a logic to make an inference based on a set of models. We may raise a different consideration raises a different question: who is the maker, the entity that causes this to come into being? In Heidegger’s

example, the chalice has one maker, the silversmith, who crafts the metal into the chalice (2012). The digital object may have at least two makers: the computational assemblage that receives and creates the object and a person, or team, who design the assemblage. The designer or coder creates the possibilities for the computational when the workflow is designed and built. This object may then be remade to read it critically.

The person who uses the interface, either through code, graphics or haptics, creates a conception of the desired form through the provided options. These may be through search terms, filters of percentages (such as network closeness) or a representational model such as a sonic or visual way of mapping the data. The computational assemblage reads these through interfaces that translate the concept into a technical form, which is then remade and mapped to the underlying data structures to create the computational object to be read. The human is an original maker who creates the initial form to be found and either provides the options within an interface or constructs a machine to process the form. The constructed machine accepts the input and creates the object according to the logics given to it by the builder and the data. Humans and machines become cognitive extensions of each other. The underlying machine processes are required to operationalise the theoretical concept and concretise the model as error or representation, depending on its fit to the interface. A cultural object is rendered artificial for the technical process. Simultaneously, the machine requires a model to begin the process of concretization from an abstract concept of a system, a scientific image (Sellars 1962) or representation (Simondon 2017), and the cultural one that might be considered as a natural object. The artificial object that is accepted is re-presented as a natural object. The materiality that is shown to be interpreted is derived from the concretisation that might not reach the abstract image that either party has in mind. It is the result of an imperfect set of transpositions and translations but it is embedded in the creation of culture.

The critical reading provided by the critical user as form becomes a negotiation with the underlying data. Once the transposed double is made, it is read by machines to be presented as seeing through a series of patterns that standardise the natural, cultural object behind the surface. As the cultural computational form is queried, it may be altered into desired response. It is a made object and there are makers that become revealed when the causality is questioned, revealing the technology involved. Having considered the forms and makers who create it, the materiality’s form can be considered as being created and altered through grammarology.

The digital object is created from the reading and writing processes that convert the original computational form from digitisation into the relevant models. An effect of the



concretisation is the transposition and translation of concepts through a series of languages. This points to the medium as a site of transcoding.

The change of medium enables the processing of the data, from the finding of patterns to reformatting and presentation, but demands a change in the way that it is written. The machine reading, hiding itself as a model, uses interfaces to mediate the translation of the cultural data between the software components. The construction of the interface defines what signifiers it can accept and so what symbols that will be created. Any deviation from these definitions suggest an (un)intended break. I use this to point to the materiality of the symbols and the role of interfaces in computational reading and writing to create the digital object. The interfaces abstract the role of the digital in interpreting the given signs through the divergent ways that they are written as code and as computer language, which itself maybe general purpose or a domain specific. The code written to determine aspects of the data, such as dominant colours or features, uses a numeric way of creating the pattern that is defined by what the underlying language allows through its own grammar. Read these against computational forms before presenting them as a model, either as a table or a colour, the pattern requires translations of the underlying data to be able to access the code to read it.

The surface image of the Next Rembrandt is derived from a remediated pattern used in a file format that is created by a reading. This format is derived from algorithmic patterns using digitised data through a series of patterns. Reading and considering these patterns relies on a different pattern language: networking. The series of services, switches and routers to allow the computation and data to meet is hidden but should be understood. Given the example that has been used of the Next Rembrandt, this network might be considered through the physical network of museums and galleries who has allowed the pictures to be digitised. The physical form can be read as an artistic work but it hides the nature of computation required in its surface. Using a series of measurements, the painting has a series of raised areas where the image may be thicker, either by choice, such as overpainting, or accident, such as a drop. In this act, the concrete behaviour of the artist might be read as the algorithmic imagination of a study of paint surfaces.

Building on the concept of the location of reading and writing, the symbols are written and developed through human and digital interactions. Having considered this through the medium, it is useful to think of this through deconstruction. There is a gap between the signifier and the symbol that creates the meaning but can also through the different types of creation machine that creates a sense of the pharmakon. The realisation of the poison and remedy is revealed in this reading. The image that is constructed is a visual representation that might use colour, which

potentially has a significance that can be read into it. Made from either a pixel or collection of pixels as a numeric entity that was derived from an internal process, the material form of a colour is altered and represented as numeric symbol on the screen but the human symbolic reading is removed. The human reading is reliant on the machine reading and the symbols that it creates and has to be recreated by the viewer. The symbol's context might be altered through a change in the underlying computational model as the computation returns the form that it has gained from its reading of the data.

A deconstructionist reading here that can reveal the processes where patterns are read and written. This has the consequence of using the underlying computational models as texts to be read and interpreted. As part of *différance*, Derrida (2016b) uses the concept of play as a deliberate destabilising of meaning can be used to consider the way that meaning is being constructed. Engaging with the languages and structures to understand how they are written and read through iteracy (Berry 2014) to view how the data becomes an image or a graph. I see the act of interrogating the symbols and forms through the interface objects as part of computational thinking to understand how the code reads and writes. As such, it becomes possible to consider how the original images are constituted as a new cultural object using the material nature of the computational object. Rather than seeing culture as material object or text that can be rewritten, we might approach the role of the medium and *techné* in the process of a creating a cultural and epistemic object.

### 3 Conclusion

Computational culture is approached in this paper through questions of causality and grammatology to engage with the materiality. A computational image, created from a reading of transposed images, is used as a way of understanding the nature of the computational to make the image through patterns and models as well as being constructed as a model to find models. I contend that machines seeing can be used to augment human seeing as a critical act but that it requires a critical reading not only of the material but the tool that creates it.

In the act of seeing culture through a computational lens, we need to understand how the medium affects the material. I build on the idea of a post-digital criticism through considering the patterns used in tools and how this affects the grammatology. Although we might not see the underlying computational object, approaching it using critical tools and destabilising the presented form provides critical insights into how the cultural text is being actively created. Yet we need to also contextualise the object within the wider role of not just the techniques but also from the entities who have

caused it to be constructed to consider why the tools have seen the way that they might. Questions of causality and grammatology enable the reader to understand the techne that underpins the cultural image as epistemic object.

**Author contributions** The author is the sole author.

**Funding** The work is self-funded.

## Compliance with ethical standards

**Conflict of interest** The author is not aware of any conflict of interest.

**Availability of data** This paper is a reading of the Next Rembrandt project (<https://www.nextrembrandt.com/>).

**Code availability** This paper does not use code.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Anderson CU, Pold S (2014) Manifesto for a post-digital interface criticism. <https://mediacommons.org/tne/pieces/manifesto-post-digital-interface-criticism>. Accessed 10 Jan 2020
- Berry DM (2014) *critical theory and the computational*. Palgrave Macmillan, London
- Berry DM, Fagerjord A (2016) *Digital humanities*. Polity Books, Cambridge
- Derrida J, Spivak GC (trans.) (2016a) *Of grammatology*. John Hopkins University Press, New York
- Derrida J, Johnson B (trans.) (2016b) *Dissemination*. Bloomsbury, London
- Dieter M (2015) *Dark patterns: interface design, augmentation and crisis*. In: Berry DM, Dieter M (eds) *Postdigital aesthetics: art, computation and design*. Macmillan, London
- Dobson J (2018) *Critical digital humanities*. University of Illinois Press, Chicago
- Dutch Digital Design (2018) *The next Rembrandt: bringing the old master back to life. Case study: behind the scenes of digital design*. <https://medium.com/@DutchDigital/the-next-rembrandt-bringing-the-old-master-back-to-life-35dfb1653597>. Accessed 10 Jan 2020
- Ellul J, Wilkinson J (trans.) (1965) *The technological society*. Jonathan Cape, London
- Galloway A (2012) *The interface effect*. Polity Books, Cambridge
- Greenblatt S (1984) *renaissance self-fashioning: from more to Shakespeare*. University of Chicago Press, Chicago and London
- Liu A (2012) *Where is cultural criticism in the digital humanities?* In: Gold M (ed) *Debates in the Digital Humanities*. University of Minnesota Press, Minneapolis
- Manovich L (2012) *How to compare one million images?* In: Berry DM (ed) *Understanding digital humanities*. Palgrave Macmillan, London
- Manovich L (2013) *Software takes command*. Bloomsbury, London
- Menkman R (2011) *The Glitch Moment (um)*. Network Notebooks 04, Institute of Network Cultures, Amsterdam
- Moretti F (2007) *Graphs, maps, trees: abstract literary models for literary history*. Verso, London
- Sellars W (1962) *Philosophy and the Scientific Image of Man*. In: Colodny R (ed) *Frontiers of Science and Philosophy*. University of Pittsburgh Press, Pittsburgh
- Simondon G, Malaspina C, Rogo J (trans.) (2017) *On the mode of existence of technical objects*. University of Minnesota Press, Minneapolis
- The Next Rembrandt (2016) <https://www.nextrembrandt.com/>. Accessed 10 Jan 2020

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Seeing like an algorithm: operative images and emergent subjects

Rebecca Uliasz<sup>1</sup>

Received: 29 July 2019 / Accepted: 18 August 2020 / Published online: 16 September 2020  
© Springer-Verlag London Ltd., part of Springer Nature 2020

## Abstract

Algorithmic vision, the computational process of making meaning from digital images or visual information, has changed the relationship between the image and the human subject. In this paper, I explicate on the role of algorithmic vision as a technique of algorithmic governance, the organization of a population by algorithmic means. With its roots in the United States post-war cybernetic sciences, the ontological status of the computational image undergoes a shift, giving way to the hegemonic use of automated facial recognition technologies towards predatory policing and profiling practices. By way of example, I argue that algorithmic vision reconfigures the philosophical links between vision, image, and truth, paradigmatically changing the way a human subject is represented through imagistic data. With algorithmic vision, the relationship between subject and representation challenges the humanistic discourse around images, calling for a critical displacement of the human subject from the center of an analysis of how computational images make meaning. I will explore the relationship between the *operative image*, the image that acts but is not seen by human eyes, and what Louise Amoore calls an “emergent subject,” a subject that is made visible through algorithmic techniques (2013). Algorithmic vision reveals subjects to power in a mode that requires a new approach towards analyzing the entanglement and invisibilization of the human in automated decision-making systems.

**Keywords** Algorithmic vision · Big data · Operational image · Machine learning · Algorithms

*Portraiture*, the act of making a portrait of one’s self or someone else, has a multifaceted meaning in an era of mass social media amidst a swell of digital images that ubiquitously flood our everyday sensorium. A portrait, especially a self-portrait, a representation of the self, has relations to the term “profile” in its colloquial use. I take a photo of my face, I post it to my Instagram story. I am adding to my profile, myself as data, an outline of myself sketched through a series of loosely connected points. My portrait dissolves into an abstract contour composed from the lines between likes, clicks, status updates, my depiction in datafied form. In contemporary discourse, computation has retooled the digital image, with the repercussion that images are not representational, but have an active capacity to perceive and produce new information through 21st-century data analytic practices.

## 1 Does one have a right to an image of one’s face?

I kept coming back to this question last spring when I learned about the controversy surrounding a data set created by the Duke University Computer Science. DukeMTMC (Multi-Target, Multi-Camera), a large-scale database of images of students and faculty captured by Duke researchers using a university-sanctioned campus surveillance camera system in 2014, was previously one of the largest and most frequently utilized data sets for video re-identification, a process that uses computer vision and artificial intelligence to identify and track targeted individuals within live video feeds (Ristani et al. 2016).

Duke researchers claim that the project, funded by the US Army Research Office and National Science Foundation, was originally intended to improve systems for motion detection of objects in video, regardless of “whether [the objects] are people, cars, fish or other” (Ristani et al. 2016), and was subsequently made available for download on the Duke Computer Vision research website (Saticky 2019). The data set was recently taken down after it came under fire in

✉ Rebecca Uliasz  
rebecca.uliasz@duke.edu

<sup>1</sup> Computational Media, Arts & Cultures, Duke University, Durham, NC, USA

April 2019 for ethical and privacy violations following an exposé by researcher Adam Harvey that prompted Duke's Institutional Review Board to revisit the terms of the collection and use of the image data (Harvey and LaPlace 2019). Harvey's research reveals that the data set has been irreversibly implemented in computer vision, body tracking, and facial recognition systems by academic, governmental, and military institutions across the globe. Significantly, the data set has been traced to research papers published by Chinese AI SaaS companies associated with the surveillance techniques used by the Chinese military to target and monitor the activities of Uyghur populations in remote northwest China (Buckley and Mozur 2019; Harvey and LaPlace 2019). As Harvey points out, this implementation is aligned with the original motivation of the Duke researchers, who published a subsequent paper in 2016 titled "Tracking Social Groups Within and Across Cameras" (Solera et al. 2016) and yet the context makes all the difference.

While this data set is no longer available through official Duke affiliated websites, by the time it was taken down it had already been cloned into many databases around the world. When I learned of this episode, I easily recovered an edited version from a publicly linked cloud drive. Scrolling through the data set, which contains images of students entering and leaving academic buildings and places of worship, one might see how this information could easily be weaponized to target and discriminate against marked individuals. In Harvey's analysis, this is but one example of "an egregious prioritization of surveillance technologies over individual rights" (2019). As I sifted through thousands of blurry images, hardly discernable faces aside from a pair of glasses here, a baseball cap there, I wondered *whose* individual rights were actually at stake? To what extent is the specificity of the images contained within the DukeMTMC data set relevant to their use?

To trace the relation between DukeMTMC's images and its implementation on the other side of the world begets an analysis of the algorithmic techniques used to process this imagistic information. In rural China, Xi Jinping's "people's war on terror" implements algorithmic profiling to doubly correlate religious extremism with all expressions of Uyghur Islam, and all Uyghur Muslims with quantifiable biometric characteristics (Sound Vision Foundation 2019). Through invasive surveillance practices like the Integrated Joint Operations Platform (IJOP) which monitors Wi-Fi, CCTV cameras, and government IDs, governmental officials have both detained targeted individuals in "re-education" centers and submitted them to further biometric profiling to form a comprehensive data print of the individual (Buckley and Mozur 2019; Human Rights Watch 2019; Cooper 2020). This data is then fed back into their system to train machine learning algorithms that are designed to target Uyghurs' physiognomy. Drawing together information that ranges

from "the color of a person's car, to their height down to the precise centimeter" to activities that would otherwise be considered lawful, like "not socializing with neighbors", officials use the IJOP to flag individuals deemed high risk to prompt further investigation (Human Rights Watch 2019: 2).

In the name of preempting "unsafe actors," the Chinese government has adopted a strategy enforced through techniques of risk—in other words, the Chinese government defines the category of risk, and feeds a system data until it is trained to automate decision making, crystalizing an ad hoc type of governmental intelligence. President Xi's "stability maintenance" initiative exemplifies an automated risk calculus that takes the entire Uyghur population as a testbed for techniques of data extraction (Byler 2019). The logic of pattern recognition is found in a 2016 researcher police report on the operations of IJOP:

*if a person usually only buys 5 kilos of chemical fertilizers, but suddenly [the amount] increased to 15 kilos, then we would send the frontline officers... to check its use. If there is not a problem [they would] input that into the system and lower the alert level (Ningning 2016, emphasis mine).*

Where there is incomplete information on each citizen, the IJOP performs a continuous disaggregation and reaggregation of noncausal data, constantly shifting the relations until a pattern deemed significant emerges. Here, algorithms fracture and fragment a data set, to the extent where it has very little to do with its underlying referent. It is recombined and projected forward to the effect that a new visualization is made to emerge—a diagram of the lines that cut across the data set rather than what is contained within.

The Chinese government calls upon knowledge provided by data analytic and security companies. These companies then can make use of infinite troves of data to form infinite correlations between samples, generating an essentially infinite number of proprietary algorithms to sell to government and corporate contractors. Among these government-contracted companies are Chinese AI SaaS SenseTime and SenseNets, both of which Harvey linked to the DukeMTMC case (Murgia 2019; Harvey and LaPlace 2019). Researchers from both companies published a paper that proposes a method of training a neural network that uses DukeMTMC to perform feature composition analysis on pedestrian images taken in crowded or noisy public spaces to overcome problems of occlusion—it fills in what cannot be seen by the machine eye (Xu et al. 2018). In a sense, the circulation of DukeMTMC has enabled the expression of state power that acts through *something that literally does not exist in the collected data set*, it fills in a gap. An unknown risk yields to the constitution of a data subject.

The paradox around contemporary demands for rights to one's data and ethical regulations around future use of

personal information becomes clear when we realize that, regarding the DukeMTMC case, we don't *see* ourselves reflected in the algorithmic system on a personal level. It does not index us as individuals. In other words, our images break from representation in the sense that they no longer mean what they appear to mean, and sometimes don't even appear at all. In the following, I argue that algorithmic vision appears in tandem with a paradigmatic shift in the status of the technological image itself. Stemming from post-war United States cybernetic sciences that led to developments in computation, algorithms operationalize images for use in systems that appear to function automatically without a "human in the loop." The presumed disappearance of the human, I suggest, is a product of a technoscientific discourse that produces a veneer of objectivity. In contrast to the full automation of human thought and eradication of human error, I argue that human experience is fundamental to the operations of digital images in systems of algorithmic governance that weaponize novel techniques of subjectification. This provocation leads to the following—the techniques of algorithmic governance today beget a critical analysis that brackets simple techno-determinism in favor of understanding how machinic systems constantly reconfigure the human subject in question, suggesting the entangled, malleable and nonhuman essence of subjectivity.

## 2 From image to data: a non-representational paradigm

For Roland Barthes, to have one's portrait taken is to experience a micro death, to feel oneself slide from subject to object position. When we pose for a portrait, we enter a "closed off field of forces," a closing in of self that at once over codes us in a field of representations and affirms us to a future viewer. When we look at a photo of a face, we witness the conquest of a subject by the apparatus of photography, an obfuscation of self undertaken to represent oneself to the world (Barthes 1981).

The conquest of the world as a picture is articulated by Vilém Flusser in his *Towards A Philosophy of Photography* (2000). The photographic universe, he tells us, is characterized by an inversion in the Cartesian model—concepts no longer signify the universe, but rather the universe is programmed to signify concepts. The camera, the first apparatus for changing symbolic meanings in the world, "knows everything and is able to do everything in a universe that was already programmed in advance for this knowledge and ability" (2000: 68). The photograph has a future effect, it casts a magic spell. It has the ever-developing ability to program culture in its image through constraining possibility and habituating action.

Enter what Flusser calls the technical image, that illusory abstraction that appears to be a direct index of reality but is really a meta abstraction of text, which is itself already an abstraction of a traditional image. Although they appear to touch the real, technical images are 'encoded' with concepts of the world. They project those concepts back out onto the world, inhibiting humans from their capacity to understand reality. Here, we are met with the crisis of representation, an overabundance of information, and a multiplication of signification. Our images and texts alike become encoded with concepts that are nested far too deep for human interpretation. Characteristic of the technical image is its ability to create information, to create *new* meaning out of thin air (or rather, un-photographed, un-informed nature). This raises the ur-question of Flusser's explication, and a paradox that maps onto our contemporary computational regime—if there is no deeper meaning underlying cultural existence, how do we separate noise from information without slipping into mass paranoia?

Although Barthes tells us the slippage of the photograph occurs in the failure to distinguish between the thing and its representation, how does the digital image complicate the status of the thing? The slippery ontology of the digital facial image warrants further contemplation in a moment, where digital images have gained new significance in the operations of global capitalism. Paradoxically, as images of faces are amassed in digital form, the face recedes from our sight. As we are flooded with news of global terror and domestic injustices validated by surveillance technologies and predatory commercial practices grounded in data analytics, we ironically lose touch with the ability to see the images that are said to be underneath such surface effects. Images become invisible with their absorption into data sets—the universe of raw matter through which algorithms touch reality. It is in this vanishing act that digital images function as one component of *algorithmic vision*.

To the eye, algorithmic vision appears to operate on an ontological plane that breaks from human experience, undergirded by a regime of digital images that do not represent an object or subject through the modes that images have been traditionally analyzed in humanistic discourse. To this extent, one popular critical response has been to understand machine learning systems technically through theoretically opening the "black box," a term used to describe their opaque and unknown latent operations (Seaver 2017; Paglen and Crawford 2019). Alternatively, however, if we are to understand algorithmic vision as a process that not only abstracts and simulates, but also imbricates human cognitive processes, we might arrive at a different configuration of the black box, and some different conclusions around what we might do with it.

For Flusser, the black box of photography didn't so much cast aside the human as it crystalized a thinking process,

concretizing and ultimately reifying a way of producing information—the camera is an apparatus that “is invented to simulate specific thought processes” for the sake of extending the human capacity to generate new information thought work (2000: 31). An apparatus is a tool unique to a postindustrial world presumed to be informatically dense. Here, the automation of labor undergirds a never-ending combinatorial game. In this world-game, shifting symbolic patterns constantly recast relations between humans and machines. To map an alternative relationality of human subjects and computation, then, is to emphasize that automation is historically bound to the division and reassemblage of human and machinic components in the postindustrial workplace that undergird the economy of digital images to come.

Automation retools the digital image such that it is not its ability to *represent* subjects, but rather, its ability to *produce* subjects that becomes a core feature of its amenability to modes of algorithmic governance. Computational imaging practices including techniques for storing, processing, and manipulating discrete imagistic information in ways that preempt human sensation are a quintessential component to the machine learning systems that are increasingly ubiquitous in both political and commercial domains (Apprigh et al. 2018; MacKenzie and Munster 2019). Analyzing computational imaging itself is immanent in analyzing the operations of algorithmic vision technologies underlying the subjective experience of algorithmic culture. Here, I suggest that *pattern-finding* is a core technique of computational imaging that structures the participation of images in algorithmic vision systems. Algorithmic vision systems classify and produce images through logics of pattern-finding that beget critical media theoretical analysis. Through patterning information, algorithmic vision not only simulates a codified notion of human cognition, but actively participates in the articulation of human subjectivity.

### 3 How does an algorithm see? From data to image

Deep learning, as described by Ian Goodfellow, is an advanced form of machine learning that uses a layered neural network to perceptualize patterns to train a classification algorithm (Goodfellow et al. 2016). Deep learning techniques, writes Goodfellow, are intended to simulate the intuitive reasoning methods that humans use to make decisions in the real world, where complex variables and incomplete information negate the possibility of relying on a pre-codified means for arriving at complex decisions.

Explains Goodfellow, neural network-based artificial intelligence technologies are developed to overcome the difficulty of computationally modeling human decision making when faced with an enormous amount of information. That

is to say, humans have subjective and intuitive knowledge of the world, and a capacity to learn from previous experience, a quality not inherent to computation. Thus, the artificial neural network was developed out of a desire to imbue computational systems the “ability to acquire their own knowledge by extracting patterns from raw data. This capability is known as machine learning.” (2016: 3) The problem of granting a computer the capacity to make meaning from a rich and densely subjective reality is solved through breaking representations down into simpler representations, ad infinitum.

To give a concrete example, contemporary object recognition algorithms often make use of what is called a convolutional neural net (CNN). In reductive terms, a CNN is trained to recognize certain objects processes a digital image by extracting low-level features based on pixel values. As features (like edges, corners, and thresholds) are fed forward, higher-level features are identified at each level until the algorithm deems the output probability high enough that it can safely determine it has recognized an object.

Important to the design of CNNs is the use of convolutional layers called tensors that are used to weight input data according to a specific array of parameters. This array is referred to as the *bias*. Other layers of the neural net perform an operation called pooling, which downsamples information to reduce dimension, retaining only the features deemed most important to the algorithm. In sum, input data is initialized with random weights and processed through a set number of convolutional and pooling layers and the output is compared to the expected result to calculate the error. This calculated error is then propagated backward into the hidden layer(s) to tune the learning algorithm, the process repeated a specified number of times or until the desired error rate is achieved. Crucial to this operation is the introduction of randomness—the multiplication of data by any number of possible states—to infer patterns within information. In a way, CNN algorithms act as a playground for programmers to tweak the variables of weighted layers until the economically desired outcome is derived (Amoore 2019).

Emergence of a form proceeds through patterning across scales. In effect, patterns are fed forward to the next level of the neuron, a layering of micro-decisions that generate new propositional structures through compressing random pixel data. Patterns drawn from noise are effectively hallucinated into something akin to a form.

This operation requires a few key components to function in tandem. Apart from a learning algorithm that must be trained up until a computational or human supervisor decides it is performing optimally, and a model application that allows it to be applied to different classification tasks, it needs a vast set of *training data* that is used to “teach” the intelligent application how to extract different associations as patterns. In other words, the data is a key component in

structuring what constitutes “meaning” for the model application. Significantly, this entire operation is undergirded by many institutional databases made possible by mass digitization efforts post World War II, stemming from academic, government, and military-funded initiatives.<sup>1</sup> That is, the mass digitization of image information was a preconditional necessity for the operation of algorithmic vision technologies today (Franklin 2015).

The humanistic idea of a semantically structured and reasonable mind is all but eliminated in the CNN, which fuses biologically metaphors of neural nets with statistically derived programs bolstered by massive troves of training data. Recent scholarship on the relevance of United States post-war cybernetic sciences has contributed to an understanding of the emergence of new ideas around consciousness and cognition during this time, and the epistemic transformations this legacy wrought across science, communication, and design (Halpern 2015; Franklin 2015; Mirowski 2002). Stated reductively, in an attempt to apply systems thinking in the name of war, cybernetic discourse displaced the question of how to describe knowledge, truth, and consciousness in favor of the question of how a structure might be operationalized. The questions of Enlightenment reason (what *is* a mind?) are displaced by a methodology of rule-based rationality (what can a mind *do* in its abstract form?).

While describing the variegated technological and conceptual experiments across the biological and social sciences during the post-war moment is beyond the scope of this paper, this period significantly recast questions around vision and perception in the light of technological automation. Importantly, early cybernetic experiments conducted by Warren McCulloch’s Research Laboratory of Electronics at MIT concluded that the perceptual circuits in the eye of a frog emit impulses to the brain in response to certain patterns of visual stimuli—variations and edges between light and dark—suggesting that visual perception itself is a type of pre-conscious cognition (Lettvin et al. 1959; Halpern 2015). In broad and generalizing sweeps, this experiment was taken by others to suggest that perception might be disembodied, modeled, and therefore, automated.

Further experiments around depth perception and form recognition by Hungarian engineer Béla Julesz at Bell Labs in 1959 aimed to locate the physical location of object detection within the human visual apparatus using a stereogram—a random-dot image created from characters produced on a microfilm plotter on a 1024 × 1024-pixel grid (Patterson

2015). Julesz concluded that depth perception in random noise could be measured and reproduced through patterns. The perception of patterns became a way to discover meaning. Cognition became about the operation of scanning a vast amount of information in a quest for meaningful correlations, where what is “meaningful” is programmatically defined. Perception in Julesz’s images is a form of thinking abstracted from a human subject and prior to linguistic signification. This type of machinic visuality does not depend on a stable or individuated subject, but rather posits that an observer might be attuned through patterns. The human subject is destabilized in pattern-finding techniques in favor of finding the “essence” of a data set, the methodology of the algorithmic vision of today. The ability to know an object in the world becomes a function of the statistical distribution of pixels on a screen, foreshadowing data analytic practices to come.

If cybernetic science charts an attempt to render visible and knowable the virtual space of data through human perception of patterns in noise, the artificial intelligence algorithms of today are more complex—they perceive patterns that surpass the limits of human perception (see Parisi 2013; Hansen 2014). CNNs show us that computational imaging fundamentally alters the concept of the image as one that represents an object to one that creates an object. Cybernetic science laid the grounds for a field of image processing that substitutes cognition with the production of information hallucinated from patterns. Cybernetic science shifts the notion of the “image” from one that has a primary representational to primarily operational function. The image of contemporary computation is, in essence, the function of a temporal and dynamic relationship between data points, giving rise to what has been called the *algorithmic* or *operative* image (Farocki 2004; Hoel 2018; Hoelzl and Marie 2015:100).

Noting the disjuncture between human and machinic perception as it played out in the development of automated warhead technologies, filmmaker Harun Farocki describes “operative images” as “images that do not represent an object, but are part of an operation” (2004). In Farocki’s well-cited account, operative images are those that *do* something in the world by providing a program for action. Key to the operative image is that it maintains an unclear relationship to its object—its representational aspects are considered somewhat arbitrary to the information it contains.

While the types of real-time target tracking and missile guidance systems that Farocki scrutinized still appeared to the human controller in the form of visualizations, an emerging paradigm in media theory takes the operative image to have vanished completely. Artist Trevor Paglen coins “invisible images” to describe those images “made by machines for machines” that effectively take humans “out of the loop” (2016). For Paglen, machine-readable images become

<sup>1</sup> For recent critical and artistic research into the origins of many of the datasets that have become standard for training facial recognition algorithms today see Adam Harvey and Jules LaPlace’s online platform project, *Megapixels* (megapixels.cc).

literally divorced from human sensibility—a feedback loop that constantly reconfigures the distribution of the sensible at any given moment. With machine learning, operative images are notably black boxed, that is, they are part of an operation shielded from the human eye. These images are operational abstractions of information that may be patterned according to a given logical method—they are an expression of a paranoid machine that functions to establish order within the incomputable mess of the world.

Simultaneously we note that although most digital images are no longer seen by human eyes, predatory and preemptive modes of algorithmic identification, such as those used for predictive policing in the United States, are used to justify decisions on human life. What role do digital images hold in straddling this apparent divide between micro-temporal computational calculations and the human experience of structural violence? If algorithmic vision is predicated on the abstraction of information and indeed the erasure of the human body, how do computational imaging practices evolve human concepts of perception and knowledge that shape human subjectivity? In the following, we extend beyond an analysis of algorithmic vision as a concretized mode of human thought to map the ways that the automation of perception destabilizes subjectivity itself.

#### 4 Algorithmic subjectivation: data reembodyed

Algorithmic vision technologies have changed our facial images. First, they disappeared into a black box. When they came back, they were bounded by a thin green box, an outline that denotes a labeled area of interest for a computer vision algorithm. For digital images to be useful in training an object recognition algorithm, they must be organized in accordance with some type of knowledge system, that is, they must be arranged in a taxonomic form so that an algorithm might deem what is and is not notable information in a certain context (Pasquinelli 2015; Paglen 2016). A taxonomy applies names to objects to reify them as part of an epistemology (Bowker and Star 1999). Boxes, in their supreme ability to standardize according to a norm, give form to the “governing of databased bodies” on a larger-than-life scale (Browne 2015).

In drawing the lines between what counts for what and what does not count at all, algorithmic vision can normalize certain types of behaviors, appearances, and codes of conduct in society. By installing a norm, algorithmic vision also defines what counts as an anomaly, obfuscating the logics that determine what counts as irregular (see Foucault 2009; Pasquinelli 2015; Amoore 2019). The use of algorithmic vision as a normative way of seeing begets urgent

aesthetic and political problems (see O’Neil 2016). Writes Jackie Wang:

If what we can perceive with our senses delimits what is politically possible, then how do we make legible forms of power that are invisible? How can we imagine ourselves out of a box that we don’t even know we’re stuck inside? Like a character in a Franz Kafka story, we are called into presence, managed, confined, and punished by an authority that we struggle to locate or identify, and every time we embark on a quest for answers, there is just infinite deferral and postponement. (2017: 52).

As Wang notes, we cannot *see* the systems that are operating on us, so how can we begin to understand what rationale they are following? Algorithmic vision technology is enigmatic as it necessarily hides its images to operate. These images do not adhere to visual critical analysis, because they complicate our given philosophical notions of the links between visuality and knowledge. When “intelligence” is recast as the algorithmic production of more information, pattern-finding becomes a methodology for generating knowledge in the face of overabundant data (see Steyerl 2014; Pasquinelli 2019). This makes any defined set of images into a program for hallucinating the future in the form of digital visual information. Algorithmic vision has a magical capacity to construe evidence from possibly meaningful correlations, patterns read into data.

Seeing like an algorithm requires a risk calculus—a derivative form that incorporates uncertainty to array possible futures. Louise Amoore explains that techniques of governance have shifted post 9/11 towards algorithmic means of calculating possible futures states (Amoore 2013). This sensibility shot through with the anxiety of potential catastrophe gave authority to new calculation techniques to incorporate unaccountable contingencies. Amoore explains that in the absence of sufficient data, the security algorithm “if a and b, in association with c, then x” is used to ontologically associate unknown values (2013: 59). This abstraction involves a continuous disaggregation and reaggregation of noncausal data, constantly shuffling the relations until a pattern deemed significant emerges. Algorithmic patterning facilitates decision-making processes based on what is *not* present to account for what *possibly* could be.

The creation of a proxy—or a stand-in that represents an unknown value—allows for the visualization of a portrait of a subject through non-visual means (Chun 2018). Algorithmic hallucination of subjects as a function of biopolitical control becomes a performance, run amok of an archive. Risk calculus is at play in the creation of a mutable norm that prioritizes economic decisions over truth so that action can be taken in the present moment (see Foucault 2008). Amoore explains how border technologies express this logic



in a space of unknown possibilities by drawing data from diverse sources, like fingerprints, linguistic analysis, travel habits, and financial records, to attach risk factors to individual subjects who are not part of this database already (see Lyon 2007). Sovereign decisions are made at the border on individuals that are created as governable subjects by the “life signatures” that are inferred onto them (see Agamben 2005; Ong 2006). The “biometric border” she explains, creates an emergent subject through combining fragments (2013). The subject is made through a correlation of elements composed of other subjects, objects, and the relations between them. She is drawn up in a profile that is taken as a governable body. She is visualized, in the sense that she is assembled in the process of computationally revealing something that isn’t actually present. Subjects are disaggregated and reaggregated as proxies made out of data hallucinated into the shape of a face.

The role of algorithmic vision technologies is key in an apparatus that images a population to render them invisible. The digital facial image, while sometimes linked to a real body, has the power to remake the body as a function of sovereign power, a life signature given form through a recognized face. A suspension of basic legal rights is justified through technological means. How does the very literal power that algorithmic systems have to define who does and does not count as a human recast the relationship between images and subjects? No longer bearing a primarily representative function, we see the power of algorithmic vision today is its ability to craft an algorithmic subject that is always open to contingency. There is always more the algorithmic subject than what meets the machinic eye.

## 5 Towards nonhuman visuality

In Farocki’s video work “I Thought I Was Seeing Convicts,” we are guided to consider the surveillance techniques used by a high-security prison in Corcoran (Farocki 2000). Farocki juxtaposes images together, showing us footage from security cameras alongside visualizations used by an analytics corporation to calculate the purchasing patterns of customers in a supermarket. “What can be accelerated and increased in prison?”, Farocki wonders. Body scan diagrams appear in the lower corner of the screen, bringing to mind the type of procedure familiar to anyone who has ever had to pass through the TSA. What can be accelerated and increased in prison?

Here, an astute viewer notices the timestamps on the security footage might assume the meeting of these live feeds in some sort of central control center, where this footage could be recorded and sorted to accumulate a profile of a certain inmate. Foucault’s panopticon levels up and gains an ability to see through both space and time (Foucault

1995). We learn that the profiles are amassed as evidence to determine what a specific inmate should and should not have access to, the yards they are permitted to occupy, and the inmates they are allowed to interact with. These images are not seen by the inmates but may be used by guards, for example, to place two inmates from opposing gangs in the same yards, resulting in deadly standoffs that are preempted by the guards for their own entertainment.

For Farocki, the security camera stands in metaphorically for the gun. Inmates need not be accosted with literal guns to be made to die. They are convicted based on patterns drawn by guards through the accumulation of data. A proxy of the inmate is sent in to standoff in the yard. Farocki reveals that this black box of decisions that caused one man to take another’s life contained nothing but a human corruption of power.

In the black box of algorithmic vision, perception is not just a mechanical program that epistemologically frames a subject, but it is also an operation. In this sense, algorithmic vision less like Barthes’ camera, and more akin to what Deleuze calls *visibilités*, or non-discursive processes that are discursively enacted to make subjects visible to power in certain assemblages (Deleuze 1988). Put differently, *visibilités* bring together power and its object in a way that makes the object visible in a determined way at any given moment. An evocation of futurity, where determination and realization are allowed to remain open until the moment they are closed in upon by power. In this sense, the subject known to computer vision is never out of the loop, so to speak, but a point along a continuum of images that constitute the assemblage from which individuals might be imaged again and again. Making an image of a subject is an operation necessarily enacted by algorithmic vision to capitalize upon virtual potentiality.

So, what of the ontological status of the operational image within algorithmic vision? When the image is considered within humanist discourses of visuality and representation, we might agree that the image has vanished from sight. However, to expound upon the operational nature of these images is to position them within both a material and technical universe, where potential images have the capacity to act on each other. To see like an algorithm is *not to perceive an image of something or someone*, but to produce a world of relations, the grounds from which subjects are made, seen, and named.

To argue for the necessity of human participation in the production of information and meaning, and ultimately as the arbiter of the development and direction of future technical objects is an ethical supplement to a critical analysis of computation that encourages the mutual coevolution of technics and culture towards more equitable ends. To claim the eradication of the human in the machine is to produce a conceptual blind spot, allowing for exploitation

and extraction to falsely appear an unfortunate side effect of technoscience, as opposed to paradigmatic. A more useful consideration might be to articulate the philosophical presuppositions of the human given through the very models of perception birthed by cybernetic science. To what extent do we need to modify our understandings of humanism in light of systems that do not have a human in the loop? The tension between the human and nonhuman contained within every system might be a starting point for critical thought to find new ways to challenge the visions created by systems of algorithmic governance.

**Acknowledgements** Thanks to Brett Zehner, Quran Karriem, Benjamin Crais, Jordan Sjol, Sophia Goodfriend, Mark Hansen, Mark Olson, Luciana Parisi, and the extended community around Computational Media, Arts & Cultures at Duke University for their input and comments.

**Funding** Not applicable.

**Data availability** Not applicable.

**Code availability** Not applicable.

## Compliance with ethical standards

**Conflict of interest** Not applicable.

## References

- Agamben G (2005) *State of Exception*. University of Chicago Press, Chicago
- Amoore L (2013) *The Politics of Possibility: Risk and Security Beyond Probability*. Duke University Press, Durham, NC
- Amoore L (2019) Doubt and the algorithm: on the partial accounts of machine learning. *Theor Cult Soc* 36(6):147–169
- Apprigh C, Steyerl H, Chun W, Cramer F (2018) *Pattern Discrimination*. University of Minnesota Press
- Barthes R (1981) *Camera Lucida: Reflections on Photography*. Hill and Wang, New York
- Bowker G, Star S (1999) *Sorting things out: Classification and its consequences*. MIT Press, Cambridge
- Browne S (2015) *Dark Matters: On the Surveillance of Blackness*. Duke University Press, Durham
- Buckley C, Mozur P (2019) How China Uses High-Tech Surveillance to Subdue Minorities. *New York Times*. <https://www.nytimes.com/2019/05/22/world/asia/china-surveillance-xinjiang.html>. Accessed Dec 2019
- Byler D (2019) China's hi-tech war on its Muslim minority. *The guardian*. <https://www.theguardian.com/news/2019/apr/11/china-hi-tech-war-on-muslim-minority-xinjiang-ughurs-surveillance-face-recognition>
- Chun W (2018) On patterns and proxies, or the perils of reconstructing the unknown. *e-flux Architecture: Accumulation*. <https://www.e-flux.com/architecture/accumulation/212275/on-patterns-and-proxies/>. Accessed Nov 2019
- Cooper D (2020) *Invisible desert*. *e-flux Architecture: New Silk Roads*. <https://www.e-flux.com/architecture/new-silk-roads/313103/invisible-desert/>. Accessed Feb 2020
- Deleuze G (1988) *Foucault*. University of Minnesota Press
- Farocki H (Director) (2000) *I Thought I Was Seeing Convicts* [Motion Picture]
- Farocki H (2004) *Phantom Images*. *PUBLIC* 29:12–22
- Flusser V (2000) *Towards a Philosophy of Photography*. Reaktion Books
- Foucault M (1995) *Panopticism. Discipline and Punish: the Birth of the Prison*. Vintage Books, New York, pp 195–230
- Foucault M (2008) *The Birth of Biopolitics*. Palgrave Macmillan, New York
- Foucault M (2009) *Security, Territory, Population*. Palgrave Macmillan, London
- Franklin S (2015) *Control: Digitality as Cultural Logic*. MIT Press, Cambridge
- Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning*. MIT Press
- Halpern O (2015) *Beautiful Data: A History of Vision and Reason since 1945*. Duke University Press, Durham
- Hansen MBN (2014) *Feed-Forward: On the Future of Twenty-First-Century Media*. University of Chicago Press, Chicago
- Harvey A, LaPlace J (2019) *DukeMTMC. MegaPixels: origins, ethics, and privacy implications of publicly available face recognition image datasets*. <https://megapixels.cc/>. Accessed Dec 2019
- Hoel A (2018) *Operative Images. Inroads to a New Paradigm of Media Theory. Image – Action – Space*. De Gruyter, Berlin, Boston. <https://doi.org/10.1515/9783110464979-002>
- Hoelzl I, Marie R (2015) *Softimage: Towards a New Theory of the Digital Image*. Intellect Ltd, Chicago
- Human Rights Watch (2019) *China's algorithms of repression: reverse engineering a Xinjiang Police Mass Surveillance App*. Human rights watch. <https://hrw.org>. Accessed Dec 2019
- Lettvin JY, Maturana HR, McCulloch WS, Pitts WH (1959) What the frog's eye tells the frogs brain. *Proc IRE* 47(11):1940–1951
- Lyon D (2007) Surveillance, security and social sorting. *Int Crim Justice Rev* 17(3):161–170
- MacKenzie A, Munster A (2019) Platform seeing: image ensembles and their invisibilities. *Theor Cult Soc* 36(5):3–22. <https://doi.org/10.1177/0263276419847508>
- Mirowski P (2002) *Machine Dreams: Economics Becomes a Cyborg Science*. Cambridge University Press, Cambridge
- Murgia M (2019) Who's using your face? The ugly truth about facial recognition. *The financial times*. <https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e>. Accessed Sept 2019
- Ningning Z (2016) Big data “out of traffic”. *Southern Magazine*. <https://epaper.southcn.com>. Accessed Sept 2019
- O'Neil C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, New York
- Ong A (2006) *Neoliberalism as Exception: Mutations in Citizenship and Sovereignty*. Duke University Press, Durham
- Paglen T (2016) *Invisible Images (Your Pictures Are Looking at You). The New Inquiry*. <https://thenewinquiry.com/invisible-image-s-your-pictures-are-looking-at-you/>. Accessed May 2019
- Paglen T, Crawford K (2019) *Excavating AI The Politics of Images in Machine Learning Training Sets*. <https://excavating.ai>. Accessed Sept 2019
- Parisi L (2013) *Contagious architecture: Computation, aesthetics, and space*. MIT Press, Cambridge
- Pasquinelli M (2015) *Anomaly detection: the mathematization of the abnormal in the metadata society*. *transmediale*. Berlin
- Pasquinelli M (2019) *How a machine learns and fails: a grammar of error for artificial intelligence*. *Spheres*, vol 5
- Patterson Z (2015) *Peripheral Vision: Bell Labs, the S-C 4020, and the Origins of Computer Art*. MIT Press, Cambridge
- Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) *Performance measures and a data set for multi-target, multi-camera tracking*. European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking

- Satcky J (2019) A Duke study recorded thousands of students' faces. Now they're being used all over the world. *The Chronicle*. <https://dukechronical.com>. Accessed June 2019
- Seaver N (2017) Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data Society* 4(2):205395171773810
- Solera F, Calderara S, Ristani E, Tomasi C, Cucchiara R (2016) Tracking social groups within and across cameras. *IEEE Trans Circ Syst Video Technol* 27(3):441–453
- Sound Vision Foundation (2019) About Uighurs. Save Uighur Campaign. <https://doi.org/saveuighur.org>. Accessed Oct 2019
- Steyerl H (2014) Proxy Politics: Signal and Noise. *e-flux journal*. <https://www.e-flux.com/journal/60/61045/proxy-politics-signal-and-noise/>. Accessed May 2019
- Wang J (2017) *Carceral Capitalism*. MIT Press, Cambridge
- Xu J, Zhao R, Zhu F, Wang H, Ouyang W (2018) Attention-aware compositional network for person re-identification. 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 2119–2128

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Seeing threats, sensing flesh: human–machine ensembles at work

Perle Møhl<sup>1</sup>

Received: 30 July 2019 / Accepted: 18 August 2020 / Published online: 9 September 2020  
© Springer-Verlag London Ltd., part of Springer Nature 2020

## Abstract

Based on detailed descriptions of human–machine ensembles, this article explores how humans and machines work together to see specific things and unsee others, and how they come to co-configure one another. For seeing is not an automated function; whether one is a human or a machine, vision is gradually enskilled and mutually co-constituted. The analysis intersects three different ways of human–machine seeing to shed further light on the workings of each one: an airport, where facial recognition algorithms collaborate with border guards to grant passage to particular travellers and not to others; a luggage-scanning system, where potential security threats are assessed by a complex of X-rays and human intro-spection; and a hospital operating room, where human–machinic surgical robots find their way and operate on the insides of human bodies, touching only by seeing. In these examples, human and machine ways of seeing merge together, seeing in particular apparatuses of material, political, organisational, economic and fleshy components. The article analyses the practical work of human–machinic collaboration and explores how the different material and social constituents, not necessarily always working from the same agenda, come to configure what can be seen and sensed and what cannot.

**Keywords** Human–machine interfaces · Visual enskillment · Sensory anthropology · Facial recognition · X-ray scanning · Robotic surgery

## 1 Introduction

This article describes how agencies engaged in border control and robotic surgery see in the technological taskscapes they are engaged in for specific predefined purposes. It is based on and largely consists of detailed descriptions of three settings in which humans and machines work together to see specific things—and unsee others—as defined by their professional tasks and the technological and human sensorialities, instruments, capabilities and automatisms they have at their disposal. A main point is that seeing is not an automated function, whether one is a human or a machine. The contention behind the article is that vision is enskilled and that the technological specifics, together with the specific tasks and material settings, co-configure the collaborative enskillment and work of seeing and sensing. Human and machine ways of seeing merge together to produce particular seeing apparatuses that are constituted by material, political,

organisational, economic and fleshy components. The aim is to explore how all these constituents, which do not necessarily work in the same direction, come to configure what can be seen and sensed and what cannot.

The article takes us to three different settings: an automated border control booth in an airport, where facial recognition algorithms collaborate with border guards to grant particular travellers passage and not others; an intricate luggage-scanning system in another airport, where potential security threats are assessed by a complex of X-rays and human eyes intro-specting the contents of suitcases; and a hospital operating room, where human–machinic surgical robots find their way and operate on the insides of human bodies, touching only by seeing. The three different ways of human–machine seeing may, when intersected, shed further light on the workings of each one.

✉ Perle Møhl  
perle.mohl@antro.org

<sup>1</sup> CAMES-Copenhagen Academy for Medical Education and Simulation, Capital Region, Copenhagen, Denmark

## 2 Fields of inquiry and some theoretical anchorages

Most of the analytical work in this article takes the form of empirical descriptions and discussions. The analytical deliberations nevertheless find affinity and anchorage in the theoretical concepts presented in the following.

A first such affinity concerns seeing. It is not because you have eyes that you can see. Seeing is not an inherent physiological capacity, nor an invariable mode of perception. It is a learned “technique of the body”, ingrained through apprenticeships that vary from practice to practice, from society to society (Mauss 1973). ‘Culturally inculcated and socially performed’ (Grasseni 2011: 20), vision undergoes a particular *enskillment* (Grasseni 2007, 2011). Enskilled seeing is a situated practice, related to particular “taskscape” and involving particular forms of expertise and particular ways of knowing that emerge in “communities of practice” (Grasseni 2011; Ingold 2000; Lave and Wenger 1991). An enskilled vision is a purposeful albeit often implicit way of scrutinising the world. A distinction is often made between looking and seeing (e.g., Okely 2001; MacDougall 2006), between unreflectively letting light waves imprint themselves on one’s retina, mostly characterised as ‘looking’, and an attentive, purposeful, meaningful way of engaging and knowing characterised as ‘seeing’. Enskillment, however, replaces this dualism with a gradual getting to see and know that clearly characterises all engagement with the world, including perceptive, and the many particular modes of vision that take place in particular settings and their particular interactions and activities. I apply this idea of enskilled vision in this article and consequently use looking and seeing indiscriminately, not as oppositions. Enskillment can also be related to a pragmatic semiotic conception of signs and meaning as variegated, positioned, partial and related to a wide range of different agentive modes of experience and interpretation (Møhl 1997; Peirce 1998). Where a tourist sees a nice green spot for a picnic, the peasant sees a patch that needs harvesting, and the cow sees food.

Cattle herders and their capacity to see, the particular enskilled vision they develop, emerges as a frequent trope or figure in descriptions of such tacit ways of knowing by seeing: the particular visual expertise that it takes to know a good dairy cow when you see one, to assess its capabilities, the quality of its muscles, fur, horn, hooves and udders, and to distinguish two cows from afar, not to mention certain practical-aesthetic forms of visual know-how (Berger 1972; Berger and Mohr 1982; Møhl 1997; Grasseni 2004; Grimshaw and Ravetz 2005). Such visual ways of knowing go before words (Berger and Mohr 1982) and sometimes also, I contend, beyond words.

If vision is not inherent in the body, if ways of seeing are enskilled and change from one setting to another, what are the implications and effects of working with and seeing through a seeing machine? How does such a collaboration further enskill and configure human ways of seeing? And what does the practice of seeing together—co-sensing—say about human–machine interlacing and co/multisensory interactions, or about the nature of the co-sensing ensemble? In other words, the discussion of visual enskillment necessarily takes a posthuman turn when practices of seeing are carried out in human–machine collaborations.

Using machines to see—see in greater detail, see further away, see things the human eye cannot look at, see from nonhuman/omniscient/god-like positions—goes far back in history, e.g., the use of magnifying lenses in Antiquity, eyeglasses and telescopes in the early seventeenth century and the photographic camera in the early nineteenth century (Hirsch 2017). In this article, direct human vision and cameras both constitute fundamental elements of the settings being analysed. As Berger notes, one thing that changed the way humans saw was the invention of the camera (Berger 1972:24). This idea runs as a red thread through visual anthropology: seeing through a camera creates other possibilities for moving in and interacting with the world (Rouch 1975; Rouch 2003), other ways of knowing and getting to know (MacDougall 1998; Grimshaw and Ravetz 2009; Møhl 2011), other ways of telling (Berger and Mohr 1982), other ways of relating to people and “presensing” those who are far away in time or space (Deger 2006; Møhl 2011).<sup>1</sup> Cameras are integrated components of the sensing ensembles described in this article, whether in relation to border control, X-ray scanning or surgical robotics, and their material specifics, therefore, contribute, along with other, more abstract constituents of seeing ensembles, to define what can be seen and sensed and what cannot.

Fundamental to any analysis of ways of machine seeing is also an understanding of human perception and sensing in general. Importantly, a fundamental point of departure for this article is the notion of the integration of the senses or, in Merleau-Ponty’s terms, the synaesthetics of perception, the fact that no sense ever works alone but is intricately interwoven with the other senses and with bodily and cognitive functions in general (2005). From a posthuman perspective, the notion of synaesthetics must be stretched beyond the human senses to integrate the full co-sensing apparatus—cameras, optic sensors, recognition algorithms, X-rays and

<sup>1</sup> My own research projects have generally been based on the practical, intersubjective and epistemological aspects of working with and creating anthropological knowledge through a camera, as well as the particular ways in which the filming process configures and frames one’s approach to and interactions with the world (Møhl 1997, 2011, 2012; Møhl and Kristensen 2018).

image analysis algorithms, etc.—whether the collaboration takes place on a border or in an operating theatre. And when it comes to the enskillment of vision and the professional practices and “taskscape” around which enskillment is organised (Ingold 2000:194; Grasseni 2004), the particular community of practice that comes to see necessarily also includes the material sensing components, as well as the people and institutions that developed and applied those components and their capacities to see certain things and not others. Thereby a broader network of agencies also comes to reconfigure the human operator’s sensory work and engagements with the seen and the unseen in particular ways. The synaesthetic effect becomes far-reaching.

Another affinity concerns synaesthetics and the sensory apparatus. Berger discusses the interlacing of vision and touch and describes the effects of looking at a painting: “Every square inch of the surface of this painting, whilst remaining purely visual, appeals to, importunes, the sense of touch”, “what the eye perceives is already translated... into the language of tactile sensation” (Berger 1972: 138–139). Analysing film’s capacity to convey a sense of touch visually, Marks (2000, 2002) uses the term “haptic visuality” to describe “the way vision itself can be tactile, as though one were touching a film with one’s eyes” (2000: xi). Although she distinguishes haptic visuality from optical visuality that sees things from a distance (idid:162), she wishes to reinstate vision in general as a form of contact, contrary to a post-Enlightenment rationality that sees vision as disembodied and distancing (Marks 2002: xiii). These two analyses consider the haptic qualities of visual media. In this article I use the term “haptic vision”<sup>2</sup> to describe the capacity to directly feel or touch with the eyes from a sensorially interlaced or synaesthetic perspective.

In the empirical descriptions, following Haraway (2017) I characterise the seeing human–machine ensembles or apparatuses as bionic or cyborgs. In her approach, the cyborg figure constitutes a methodological tool and an epistemological challenge that dissolves dichotomies and other blunting oppositions—between nature and culture, human and nonhuman, machinic and organic, as well as between the senses—a tool that reconceptualises the agential object (Suchman 2007; Lynes 2016; Haraway 1988). The cyborg is a good figure of speech. But I also see the cyborg as a matter of fact, a concretely existing working body in the present examples of airport control and robotic surgery, where humans and technologies merge together to perform specific tasks, mutually formatting and enskilling one another. I use the cyborg

figure, because it illuminates how machine and human ways of sensing merge, extending a purely human synaesthetic effect to its material components. Indeed, the cyborg makes sense not only as an analytical tool or an allegorical figure installed to think and disrupt dichotomies, but here as an actual material-semiotic working body of flesh and metal and algorithms, where some parts make other parts sense in new ways. The cyborg is a more concrete figuration of the other concepts I also use, namely ensemble or apparatus, and a question, therefore, imposes itself: where the cyborg ends, if it does. For the agencies that take part in the tasks and sensory work are far-reaching.

### 3 Empirical instances of seeing and co-sensing

In the following I present three settings in which humans and machines work and merge together in a sensory task of introspective vision work. The three examples are taken from my two most recent research projects on the visual, technological and sensory aspects of border and security control and robotic surgery, respectively. In all three cases, human and machine sensitivities coalesce to form vision apparatuses that develop particular ways of seeing that are adapted to the tasks at hand. These tasks are carried out in very different material, sensory, political and economic conditions that co-configure ways of seeing and objects seen and unseen. Intersecting the empirical descriptions of these three different interlaced human–machine ways of seeing is a diffractive move that serves to highlight the sensory effects created by those differences rather than the differences themselves (Barad 2007; Haraway 1992).<sup>3</sup> The intersections also allow us to discern some patterns in the political economies of the sensing ensembles and the difficulties involved in determining their limits.

The analysis starts with airport and border control, where humans and machines in different constellations inspect faces and luggage in the search for threats. As we shall see, three different modes of seeing are in action at the airport, and three different effects of seeing emerge. With the final example, these differences are further diffracted by yet another mode of machine seeing and touching by seeing, namely robotic surgery.

<sup>2</sup> The term “haptic vision” is also sometimes used in the inverse sense to describe the capacity to see an object by touching it, a common example being the blind person. This does not, however, interfere with the project of dissolving the distinctions between the senses—on the contrary.

<sup>3</sup> Diffraction, a term from classical physics, describes the effect of waves, e.g. light waves, being bent when they hit an obstacle. Different wavelengths are not bent to the same degree, resulting in a particular light pattern. What we see are not the differences in wavelengths but the effects of those differences when they are curved around an obstacle. As Haraway notes, “A diffraction pattern does not map where differences appear, but rather maps where the effects of difference appear.” (Haraway 1992:300).

### 3.1 ABC vision

Airports are sites of augmented vigilance and perpetual monitoring, carried out for multiple purposes by a human–material and to varying degrees automated constellation of cameras, one-way mirrors, optic and heat sensors, radars, X-ray scanners and human scrutiny, to mention just some. There are many ways of seeing. In surveillance and control, there is no purposeless looking, no casual or unintentional glances; there is always some kind of analysis taking place, some kind of guided interpretation of what is seen.

At Copenhagen International Airport, a police officer is surveying passengers leaving Schengen through the Automated Border Control (ABC), a system using passport-scanning and facial recognition. The officer, Hanne, has logged into the system, including the different national and international police databases, and has two screens at her disposal, as well as a full view of the six open eGates. Her auxiliary screen lights up if the ABC detects a hit between data in the scanned passport and the databases, e.g., full records or sparse data on wanted or discretely surveyed persons. But Hanne’s attention is focused on the passengers and on her main screen, where she can follow the workings of the automated facial recognition system.

Travellers scan their passports, the first glass door opens, and the traveller is requested to stand on a pair of yellow footprints and face the camera. The ABC system scans the traveller’s face and extracts a series of “minutiae points” that are turned into facial recognition templates, which are compared with the equivalent template stored in the passport chip. The system requires a certain level of resemblance—the recognition threshold—between the live face in the ABC, the passport chip and the ID photo. On the screen, Hanne and I can see how the comparison score moves up and down until it eventually meets the threshold and the second glass door opens. In sum, the system requires that the person in the eGate can present a face that sufficiently resembles the ID photo in the passport. This entails complying with the requirements of the recognition system by removing hats, veils or earphones, not smiling too much and looking straight at the camera. Hanne explains that if the sunlight hits the camera it can’t see and if people have too much hand-luggage or wear t-shirts with printed faces, “it gets confused”. “It’s very sensitive”. If there’s too much confusion, it will spread to the other eGates like a virus, and they will have to restart the whole system, she explains.

So Hanne’s job is to keep an eye on people, but also on the machine’s state of mind. She has both to look through the glass at people “with her own eyes”, as she says, and also be able to “see like the machine”: what it detects, what it gets wrong, what will confuse it, and how can she keep confusion to a minimum so it doesn’t glitch. If the machine starts hesitating—shifting between recognition scores or exhibiting

red alert signs, indicating that it is seeing “more than one person in the eGate” or no resemblance whatsoever—she can override the machine and relieve it of its qualms. This requires her to be able to read the ID photo and the facial scan on her screen and compare them with the actual face she can see in the eGate. To this effect she is “learning to see in 2D”, she says, comparing the 2D ID-photos on her screen with the actual physical 3D faces out there and trying to see both resemblances and dissemblances. Checking if a person sufficiently resembles an ID photo is, of course, a recurrent task for most police and security controls, something Hanne herself often does in the “manual” passport control booth. But here she has to add several other layers to her seeing, mimicking the seeing of the machine and searching for possible dissemblances or other interferences that might be triggering its dissatisfaction or lowering the resemblance score. In that sense, she has to see like but also see better than the machine, even though she does not dispose of its optic and computational capacities.

In this process of interlaced human–machine seeing, the officer learns to see what the machine sees to understand and override it. Where the police officer sees faces and crevices, wrinkles, bumps, colours and shadow zones, the machine sees surfaces, and extracts and compares minutiae points. Where the machine sees two persons in the eGate, because it detects faces and masses, the officer sees a woman with a lot of handbags or a person with a magazine front page showing a natural-size face. The operators gradually “get to know” how the machine sees, what it registers and how it gets “confused”, and they learn to “see in 2D” like the machine.

Zooming out from Hanne’s booth, other issues and other types of machine seeing enter into the field. Passport control is an obstacle to steady passenger flows in the airport and a frequent issue of debate and negotiations between the police authorities and the private company running the airport. The company would like to see the resemblance threshold lowered so people can go across the border more smoothly. To this effect, a large screen in the police HQ displays the queueing for passport control “to urge police officers to lower their vigilance”, as some of Hanne colleagues explain it (see also Møhl 2019). They are worried about the pressure and negotiations. As one of the critics of the ABC system says: “its threshold can be lowered by the click of a mouse; mine can’t!” They clearly feel that their authority is being undercut by the automated system and by the economic and political pressures behind it. They also recognise that the ABC is not efficient enough to run the border by itself, at least not in the immediate future. But the unease about the future and about being surveyed by the airport company and other public and supra-state institutions regulating border control is palpable. There are cameras and presence sensors everywhere, as well as seeing at many levels and of many sorts in the airport passport control.

### 3.2 X-ray scanner visions

In the following I describe two types of X-ray luggage-scanning that take place at Gibraltar International Airport. The two processes might seem very similar, but they each implicate particular ways of seeing that come out of and are configured by technological specificities, temporal constraints, organisational requirements and preconfigured images of threats.

Deep inside the newly built airport building, luggage is transported from passengers' hands into the departing airplanes through a complex system of conveyor belts that run up and down at many levels, circumventing the lounges and transit areas, where passengers move through the airport. The back-stage conveyor belt system was clearly designed and assembled after the front-stage passenger areas were built, creating this intricate cobweb of impressive metal structures. The clanking of pulleys and rubber belts, strewn with sensors, cameras and scanners, rises to infernal levels and makes talking almost impossible. But above all we are here to see.

Around the conveyor belts, a four-layered system of security levels, both computed and human and of varying degrees of automation, has been established. The first level, security level 1, is wholly automated: all luggage is run through an X-ray scanner that is set to make out objects and densities, and determine whether the pieces of luggage can go directly to the plane or require closer inspection. If known threats are detected (e.g., weapons, organic explosives, electronic circuits) or if the scanner cannot make out the details, the suitcase is pushed off the conveyor belt and onto a separate track.

Far from the conveyor belts and the noise, in an isolated room with thick walls that from the outside has the appearance of a bomb shelter and may have been designed to give that impression of invulnerability, Jane, a police officer, is inspecting the X-ray images of the suitcases that the automated scanning system has deemed suspect and sent down the alternative track. A sign on the door reads "security level 2". The amount of luggage being sent on to this level is quite high, not because luggage often contains threats to flight security—in fact, very few if any have ever been found—but because they contain objects or meshworks of electric gear that cannot be identified or simply because they are so densely packed that the scanner cannot make out the details. Therefore, Jane and her vision take over. She has twelve seconds to look through each image and decide whether the piece of luggage needs more meticulous inspection—if so, she presses the red button and sends it off to security level 3—or whether it can be considered safe and sent back into the normal circulation track, in which case she hits the green button. Jane can shift through the layers and change between different X-ray frequencies to reveal some of the

compositions, densities and contours of the contents, but she rarely has the time to do so. If she takes too long, hesitates or is distracted, the virtual 12-s hourglass on her screen runs out, and the luggage is automatically sent on to level 3. Such instances of hesitation or unresponsiveness are analysed as inattention and are audited and presented to her at the end of the month, along with her pay check. In other words, as in passport control, it is not only the luggage but also the operators that are being screened and monitored here.

A new suitcase pops up on Jane's screen, and I hastily make out some screwdrivers and what looks like safety shoes before the hourglass runs out and Jane hits the green button. Before a new X-ray image pops up on her screen, I ask her about the safety shoes to check my own visual skills, but she didn't see them. In fact, she is incapable of telling me what was in the suitcase she's just cleared and sent back to the ordinary conveyor track. "I don't really look at the objects", she says, "I don't have time for that. I look for the usual threats, but I mainly try to figure out why the machine sent this suitcase on to me." So in sum, her work is to try to see—figure out—what the machine saw or, in most cases, indeed, didn't see. She is not looking for particular things but for the fuzzy zones, the questionable unknowns, turning her sight into a kind of inverted vision. In that sense, she and her scanner deploy a complicit form of non-seeing, carried out in a tiny human-machine community of practice, adapted to the very specific task at hand, and formatted by the fact that there are very rarely, if even ever, any real threats to be seen. This way of seeing emerges in the interaction or, more accurately, intra-action (Barad 2003), a searching for what cannot be seen that is formatted in this particular setting between Jane, the X-ray scanner, the densities, the tasks, and the time restriction and hourglass monitoring system.

There is, however, one particular type of object that does appear regularly on Jane's screen and that she is required to notice and mark as seen by hitting the red button. This "object" is a so-called threat image projection (TIP), an image of a known threat—often handguns—projected onto a piece of luggage. These projections have a double function, according to the producers (Rapiscan 2017): to keep the operators alert to threats and to remind them what threats look like, exactly because they hardly ever see any real threats—as well as to check that they are alert and looking at the screen. Indeed, missed TIPs are counted and presented to Jane at the end of the month, she explains. The TIP system uses a bank of images of known threats and thereby requires that the officers are able to detect only those known threats at the risk of being inattentive to hitherto unknown threats. This produces what I have called a visual agnosia to the unknown, because it undercuts what the police officers describe as their creative capacity for imagining things that are unknown and projected into the future, exactly what the pre-programmed vision of the machines cannot do (Møhl



2019, 2020a, 2020b). Besides configuring this function of recognising only the known, the TIP system also produces a particular form of operator vision that is alert to the system's internal threats, its built-in control mechanisms, and the particular colours and strange contours of the artificial TIP objects. Indeed, from what they say, when officers detect TIP images, it is actually not the object itself that they perceive but the uncanniness of its projection, especially at this level of security, where time is a scarcity.

The way Jane comes to see mainly fuzzy zones and what cannot be seen in collaboration with her level 1 scanner—their particular enskillment of an inverted vision—stands out, because it so clearly contrasts with the way the police officer and the more elaborate scanner on security level 3 come to see. In addition, where these two modes of seeing meet and produce a diffractive potential, time seems to be of the essence to the different forms of seeing that appear.

When Jane hits the red button or lingers too long, the scanned piece of luggage is sent on to yet another conveyor belt track that rolls it right into security level 3, situated in the middle of the conveyor belt bustle and noise. An alert sounds, a lamp starts blinking, and the piece of luggage appears physically in front of the police officer who has to lift it into a huge and more proficient X-ray scanner. Here, a very different type of intro-spection starts, for at this level, there are fewer pieces to look at and much more time to go through the details. Paul, the officer, not only sees the outside of each piece of luggage, but can also minutely go through all the objects contained within it with the X-ray scanner. The X-ray scanner can show more details and scan the piece of luggage from all sides. But mainly, Paul has time to try out different visualisations, zoom in and out, and to look attentively at and identify the objects in the pieces of luggage. Paul is looking at content, not form. And only when the fuzzy zones persist and he cannot make out the contents is the piece of luggage sent onto the ultimate security level, security level 4, where the passenger is called down to an explosion-proof room to open the luggage and go through the contents. This happens only rarely, maybe once a week, Paul estimates.

Paul and his scanner are able to make out and identify objects, because some X-ray frequencies make certain parts stand out more clearly while making other parts invisible, and other frequencies make exactly those formerly invisible parts clearer. A pair of curved metal wires show up several times. He quizzes me, and I learn that they are metal wires in bras. Strange square plates with holes in them turn out to be metal inserts in leather shoe soles; Paul can identify more expensive, robust leather shoes by the fact that they do not have these metal inserts, he explains.

In making decisions, Paul works through qualified guesswork and induction by positively identifying certain objects and inferring what the others might be, or by explaining

the nature of the composition of objects by who the owner might be. He puts together life stories, as he says. He reads coherent life stories out of a selection of the material fragments of those lives. A duffle bag held together by rope turns out to contain a series of long knives and some unidentifiable objects entangled in electronics that were probably what made his colleague in level 2 hit the red button. Paul infers from the duffle bag and the knives that the traveller is a Philippine chef disembarking in Gibraltar, and also that the unidentifiable objects in the bag must be some kind of kitchen utensil. A densely packed suitcase contains both women's and men's shoes, from which he infers that this is a couple travelling together with only one suitcase, which is why it is so heavily packed and difficult for the human and machinic seers on the former security levels to see through. There is a logical explanation to the particular constellation of objects, known and unknown, visible and fuzzy, as well as to the fuzziness itself, which clears the pieces of luggage of suspicion. It's all about seeing the logics.  $2 + 2 = 4$ . And it's about having the necessary time to exploit the intro-spective scanner vision technologies to their fullest.

### 3.3 Robot visions

At Herlev Hospital in the Capital Region of Denmark, surgeons operate with robots in three main domains, urology, gynaecology and gastrointestinal surgery. Herlev has four robots in the operating rooms, and an older model is used for training and simulation in the basement. They are all “da Vinci” robots produced by an American company, Intuitive Robotics, which has international monopoly of surgical robots. “But other companies are biting at their heels”, Niels, chief urologist and experienced robotic surgeon explains to his trainees, surgeons and surgical nurses who will soon be working with the robot. During the training program they learn to wrap the robot's arm in disposable drapes, install it by the operating table, position the patient and insert the robotic instruments inside the patient. They also learn to deinstall everything very quickly in case something goes wrong and they have to switch to open surgery. Once everything is ready, they start learning to use the robot, taking turns at the console in operating on a dummy, working at a distance with hand controllers and foot pedals. They train to move the robotic arms and the camera, switch between instruments and make stitches on a phantom silicone organ positioned inside the dummy patient. Later in the training program, they move on to an animal model, where things become more lifelike and the tasks more acute. There is bleeding, real organs that can be injured and a real risk of killing the “patient”. From the console, they learn to use the robot to find their way inside the body, identify tissue and organs by means of the camera, and perform operative procedures such as separating connective tissue and partially

or totally removing organs, blood vessels and nerves, stopping bleeding by cauterising and clamping vessels, making stitches and reconnecting tubular organs—all this at a distance by connecting to the robotic arms and instruments and seeing through the endoscopic camera.

The viewing experience is astonishing. It is 3D—3DHD in the latest version—and you can sense the depth and relative distances, you can move around very narrow passages, open tiny crevices, grab minuscule nerve-vessel bundles, and push tissue and organs with your small robotic hands. The camera is equipped with a strong light and works with a factor 10 magnification. Following a kidney operation on the screen in the OR, the movements of the hands in the abdominal cavity make me think of a miniature speleologist gradually progressing, cutting and slicing and searching for openings, in a close-up subjective shot. The imagery produces a sensation of intimacy and immediacy, a very uncanny feeling of being there, inside the patient's body.

Although surgical robots were originally built for tele-operations, the surgeon consoles at Herlev are positioned inside the ORs at approximately 2–3 m from the operating table. The robotic system is equipped with a two-way sound system with microphones and loudspeakers at both ends, but the sound quality is not always optimal, and they are often turned down or off. This means the surgeons can only vaguely hear the discussions, comments and sounds from the operating field and patient. To communicate with the team, they often have to raise their heads from the console, momentarily halting the operation.

Surgeons speak with satisfaction of heightened precision and concentration, and how they actually like the distance and relative sensory isolation the robot provides. They can work for hours on end without getting tired. The robot arms even eliminate any possible trembles. Although the robot is not intelligent or even automated, and hence not a robot in the classic sense, there is a bionic quality to this surgeon–robot assemblage, where the robot's many arms, 3DHD vision, force and steadiness merge together with the human operator and enhance the surgeon's acuity, precision and endurance.

When sitting at the console, looking down through the stereoscopic 3D viewfinder into the patient's abdominal cavity, the small robotic instruments move just like one's own articulated hands, turning 360°, grasping tissue and passing needles, and the position of the “hands” in the viewfinder corresponds to where one's own fleshy hands are manipulating the console controllers. It is all very intuitive, as the company name also indicates. The robotic hands are more than bodily extensions; they take over the place and role of one's own arms and hands, like a prosthesis, making the robotic ensemble look convincingly like a cyborg.

But like a prosthesis, the hands lack one essential feature: the sense of touch. The prosthetic hands do not feel anything,

and the surgeon has no haptic feedback whatsoever. Because of the enormous force of the robot's hands, this calls for extreme attention to movements and operative interventions inside the patient. This lack of haptic feedback is considered particularly challenging by robotic surgeons, both at Herlev and in medical scholarship in general (e.g., Bethea et al 2004; Rangarajan et al 2020).

Thus the surgeon in the console cannot feel what the robotic hands are touching, what the instruments are doing and the force they are applying, and can barely hear what is going on in and around the patient. All the sensations that surgeons normally rely upon in their interactions with patients, including smell and proprio- and alteroception, are in sum reduced to and rely upon one single sense: vision. And although designers and industry are trying to remedy this limitation by searching for technological solutions, I believe the solution is to be found within the pre-existing robotic ensembles, where surgeons enskill their vision to carry out this haptic perception. This becomes apparent both in the training programs and in the OR, where the participants tacitly enskill their eyes to sense the densities and resistances of the different types of bodily organs and tissues. They are perfecting what can only be called a synaesthetic haptic vision. In keeping with Merleau-Ponty's notion of synaesthetic perception (2005), this faculty to feel with one's eyes is already a human capacity that is then further enskilled, in the robot ensemble and its particular human–machinic constellation, to the particular tasks it needs to carry out. The phenomenon of interlaced senses contrasts with a prevalent taxonomic and anatomical ontology of the senses as perfectly isolated, correlating with a fragmented body and anatomically isolated organs—a taxonomic and functional mapping out that Leonardo da Vinci incidentally partook in (Foucault 1972; Hillman and Mazzio 1997). It replaces sensory inputs of pressure magnitude with sensory inputs of the spectrum of electromagnetic waves of light. There is an inverted parallel in the way an STM (a scanning tunnelling microscope) forms an image of an atomic specimen by touching it with a tungsten tip, “an encounter that engages the sense of touch rather than sight” (Barad 2007: 52).

The way this bionic surgical robot feels with its human eyes is generally tacit and non-explicit. During training the problem of lacking haptics is usually mentioned, but there are no specific words for describing it, simply frequent calls to attention, “watch out, you're pulling too hard”, “be careful near the kidney”, “don't push the spleen so much”. Digital simulation tools built into the da Vinci robot are used to train and evaluate movement and precision, but not haptic vision as such. Nevertheless, the surgeons seem to rely on their prior open and laparoscopic experience, where they couple vision with a direct haptic sense of the instruments, tissues and organs they touch. They also learn

to estimate pressure from different kinds of visual signs. For one thing, they can directly see the movement of the tissue that the hands are touching and estimate the force applied—although in training they initially push too hard and sometimes perforate organs or pull stitching sutures through tissue. Secondly, some tissue reacts to pressure by changing colour, generally because of altered blood flow, and this is also used as an indication of pressure. Thirdly, pressure on organs can be visually evaluated by the light reflections on the reflective surfaces. But again, all these visual skills are generally tacit, only occasionally being made explicit in training situations and supervisions in the OR, as well as in my discussions with robotic surgeons. Indeed, vision not only comes before words, as Berger reminds us (1972), vision often also elides wording, especially when it comes to haptic vision.

In the robotic cyborg, human and machine parts are connected and interlaced, just as the different sensory qualities operating in the cyborg are synaesthetically interlaced. Human vision, as an integral part of the bionic ensemble, is co-configured by the machinic and the optical parts that it operates through, and is further enskilled to become a proficient sense of touch. And this assertion comes not from applying an abstract theoretical concept such as the cyborg, but out of the analysis of lived experience in the daily integrated human–machine practice of training and operating with-in-as robot.

#### 4 Analytical deliberations: ways of human–machine seeing

Through these three different examples of human–machine ways of seeing, some common traits, as well as some interesting differences, become apparent that may provide further insights about the ways in which the three ensembles and their constituents produce particular forms of human–machine co-sensing. These common traits concern the interlaced process of visual enskilment in human–machine ensembles and the resulting forms of vision that arise, the sensory entanglements and, zooming out, what can be said about the nature of the sensing ensembles.

In the surgical cyborg ensemble, as well as in the ABC and luggage-scanning ensembles, not only are the senses interlaced, but machine and operator skills and capacities mutually configure one another. The cyborg figures of flesh-and-mechanics operate in human–material mutuality, but that does not necessarily imply a symmetry (Suchman 2007). Indeed, the machines are designed and configured by humans, often in ways that mimic human ways. But they impose their machinic particularities on the operators sitting in and in many respects melting together with them.

#### 4.1 Tele-visions

The taskscapes described in the examples all involve seeing, sensing and identifying particular objects: photos and faces, threats and densities, crevices and organs. In all the examples there is a distance between observer and observed, and some kind of tele-vision involved, but it is not necessarily a distanced gaze. We are often up close, even touching the scrutinized objects with our eyes. It is an interactive embodied vision, even in those cases, where it is carried out for control purposes.

Zooming out, however, other types and agencies of tele-vision appear. Although the robot does not “see” in the sense of ongoing image analysis that it acts on, as the ABC and the X-ray scanners on Level 1 do, the videos of the operations are being recorded and are continually accessible to an Intuitive support team in Switzerland. If the surgeon has logged into the console, the video and data relating to each operation are linked to her or his profile and can subsequently be reviewed. Thus in this setting there is also a type of “machine seeing” going on where user performances are being monitored, just as flow, operator efficiency and alertness on the borders and in security control are being monitored by external instances.

The examples above thus highlight how the notion of a visually enskilled community of practice must necessarily be broadened to comprise not only the technologies humans work with and that do their part of the seeing, but also the political, organisational and economic forces that surround and partake in defining the tasks and necessary technologies and the people who develop, install and maintain the machines, to mention only a few. The sun hitting the eye of the ABC camera also modifies the process of co-seeing, as does a momentary glitch in the Interpol connection or the rupture of a small artery blinding the tiny endoscopic camera eye. Such incidents and factors also have their role in the processes of seeing and defining what can be seen.

From that perspective, ways of machine seeing necessarily take into account all the various agencies and factors that go into seeing, besides the machines and their material particularities. They do not simply oblige users to see particular objects in particular ways. Furthermore, the technologies may have been tailored for specific purposes, but those purposes often slip, the technologies come to be used in other ways, and they malfunction in particular settings or for particular purposes. They have deficiencies, they cannot feel, they mistake a printed face for a person. And they necessarily have to work together both with complex and interlaced sensory apparatuses of the humans who are engaging with them and with a variety of contrasting agendas, like those that exist, for example, between employers and the employed, and private and public interests. Finally, passengers seeking passage and patients being operated on

also partake in the processes of seeing by complying more or less efficiently to those processes and contributing their own agendas and bodily particularities (see also Fog et al. 2020). In that sense, it becomes impossible to say how far out the bionic ensemble reaches and where it ends, even if, in the immediate setting, we can identify a central node of co-sensing collaborators.

## 4.2 Tunnel visions

Tunnels bring you from one place to another by eroding the obstacles in your way. But they also seem to lock your attention into to the immediate surroundings and make you focus only on your own movement and the light at the end of the tunnel. In several of the settings described here, different forms of tunnel vision are at work that strongly reduce what can be seen and even create certain forms of blindness. This happens when the police officer in the ABC comes to see in 2D like the machine, when the security officer sees only some objects, whereas others fall out of their field of attention and become invisible and, for the robotic surgeon, when the surroundings disappear from view, because seeing is up so close, focused on the centre of the action. This tunnelling comes out of the technologies involved, as well as out of the organisational, political and temporal factors at play. The resulting partial blindness comes from a temporal and organisationally installed necessity to apply what we could call a selective, ‘protective eye’ that filters away any excess information (Benjamin 1985 in Latham 1999: 464). The more information, the greater the need for protective filtering of what can be seen and sensed. The automation of visual tasks seems to be a common technological answer to this problem, but in the examples above it merely gives the human operator additional and more complex visual tasks.

## 4.3 Automat visions

In the border and airport examples, machines do an important part of the seeing and have their share of decision-making. In the bionic robot, at least in its current version, the surgeon component makes the decisions, strengthened by the stereo-eyes and the dexterity and stabilising effects of the mechanics. But augmented reality, more proficient real-time visualisations, and automation based on artificial intelligence are making their way into also robotic surgery. However, as we see in the examples, higher levels of automation require different modes of seeing and analysis on the part of the human operators involved. Whereas the surgeon’s eyes, by seeing through the stereoscopic eyes, can fully concentrate on the object in focus, the particular organ being seen and touched with the eyes, the border and security guards collaborating with facial recognition

systems and luggage X-ray scanners use a considerable amount of their attention to understand how the machine sees, why it makes the decisions it does and when it needs to be helped or overridden. They deploy a split vision that the surgeon does not, at least for the moment. And they are also attentive to the visual signs of control and monitoring to which their own performance is subject. In that sense, heightened automation may lead to a progressive blurring or even blinding of human vision, with light coming in from too many directions, levels, instances and structures, and where sunglasses are not enough.

## References

- Barad K (2003) Posthumanist performativity: toward an understanding of how matter comes to matter. *Signs J Women Cult Soc* 28(3):801–831. <https://doi.org/10.1086/345321>
- Barad K (2007) Meeting the universe halfway quantum physics and the entanglement of matter and meaning. Duke University Press, Durham
- Berger J (1972) *Ways of seeing*. Penguin, Harmondsworth
- Berger J, Mohr J (1982) *Another way of telling*. Pantheon, New York
- Bethea BT, Okamura AM et al (2004) Application of haptic feedback to robotic surgery. *J Laparoendosc Adv Surg Tech* 14(3):191–195. <https://doi.org/10.1089/1092642041255441>
- Deger J (2006) *Shimmering screens: making media in an aboriginal community*. University of Minnesota Press, Minneapolis
- Foucault M (1972) *Naissance de la clinique*. Presses Universitaires de France, Paris
- Grasseni C (2004) Skilled vision: an apprenticeship in breeding aesthetics. *Soc Anthropol* 12(1):41–55. <https://doi.org/10.1017/S0964028204000035>
- Grasseni C (2007) Introduction. In: Grasseni C (ed) *Skilled visions: between apprenticeship and standards*. Berghahn Books, New York, pp 1–19
- Grasseni C (2011) Skilled visions: toward an ecology of visual inscriptions. In: Banks M, Ruby J (eds) *Made to be seen: perspectives on the history of visual anthropology*. The University of Chicago Press, Chicago, pp 19–44
- Grimshaw A, Ravetz A (2005) *Visualizing anthropology*. Intellect, Bristol
- Grimshaw A, Ravetz A (2009) *Observational cinema: anthropology, film and the observation of social life*. Indiana University Press, Bloomington
- Haraway D (1988) Situated knowledges: the science question in feminism and the privilege of partial perspective. *Fem Stud* 14(3):575–599
- Haraway D (1992) The promises of monsters: a regenerative politics for inappropriate/d others. In: Grossberg L, Nelson C, Treichler PA (eds) *Cultural studies*. Routledge, London, pp 295–337
- Haraway D (2017) *A cyborg manifesto*. In: Wolfe C (ed) *Manifestly haraway*. University of Minnesota Press, Minneapolis, pp 3–90. <https://doi.org/10.2307/3178066>
- Hillman D, Mazzio C (eds) (1997) Introduction. In: *The body in parts: fantasies of corporeality in early modern Europe*. Routledge, London, pp xii–xiii
- Hirsch R (2017) *Seizing the light: a history of photography*. Routledge, London
- Ingold T (2000) *The perception of the environment: essays in livelihood, dwelling and skill*. Routledge, London

- Latham A (1999) The power of distraction: distraction, tactility, and habit in the work of Walter Benjamin. *Environ Plan D Soc Space* 17(4):451–473. <https://doi.org/10.1068/d170451>
- Lave J, Wenger E (1991) *Situated learning: legitimate peripheral participation*. Cambridge University Press, Cambridge
- Lynes K (2016) Cyborgs and virtual bodies. In: Disch L, Hawkesworth M (eds) *The oxford handbook of feminist theory*, pp 1–21. <https://doi.org/10.1093/oxfordhb/9780199328581.013.7>
- MacDougall D (1998) *Transcultural cinema*. Princeton University Press, Princeton
- MacDougall D (2006) *The corporeal image: film, ethnography and the senses*. Princeton University Press, Princeton
- Marks LU (2000) *The skin of the film: intercultural cinema, embodiment, and the senses*. Duke University Press, Durham
- Marks LU (2002) *Touch: sensuous theory and multisensory media*. University of Minnesota Press, Minneapolis
- Mauss M (1973) Techniques of the body. *Econ Soc* 2(1):70–88. <https://doi.org/10.1080/03085147300000003>
- Merleau-Ponty M (2005) *Phenomenology of perception*. Routledge, London
- Møhl P (1997) *Village voices: coexistence and communication in a rural community in Central France*. Museum Tusulanum Press, Copenhagen
- Møhl P (2011) Mise en scène, knowledge and participation: considerations of a filming anthropologist. *Vis Anthropol* 24(3):227–245. <https://doi.org/10.1080/08949468.2010.508707>
- Møhl P (2012) Omens and effect: divergent perspectives on emerillon time, space and existence. Semeion Editions, Meaulne
- Møhl P (2019) Border control and blurred responsibilities at the airport. In: Diphooorn T, Grassiani E (eds) *Security blurs: the politics of plural security provision*. Routledge, London, pp 118–135
- Møhl P (2020a) Vision, faces, identities: technologies of recognition. In: Olwig KF, Grøenberk K, Møhl P, Simonsen A (eds) *The biometric border world: technology, bodies and identities on the move*. Routledge, London, pp 83–99
- Møhl P (2020b) ID-entities, data and the sensory work of border control. *Ethnos J Anthropol*. <https://doi.org/10.1080/00141844.2019.1696858>
- Møhl P, Kristensen NH (2018) At Bruge Billeder og Lyd: Sensorisk Antropologi. In: Bundgaard H, Mogensen H, Rubow C (eds) *Antropologiske Projekter: En Grundbog*. Samfundslitteratur, København, pp 224–244
- Okely J (2001) Visualism and landscape: looking and seeing in Normandy. *Ethnos J Anthropol* 66(1):99–120
- Peirce CS (1998) What is a sign? In: N. In: Houser N, Kloesel C (eds) *The essential peirce: selected philosophical writings*, vol 2. Indiana University Press, Bloomington, pp 1893–1913
- Rangarajan K, Davis H, Pucher PH (2020) Systematic review of virtual haptics in surgical simulation: a valid educational tool? *J Surg Educ* 77(2):337–347. <https://doi.org/10.1016/J.JSURG.2019.09.006>
- Rapiscan (2017) Threat image projection. <https://www.rapiscansystems.com/en/products/rapiscan-threat-image-projection>. Accessed 4 Dec 2017
- Rouch J (1975) The camera and man. In: Hockings P (ed) *Principles of visual anthropology*. Mouton, The Hague & Paris, pp 79–98
- Rouch J (2003) Jaguar. In: Rouch J, Feld S (eds) *Ciné-ethnography*. University of Minnesota Press, Minneapolis, pp 204–209
- Suchman L (2007) *Human-machine reconfigurations: plans and situated actions*. Cambridge University Press, Cambridge

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Ground truth to fake geographies: machine vision and learning in visual practices

Abelardo Gil-Fournier<sup>1</sup> · Jussi Parikka<sup>1,2</sup>

Received: 13 September 2019 / Accepted: 18 August 2020 / Published online: 7 November 2020  
© The Author(s) 2020

## Abstract

This article investigates the concept of the ground truth as both an epistemic and technical figure of knowledge that is central to discussions of machine vision and media techniques of visibility. While ground truth refers to a set of remote sensing practices, it has a longer history in operational photography, such as aerial reconnaissance. Building on a discussion of this history, this article argues that ground truth has shifted from a reference to the physical, geographical ground to the surface of the images echoing earlier points raised by philosopher Jean-Luc Nancy that there is a ground of the image that is central to the task of analysis beyond representational practices. Furthermore, building on the practices of pattern recognition, composite imaging, and different interpretational techniques, we discuss contemporary practices of machine learning that mobilizes geographical earth observation datasets for experimental purposes, including tests such as “fake geography” as well as artistic practices, to show how ground truth is operationalized in such contexts of AI and visual arts.

**Keywords** Remote sensing · Machine vision · Machine learning · Visual culture · Operational image

## 1 Introduction

“Knowing how to discern a groundless image from an image that is nothing but a blow is an entire art in itself”—Jean-Luc Nancy (2005: 25).

Geographical knowledge starts with how we see, or even more accurately, with the production of images through which we see, observe, analyze, and identify. Images are the supportive instrument for understanding territorial formations, and their mediating role is crucial in establishing the seeing that defines geographical entities of knowledge. This can include the most (seemingly) inconspicuous practices, such as coloring maps or populating them with place and site names. It can include the observation of how everyday life is filled with a variety of forms of geographical knowledge embedded in digital platforms for navigation and other purposes. Geographic information systems are the mainstay

of such practices that emerge through the mobilization of data and electronic communication technologies where the physical and the virtual sign entangle (see Pickles 1994). What has been established by decades of critical research is that the relationship between geography and images is heavily overdetermined: the visual and epistemic systems giving a sense of landscape formations are embedded in multiple social, colonial, gendered, and other forms of representational biases (see Rose 1993; Rogoff 2000; Thrift 2008). What’s more, this complex role of images in geographical knowledge has also given rise to various forms of epistemic transfers that are addressed in media theory: maps are understood as media (Siegert 2011) and cities are mediated by maps that themselves are materially situated as part of multiple layers of technologies new and old (Mattern 2017). Building on this work and related to questions of automation, calculation, and AI techniques, we are interested in how analytical and synthetic knowledge about surfaces of the world—landscapes and territories—shifts to knowledge about the surface of images.

This article focuses on the concept of *ground truth* that has both a technical and symbolic meaning in how it negotiates relations between images, material surfaces (geographical, landscape, territorial) and their entangled relations—to echo Michel Foucault’s phrasing—in various institutional

✉ Abelardo Gil-Fournier  
abelardo.gilfourniermartinez@famucz

Jussi Parikka  
j.parikka@soton.ac.uk

<sup>1</sup> FAMU, Prague, Czechia

<sup>2</sup> University of Southampton, Southampton, UK

arrangements of power and knowledge. Starting from ground truth as a grounding figure of knowledge in remote sensing, we build an argument about the synthetic landscapes experimented within current contexts of AI, which we will refer to as “fake geographies” following a term already proposed in computer science research (Xu 2018). Such fabricated and speculative landscapes are intriguing experiments in the creative use of machine learning techniques that deal with, for example, geographical datasets, as they also relate to the figure of the ground truth that is not anymore grounded on Earth, and have become a shifting, ungrounded reference point that has epistemic and aesthetic value. As a contribution to issues of machine vision (as per this special issue), these questions deal with what becomes decipherable as an image and as a landscape in systems that function primarily through data as their input. In short, we are interested in analyzing the shift where the notion of ground truth is no longer specific to the surface of the ground as a geological or geographic reference point. Instead, ground truth becomes read through the “ground” of the synthetic AI images themselves and how datasets are mobilized in machine learning techniques for visual ends.

Jean-Luc Nancy’s (2005) *The Ground of the Image* proposes a similar shift that troubles a rigid distinction between the figure and the ground, the ground and its representation. For Nancy, the image already contains a ground. Even if Nancy’s focus is on classical art history and the philosophy of images that stems from Western art, it becomes a useful reference point for considering the image itself as containing the material ground imprinted onto it but also what is being cut into existence by the image. Nancy writes:

The image does not stand before the ground like a net or a screen. We do not sink; rather, the ground rises to us in the image. The double separation of the image, its pulling away and its cutting out, form both a protection against the ground and an opening onto it. In reality, the ground is not distinct as ground except in the image: without the image, there would only be indistinct adherence. More precisely: in the image, the ground is distinguished by being doubled (Nancy 2005: 13).

This doubling is both philosophical and technical, as our article aims to show. This argument relates essentially to a range of contemporary data and computational techniques and shifts as part of the broader framework of how we engage with questions of AI—such as different machine learning techniques—as part of operational images: images that do not primarily represent but operate in scientific, military, and other technical systems and institutions (see Farocki 2004).

This article is structured around three key points: first, to address the scope of the notion of ground truth, we map the shift from the truth of the ground to pattern recognition as a significant transformation that also relates to questions of machine vision and machine learning techniques (even if we are not able to go into technical specifics in this article). Second, we show how, from the recognition of patterns, we can move to the building of datasets as a relevant part of the infrastructures of ground truth and machine vision. Third, we look at examples of synthetic geographies as experiments that help to understand the ensemble of images in which the ground becomes synthesized with meaningful aesthetic and epistemological consequences.

At stake in our discussion is the claim that ground truth is read from a mass of images, instead of comparatively off the ground. This leads to the question of how these concepts and terms relate to the contemporary situation of machine vision and, more specifically, machine learning and the production of synthetic landscapes, as we argue at the end of the article.

## 2 Evidential paradigm: from truth on ground to pattern recognition

Ground truth, as used in geography and environmental sciences, designates the information provided by direct observation—usually at the level of the literal ground, the surface of the Earth—in relation to maps, models, and remote sensing technologies. It is a concept that operates by recognizing a distinction between different sources of data, which, by comparison, can be brought to verify certain features of geography. Ground truths emerge on location; they are local, specific, and situated so as to be able to offer a grounding for the network of technologies of sense and location.

Ground truth is premised on calibration as a central feature of remote sensing that includes how distances can be negotiated and become standardized against a set of features that are assumed to stay regular. Hence, ground truth is itself constantly situated in a set of dynamic processes, which can be argued to be also about forms of social and economic power as some discourses in critical geography pointed out in the 1990s (Pickles 1994). Already in this phase of earlier research, it was noted that the computerized environments of geographical systems might fundamentally change the nature of ground truth, where the mediations are becoming distanced from the actual ground as a material and lived environment: “The computer promotes a remote, detached view of the world

as seen through the filter of the computer database. Intimate knowledge of the world recedes into the background of ‘ground truth’ as the computer screen becomes the medium through which the geographer interacts with the world” (Veregin 1994: 100–101).

However, besides the discussions about the computerization and mediatization of material landscapes, the notion of ground truth has found a new milieu in practices linked to machine learning, where it has become part of the standard vocabulary. In these contexts, ground truth refers to the data provided as sample output values to the training and testing phases. That is, it is the set of given outcomes—obtained by any means—used to build the model during the so-called learning process. Ground truth does not distinguish between sources of data but refers to a distinction between the outcomes produced by a model and the data provided as expected values to be compared with. Hence, this incorporation of the ground into the machine learning operations of producing models becomes a central part—of not only this specialist practice—of how we understand image operations in AI culture: as an interplay of images, data, and material environments.

These two different contexts of use of ground truth unveil a significant transition in the role played by images as part of the monitoring complex of remote sensing of the Earth. In Earth Observation and other instrument and discursive practices, the surface becomes a site of grounding particular truths (see, e.g., Bishop 2011). As we will show next, these truths highlight how the image complex has evolved and displaced the Earth as an object of knowledge to Earth-data and datasets of images that constitute the primary reference point.<sup>1</sup>

Already as a linguistic term, ground truth can be recognized to contain an oxymoronic heterogeneity. While references to truth locate the concept in the seemingly immaterial space of epistemological values, the ground part of the concept alludes to a presumed tangible substrate of firm evidence included in its multiple uses across different philosophical discourses too. As Caren Kaplan observes, “‘Ground truth’ anchors contemporary preconceptions about physical geography to the comforting solid matter of the earth’s crust” (Kaplan 2018: 34). The idea of witnessing and proximity is closely related to this epistemological trope of ground truth, which thus resonates with Kaplan’s note about the implied solidity of truth, like a permanent and stable geological formation.

But while the rhetoric of epistemological positions is here already heavily laden with different layers, it is tempting to further interrogate how the solidity of the earth’s crust can

be related to the practices of ground truth. Our approach comes from visual studies and critical artistic studies of AI culture while also drawing on material that has investigated questions of surfaces in both screen culture and in architectural and environmental approaches to territories. This means to account for the ways material surfaces become not only the object of systems of perception and analysis—like in aerial photography—but how they become its core element, as the ground truth is subsumed in their operative logic: not only analytical observing but synthetic creation of images from images.

In this regard, it is important to recall that while the notion of ground truth seems to deal with a sort of unmediated set of facts that emerged directly from the earth, the idea of a pure observation has been extensively contested in the domain of science and technology studies (STS). Indeed, as is convincingly shown in many studies and contexts, observations are always theory-dependent and part of a more detailed back and forth movement of comparison and synthesis in contexts of the materiality of epistemic practices (Knorr-Cetina and Mulkay 1983; Hacking 1983; Latour and Woolgar 1986). Similarly, models are embedded in remote sensing from the first point of contact, so to speak, which makes it impossible to maintain a binary between a (data) model and data (capture) through sensors. As Edwards (2013: xiii) puts it: “Today, no collection of signals or observations—even from satellites, which can ‘see’ the whole planet—becomes global in time and space without first passing through a series of data models”. To name an example: in environmental monitoring, the ground truth of measuring instruments in meteorological stations cannot be separated from the weather forecast modeling practices they are part of. Similarly, seismic data is collected as ground truth as it feeds the prediction of seismic movements.

In such domains, ground truth becomes an operation of validating and adjusting a model to a set of facts measured on the ground. While this intervention of the context of the research in the practices of collecting observations has been extensively acknowledged as mentioned above, we would like to emphasize how the notion of ground truth involves an epistemic realm linked to what Carlo Ginzburg (2013) named the “evidential paradigm” which emerged towards the end of the nineteenth century. In other words, ground truth has its own epistemic history as a figure of knowledge.

In the evidential paradigm, the observer—presented under the persona of the detective—is, on the one hand, able to identify a layer of clues and patterns on top of the undifferentiated roughness of matter, and on the other, to produce a plausible reconstruction of events taking place. In this domain, the analytical and the synthetic coalesce. While emerging in a different context than that of visual epistemology or remote sensing, let alone the synthetic technologies of “fake landscapes” in AI techniques that we will turn to

<sup>1</sup> See the notion of data ensemble in Hoelzl and Marie (2014) or the one of image ensembles in Mackenzie and Munster (2019).



later, this reference to Ginzburg helps to draw attention to a common trait that characterizes practices linked to ground truth: the idea that ground speaks through clues, signs, and evidence that a careful observation of the ground as a register is able to distinguish. Among these are also forensic practices that are again part of the contemporary landscape of technical analysis of surfaces (Weizman 2017) and the mediated practices of witnessing (Schuppli 2020), which both seem to carry forward the evidential paradigm even if in more (technical) media-specific ways.

Ground truth applications are found in disciplines such as archeology, paleontology, and forensic research. In such knowledge practices emerging from remote sensing, ground truth is evoked in various contexts: settlement evidence is checked against their images taken from the air (St. Joseph 1945), mass graves are unearthed in relation to the indexical appearance of certain species of plants (Cox, Flavel and Hanson 2008), and agricultural sites are correlated to the detection of phytoliths in soil probes under the microscope (Lombardo et al. 2020). Ground truth is actually a broader term for knowledge verification and calibration that circulates in diverse contexts and practices. In other words, ground truth surfaces as an operation where sets of material traces are distinguished as registers of information. If the detective recognizes and operationalizes the material arrangement of a footprint, the dust on a shoe or the ash of a cigarette as clues indicative of a potential event, in a similar way ground truth encapsulates a set of filtered objects to be mobilized as data.

Ground truth relates closely to Schuppli's (2020) operative concept of *material witness*. This epistemological tool acknowledges not only the evidential role of matter as an active register of external events but also the intervention of explicit acts of scrutiny in the reading of imprints. Here, the "truth" of these "grounds" relies on the application of a series of techniques of demarcation, filtering, and observation, that is, a domain of practitioners and a material culture "that enable such matter to bear witness" (Schuppli 2020: 3). The forensic method of reading material culture acknowledges that ground alone tells no message; hence we are dealing with "impure matter" (Schuppli 2012) affected by the acts of looking. Furthermore, in such contexts of ground truth being established by comparison and other methods, questions of analysis pick up another take on the evidential paradigm as it becomes involved in advanced computational techniques including machine learning—and how it stems from pattern recognition.

Not by chance, since the early 2000s, an educational game by NASA—"Where on Earth...?"—has been inviting players to become "geographical detectives" (California Institute of Technology 2019). The game consists of quizzes where users are asked to locate the geographical area shown on a satellite image by using their abilities to extract visual

clues from the image to recognize the place. The archive of quizzes displays islands, mountain ranges, deltas, volcanos, and other geographical landmarks, pictured from above and presented without any revealing textual key. This case would seem anecdotal if it were not for a recent machine learning project that aims to do something similar on Google's platform, the PlaNet neuronal network (Weyand, Kostrikov, and Philbin 2016). The project aims to build a machine learning model with the ability to determine the location of a photograph by looking at its pixels, that is, without accessing any image metadata such as GPS information.

When comparing NASA's educational game to Google's PlaNet, the task's similarity highlights the main difference: the detective player has been replaced by an algorithmic, big-data-driven process. Beyond the characteristic automatization of pattern recognition in machine learning systems (Mackenzie 2017), we want to address an additional noticeable difference between the two examples. What is interesting here is an assumption underlying the context of the PlaNet project: that any image is supposed to contain in itself enough visual clues and patterns for a sufficiently trained AI model to be able to recognize the place on Earth where it was taken. While for the player in NASA's game, the knowledge of geography linked to her pattern recognition skills is enough to complete the task, in the machine learning project the computational model is supposed to be able to identify the place shown in a photograph—once it has been trained with a large enough dataset of all sorts of outdoor images that are labeled with their geolocation. The physical immutability of the ground in geographical knowledge is replaced by a machine-readable statistical correlation to be found among images when comparing one to another. This is what Adrian Mackenzie and Anna Munster name as the "invisuality" of "platform seeing": "Collections of images operate within and help form a field of distributed invisuality in which relations between images count more than any indexicality or iconicity of an image" (2019: 16). That is, the ground for geographical detectives is replaced by an invisible ground of relations amidst the dataset in the context of statistical learning.

### 3 Photomosaics and stitching ground truths

Although the term "ground truth" was not used widely until the 1960s,<sup>2</sup> the topographical techniques of "ground truthing"—that is, synchronizing images and maps as part of epistemic procedures of verification and calibration—as well

<sup>2</sup> A Google Ngram search showing the use of several wordings for ground truth displays the growing popularity of the term since the 1960s. For more on n-grams see Michel et al. (2011).

as their relation to matters of triangulation as in telemetry, were deployed shortly after the invention of photography. Only ten years after Daguerre's invention, French Army officer Aimé Laussedat produced the first aerial surveys with balloons; five years later, French photographer Nadar filed for a patent on the use of overlapping photos in these surveys (Cosgrove and Fox 2010: 24). Besides the early photography context that is interesting as part of the history of photogrammetry, we want to emphasize the centrality of the early twentieth century and the First World War as far as the operationalization of images about landscapes is concerned. As shown in detail by Saint-Amour (2003, 2011, 2014), after the development of airplanes, the military contexts on photography become instrumental in the image-map complex that shifted the ground of ground truthing.

Besides trained personnel with the fine-tuned capacity to read terrains and images, new technologies supported the task of image comparison and synthetic knowledge. On the one hand, interpreters were considered to be a “highly trained interpretive elite,” often compared to detectives (2003: 356) as they had learned to extract as many visual clues as possible from single aerial photographs. Here the link to the evidential paradigm persists clearly. On the other hand, “a complex technological matrix” (2003: 354) was set up to help in the execution of this task. This matrix included technologies such as the stereoscope, used with pairs of aerial images, the hyperstereoscope, an improved version of the latter relying on the constant speed of planes, which used two pictures separated by a known temporal gap (2003: 360–361), and a specific adaptation of the body and the perceptive skills of the interpreters needed to operate these techniques. With the aid of this reconnaissance matrix—the “deadliest weapon in the war” (2003: 357)—armies were able to distinguish features on the images related to elevation, the third dimension of landscapes, such as differentiating trenches from embankments, as well as seeing even what was hidden underneath bridges and forests (2003: 358). Thus, while the aerial image provides a way of transforming landscapes into readable surfaces (Scott 1999), examples prove that when the dimensions of landscape exceeded the representational and encoding capabilities of isolated pictures, a set of operations involving the use of multiple images simultaneously had to be put into practice.

In addition to making elevations visible—in a way, interpreters accessed “a three-dimensional scale model” of them (Saint-Amour 2003: 358)—other technologies helped to produce pictures of areas of a large-scale that would have otherwise been impossible to portrait in a single shot, such as the areas occupied by trenches. In these cases, different images were stitched together to build a large photomosaic. As Saint-Amour has shown (2014), technicians relied on the appearance of several recognizable objects in the images. As if they were ground truthing the images, they identified

these features as reference objects and used them as anchor points when stitching them together. In other words, techniques of ground truthing had two functions in photomosaics: referential objects were usually tracked to compare and link a particular image to the map of the ground, but here they were also employed to connect images to each other. That is, the ground truthing techniques used to keep images linked as maps that are useful for navigating and reading the surface of the earth also operated to keep images linked to each other. This is particularly relevant in this article's scope, as it shows how the same techniques were used in two different operations. The same techniques used to verify the correlation between data (images) and ground were also in operation to keep hold of the domain of images itself, that is, to keep images connected, not only to the surface of the world but to each other as well. It is these operations of comparison, synthesis, synchronization, calibration that define the scope of ground truth as it emerges as a media technique even before contemporary versions of machine vision and machine learning.

The example of the photomosaic shows how techniques involved in the concept of ground truth were also central to enabling the up-scaling of photographic images by stitching them together. Ground truthing becomes—in this case—relational, traveling from the stabilization of the image in relation to the map, onward to the interweaving of images while keeping them as legitimate geographical tools. The epistemic dimension of what is verifiably “there” is managed through the media techniques mentioned above. Interestingly, this is a relational dimension of the concept of ground truth that has been highlighted elsewhere. Writing about the concept of ground truth as used in the domain of contemporary planetary remote sensing, Jennifer Gabrys has observed how “the ground of ground truth is not, however, the final point of resolution in these sensor environments. Instead, it is a reminder of the constant need to draw connections across phenomena. Ground here is connection and concretization” (Gabrys 2016: 71), which points to the similar traits we have put forward through the examples above.

Continuing on, we move to the question of media techniques of ground truth by discussing another influential project of the past decade, leading us to consider how aerial images are integrated into complex data-synthesizing environments that fluctuate between the visual and invisible. Following Mackenzie and Munster (2019), we argue that this relates to how data is being prepared to be platform-ready and that visual data operates in invisible ways. This relates to the processes of synchronization of visual data, made operative for navigational and other purposes.

#### 4 Environments of images: Google Ground truth

Continuing the recycling of existing terms in data-driven platform contexts, Ground Truth was also the name of one of Google's core projects of the 2010s. First publicly described in *The Atlantic* (Madrigal 2012), the strategic relevance of the availability of detailed and accurate GIS systems in the context of the emergence of all-encompassing digital platforms (Gillespie 2010) fuelled development initiatives. These Google projects aimed to extract data from images at a massive scale, such as the reCAPTCHA project (Strauß 2018: 11). Project Ground Truth focused on creating “accurate and comprehensive map data, by conflating multiple inputs, via algorithms and elbow grease” (Lookingbill and Weiss-Malik 2013). Alongside aerial images, creating access to other combinable inputs—not least of which were Street View cars—was instrumental in offering the epistemologically significant synthesis operating at the back of the map services. The Street View images featured three particularly valuable characteristics: they were regularly updated, accurately geolocated, and displayed map-related information such as traffic signs, street names, and brands' logotypes, among others. The Ground Truth Project has been responsible for the developments geared at reading—as in Optical Character Recognition (OCR)—the information printed in the physical world, pictured afterward on Street View images. Notably, these developments involved much more than software engineering, functioning on the coordinated extraction of vast amounts of hours of human cognitive labor—typical of contemporary AI projects (Crawford and Joler 2018; Ganesh 2020; Joler and Pasquinelli 2020). However, the relevance of the Ground Truth Project in relation to this article is not who or what is in charge of recognizing patterns, but instead, where this activity is performed. Significant for our argument about the machine learning version of ground truth is that the surface of images is the key holder of information. The remarkable aspect of this case is precisely the circulation where data from images in datasets is transferred to the images used as geographical maps. The evidence—the clues and signs—are extracted from the surfaces of the images and projected onto the tiled images that make up the map service.

Furthermore, this circulation presents a version of how the stitching of images mentioned earlier persists as a key trait of image analysis and machine vision from aerial images to a multitude of data-points where the ground becomes not a ground but a shifting set of techniques in which the ground is constantly established and calibrated. In this regard, drawing on Nancy and other sources, Ryan Bishop (2011: 276) argues that “[a]erial visual technologies and aesthetics are almost sole grounded, literally, in the

terrestrial”. The inverted is also true when we consider the role of media that establishes the ground as an epistemologically existing composite: the terrestrial is grounded in the aerial (technologies) and, broadly speaking, in the circulation of images. Aerial photography did, as a matter of fact, persist as a key reference point and infrastructural anchor for the development of remote sensing techniques. Early research on remote sensing in the 1960s shows how the need for geolocating the readings of sensors carried by surveying aircraft—such as spectrometers or radiometers—involved the same hardware as used in aerial photography. Before the availability of satellite-based positioning systems such as GPS, aerial images were the means to geolocate readings of in-flight sensors, thus connecting “the recorded sensor signals to the ground truth visible in or derived from aerial photographs taken in the course of the flight test” (Eppler and Merrill 1969: 665). Aerial photographs were shot simultaneously as the sensor measurements were produced, printing, in some cases, the image of the ground next to the image of the sensor in the same plate (Grossman and Marlatt 1966). Then, be it through the bare ocular inspection of an investigator, a computer-aided “photointerpreter” with a light-pen, or the operations of an automated system, all of the sensor data was printed on top of a map or an aerial image of the surveyed zone (Eppler and Merrill 1969) as if it were a layer of geolocated data in a geographical information system. Not by accident, these were works published simultaneous to Ian McHarg's seminal book, *Design with Nature* (1991), which proposed the layer-cake model, acknowledged as a forerunner of GIS (Steiner and Fleming 2019: 173).

The discussion in geography concerning the mediated “remote, detached view of the world” in computational geographical databases (Veregin 1994: 100–101) resonates again in the context of image circulations that precede GIS systems. The use of aerial photographs as a geolocating tool “projects an image of a de-materializing world” (Virilio 1994: 13), where spatio-temporal coordinates are replaced with the circulation of images. While the reference to dematerialization is a characteristic part of the 1980s–2000s discourse concerning digital technologies, the way images are being understood is nowadays approached with a focus on the materiality of the media techniques that are formative of this image-complex. The ability to picture the ground and the sphere and needle of a measuring instrument was already used in the first aerial photograph surveys, where each shot of the ground included the image of an altimeter and clock placed under the camera. Framed initially by the altimeter and the clock, the aerial image acquired the role of a navigational tool itself, while later becoming replaced by some of the platforms already mentioned. The main point is, however, the focus on the shift from ground to images to data, which in the current context of experimental media arts and experimental use of geographical datasets, becomes

included in the construction of “fake” geographies that we turn to next.

## 5 Ground truth and synthetic (“fake”) geographies

In this article, we have analyzed how the concept of ground truth entails a movement from observations practiced at the ground level to operations at the surface of the image. From images produced on aircraft to Google Street View vehicles, the priority of datasets becomes emphasized as a core feature of ground truthing that is tied closely to environments of images. As such, the primacy of images ties two sets of recent contexts of imaging and ground truth in surprising ways that also reveals something fundamental about such operational images and how they are also mobilized in contemporary experiments that further shift the notion of the ground truth to fictitious, and even extraterrestrial, land surfaces.

In extraterrestrial remote sensing, we are faced with image analysis where the lack of access to the ground means a complete absence of “absolute ground truth” (Smyth et al. 1995: 109). Operations such as the exploration of Venus’ surface through the images taken by the Magellan Mission during the 1990s are peculiarly similar to examples such as Google’s PlaNet. A dataset of human-labeled images of planet surfaces—samples of images of craters and other patterns of landscape filtered by expert observers—is separated from the ensemble of images obtained from the mission and distinguished as ground truth for a statistical learning process aimed at classifying, at a massive scale, the geographic features on the surface of the planet (Smyth et al. 1995). In outer space, the ground of extraterrestrial planets emerges from the techniques embedded in the technological infrastructure of orbiting vehicles. Ground truth is reliant on spacecraft systems, just as earlier cultural techniques were carried by travelers and colonizers on their ships (Siegert 2015).

Complementing the complex and costly procedures of extraterrestrial imaging projects and “comparative planetology” (Likavcan 2019), contemporary experimental media arts projects, including work that elaborates calibration such as Geocinema’s *Framing Territories*,<sup>3</sup> deal with AI methods too; they produce a version of synthetic “fake” landscapes. For instance, in the context of experimental computer science, the *Satellite Image Spoofing project* (Xu and Zhao

2018), proposes a technique aimed at creating fake datasets of satellite images, just as deep fakes produce the illusion of portraying non-existing faces. Deep fake landscapes shift both the focus of machine vision and AI systems from the individual face and demonstrate that any image surface—face, landscape, earth, or extraterrestrial—can be treated in similar ways and subject to similar considerations that push questions of ground truth off the ground. In a way, *Asunder*, the art installation by Tega Brain, Julian Oliver, and Bengt Sjöln (2019), works in this way too. The machine-learning driven simulation of imaginary (future) landscapes are examples of how a fictional AI Environmental Manager not only observes but reorganizes specific locations on Earth based on existing environmental data assembled into (at times absurd) projections. While this work is more about the variety of assumptions of rationality built into climate models and projections, it also works with the “machine vision” of fabricating images of terraforming.

Also within techniques designated “fake geography,” the relation between the aerial view and its intrinsic calculability is explored in generative artworks where images of lands are merged with algorithmic textures, such as in *Neural Landscape Network* by Gregory Chatonsky (2016) or *Invisible Cities* by Gene Kogan (2016). With similar techniques, Shi Weili’s *Terra Mars* mobilized artificial neural networks (conditional GAN in this case) and trained it with “with topographical data and satellite imagery of Earth.” This model was then applied to “see Mars differently,” to make it look like Earth, as one visual commentary on imaginaries of terraforming and, as per the artist’s own words, creative use of AI technologies. In a similar vein, the *Terraformed Mars* twitter bot by the physicist Casey Handmer (2018) offers images of “simulated terraformed Mars landscapes every six hours” that are based on datasets such as the Mars Orbital Laser Altimeter (MOLA) dataset.

Often, such works are discussed in terms of the creative uses of algorithmic techniques and AI. Instead, we want to highlight how the image environments themselves are generative, based on data and details extracted and mobilized from comparative techniques. Instead of merely creative AI, we want to refer to Russian filmmaker Lev Kuleshov’s concept of “creative geography” from the 1920s, a concept that already formulated the ability to build “unique spatial realities [...] out of shots taken in different geographical locations or at different times” (Bozak 2011: 97). This principle of Soviet montage recalls the importance of the relational space opened when ensembles of images are taken together, interweaved in technical operations, and made explicit by Harun Farocki through his practice on the soft-montage (Farocki 2009; Pantenburg 2017). Hence, the synthetic nature of images and landscapes revolves not merely around current versions of machine vision and the creative use of datasets in different AI techniques, but the longer legacy of

<sup>3</sup> On the question of calibration both as a technical and artistic notion, see Geocinema’s work *Framing Territories* (2019) that focuses on remote sensing and science infrastructures of the Digital Belt and Road, and Hito Steyerl’s *How Not to Be Seen: A Fucking Didactic Educational.MOV File* (2013).

how techniques of ground truths afford a synthetic creation of truths that rise “to us in the image,” to return to Nancy’s phrasing. Indeed, Nancy’s (2005) take on the ground being doubled and framed in the images becomes an essential guideline—although we must add a further note that it is not only a doubling but a radical synthetic multiplication of grounds that takes place in images as they are mobilized in massive quantities of datasets.

## 6 Conclusion

In this article, we addressed a set of imaging practices related to the production of geographical knowledge, and we have focused on an analysis of the techniques as they relate to a broader domain of the image. The aim has been to address the shift that contemporary AI culture operationalizes from the surfaces of the world—such as landscapes and territories—to environments of images. This relation is more than representational, and it has, over a longer period, been conditioned by a range of media techniques, and this relates to a shift we have tracked through ground truthing, where the concept of ground truth has been shown to leave the surface of the earth, to be read through the operations and decoding—and synthetic combining—of image surfaces.

Furthermore, and focusing on the relevance of aerial photography and photomosaics, we have shown how the notion of ground truth, despite not being mentioned in literature before the 1950s, has an epistemic history as a figure of knowledge which can be traced back to those photographic techniques linked to the first aerial surveys. We have contextualized these with what Carlo Ginzburg exposed as an evidential paradigm (2013), which can also be described using Adrian Mackenzie’s words while quoting Hannah Arendt in relation to artificial intelligence: “The crux of the problem rests on the ‘treatment’ or operations that ‘reduce terrestrial sensibilities and movements’ to symbols” (Mackenzie 2017: 53). Following approaches that discuss the role of images in the contexts of machine learning (Mackenzie and Munster 2019), we have emphasized the importance of the invisible and non-representational domain of relations between images as elements of the data ensembles involved.

In this regard, ground truth has been shown as the set of techniques where these symbols are related to each other as a media operation that, in addition to grounding, current geographical systems are also able to give rise to what we have addressed as fake or synthetic geographies. Existing critical work in geography has articulated similar claims, such as John Pickles who, writing on GIS and “Benjamin’s law of assembling images” affirmed: “In this sense, as well as legitimizing claims to verisimilitude, digital mapping signals the end of mapping as evidence for anything, or at least the emergence of a representational economy whose

illusions—Baudrillard tells us—will be so powerful that it won’t be possible to tell what is real and what is not” (Pickles 2004: 159). However, ground truth is an operation that goes beyond geography and has in AI techniques and machine vision its main domain of application, as also demonstrated in relation to artistic practices that address such ideas of the ground as a speculative, calculated, hypothesized entity. Thus, broadly speaking, the discussion also concerns contemporary AI-based image cultures in widespread terms, when it comes to technologies and the institutions of the verification of data—of ground truths.

**Acknowledgements** Thank you to Elise Hunchuck for her copyediting and feedback on the draft article and to the special issue editors and reviewers for their feedback. This research has also been supported by Czech Science Foundation funded project 19-26865X. “Operational Images and Visual Culture: Media Archeological Investigations”.

**Funding** The research has been supported by Czech Science Foundation funded project 19-26865X “Operational Images and Visual Culture: Media Archeological Investigations”.

**Code availability** Not applicable.

## Compliance with ethical standard

**Conflicts of interest** Not applicable.

**Availability of data and material** Not applicable.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Bishop R (2011) Project ‘Transparent Earth’ and the Autopsy of Aerial Targeting The Visual Geopolitics of the Underground. *Theory Culture Society* 28:270–286. <https://doi.org/10.1177/0263276411424918>
- Bozak PN (2011) *The Cinematic Footprint: Lights, Camera*. Rutgers University Press, New Brunswick NJ, Natural Resources
- Brain T, Oliver J, and Sjölen B (2019) *Asunder – project website*. <https://asunder.earth/>. Accessed 30 June 2020
- California Institute of Technology (2019) *Where on Earth? Quizzes*. MISR: Jet Propulsion Laboratory. <https://misr.jpl.nasa.gov/quizzes/index.cfm>. Accessed 22 June 2020
- Chatonsky G (2016) *Neural Landscape Network*. Author’s website. <https://chatonsky.net/nln/>. Accessed 22 June 2020

- Crawford K, Joler V (2020) Anatomy of an AI System: The Amazon Echo as an anatomical map of human labor, data and planetary resources. <https://anatomyof.ai/>. Accessed 22 June 2020
- Cosgrove D, Fox WL (2010) *Photography and Flight*. Reaktion Books, London
- Cox M, Flavel A, Hanson I (2008) *The Scientific Investigation of Mass Graves: Towards Protocols and Standard Operating Procedures*. Cambridge University Press, Cambridge
- Edwards PN (2013) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. The MIT Press, Cambridge
- Eppler WG, Merrill RD (1969) Relating remote sensor signals to ground-truth information. *Proc IEEE* 57:665–675. <https://doi.org/10.1109/PROC.1969.7021>
- Farocki H (2004) Phantom Images Public 29:12–22
- Farocki H (2009) Cross Influence / Soft Montage. In: Antje Ehmann, Kodwo Eshun (Eds.), *Harun Farocki. Against What? Against Whom?* Koenig Books, London, pp. 64–79
- Ganesh MI (2020) Intelligence Work. A is for Another: A Dictionary of AI. <https://aisforanother.net/pages/article16.html>. Accessed 22 June 2020
- Gabrys J (2016) *Program Earth: Environmental Sensing Technology and the Making of a Computational Planet*. University of Minnesota Press, Minneapolis
- Geocinema (2019) Framing Territories. Collective's website. <https://geocinema.network/>. Accessed 30 June 2020
- Gillespie T (2010) The politics of 'platforms'. *New Media and Society* 12:347–364. <https://doi.org/10.1177/1461444809342738>
- Ginzburg C (2013) *Clues, Myths, and the Historical Method*. Johns Hopkins University Press, Baltimore
- Grossman RL, Marlatt WE (1966) A method of showing what a radiometer 'sees' during an aircraft survey. *Proc 4th Symp on Remote Sensing of Environment*. University of Michigan, Ann Arbor MI, pp 571–574
- Hacking I (1983) *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press, Cambridge
- Handmer C (2018) Terraformed Mars. Twitter bot. <https://twitter.com/terraformedmars>. Accessed 22 June 2020
- Hoelzl I, Marie R (2014) Google Street View: navigating the operative image. *Visual Studies* 29:261–271. <https://doi.org/10.1080/1472586X.2014.941559>
- Kaplan C (2018) *Aerial Aftermaths: Wartime from Above*. Duke University Press, Durham
- Kogan G (2016) Invisible Cities. Author's website. <https://opendot.github.io/ml4a-invisible-cities/>. Accessed 22 June 2020
- Knorr-Cetina K, Mulkay M (1983) *Science Observed: Perspectives on the Social Study of Science*. SAGE, London
- Latour B, Woolgar S (1986) *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, Princeton
- Likavčan L (2019) *Introduction to Comparative Planetology*. Strelka Press, Moscow
- Lookingbill A, Weiss-Malik M (2013) Google I/O 2013 - Project Ground Truth: Accurate Maps Via Algorithms and Elbow Grease. Google Developers Youtube Channel. <https://www.youtube.com/watch?v=FsbLEtS0uls>. Accessed 22 June 2020
- Lombardo U, Iriarte J, Hilbert L, Ruiz-Pérez J, Capriles JM, Veit H (2020) Early Holocene crop cultivation and landscape modification in Amazonia. *Nature* 581:190–193. <https://doi.org/10.1038/s41586-020-2162-7>
- Mackenzie A (2017) *Machine Learners: Archaeology of a Data Practice*. MIT Press, Cambridge
- MacKenzie A, Munster A (2019) Platform Seeing: Image Ensembles and Their Invisibilities. *Theory Culture Society*. <https://doi.org/10.1177/0263276419847508>
- Madrigal AC (2012) How Google Builds Its Maps—and What It Means for the Future of Everything. *The Atlantic*. <https://www.theatlantic.com/technology/archive/2012/09/how-google-builds-its-maps-and-what-it-means-for-the-future-of-everything/261913/>. Accessed 22 June 2020
- Mattern S (2017) *Code and Clay, Data and Dirt: Five Thousand Years of Urban Media*. University of Minnesota Press, Minneapolis
- McHarg IL (1991) *Design with Nature*. John Wiley & Sons, New York
- Michel J-B, Shen YK, Aiden AP et al (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331:176–182. <https://doi.org/10.1126/science.1199644>
- Nancy J-L (2005) *The Ground of the Image*. Fordham University Press, New York
- Pantenburg V (2017) Working images: Harun Farocki and the operational image. In: Eder J, Klonk C (eds) *Image Operations: Visual Media and Political Conflict*. Manchester University Press, Manchester, pp 49–62
- Pasquinelli M, Joler V (2020) The Nooscope Manifested: AI as Instrument of Knowledge Extractivism. <https://nooscope.ai/>. Accessed 22 June 2020
- Pickles J (ed) (1994) *Ground Truth: The Social Implications of Geographic Information Systems*. The Guilford Press, New York
- Pickles J (2004) *A History of Spaces: Cartographic Reason. Mapping and the Geo-Coded World*, Routledge
- Rogoff I (2000) *Terra Infirma: Geography's Visual Culture*. Routledge, London
- Rose G (1993) *Feminism & Geography: The Limits of Geographical Knowledge*. University of Minnesota Press, Minneapolis
- Saint-Amour PK (2014) Photomosaics: Mapping the Front, Mapping the City. In: Adey P, Whitehead M, Williams A (eds) *From Above. War, Violence and Verticality*. Hurst, London, pp 119–142
- Saint-Amour PK (2011) Applied modernism military and civilian uses of the aerial photomosaic. *Theory Culture Society* 28:241–269. <https://doi.org/10.1177/0263276411423938>
- Saint-Amour PK (2003) Modernist Reconnaissance. *Modernism/modernity* 10(2):349–380. <https://doi.org/10.1353/mod.2003.0047>
- Schuppli S (2012) Impure Matter: A Forensics of WTC Dust. In: Pereira G (ed) *Savage Objects*. Imprensa Nacional Casa da Moeda, Lisbon, pp 120–140
- Schuppli S (2020) *Material Witness: Media, Forensics*. The MIT Press, Cambridge
- Scott JC (1999) *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven
- Siebert B (2011) The map is the territory. *Rad Philos* 169:13–16
- Siebert B (2015) *Cultural Techniques: Grids, Filters, Doors, and Other Articulations of the Real*. Fordham University Press, New York
- Smyth P, Fayyad UM, Burl MC, Perona P, Baldi P (1995) Inferring ground truth from subjective labelling of venus images. In: Tesauro G, Touretzky DS, Leen TK (eds) *Advances in Neural Information Processing Systems 7*. The MIT Press, Cambridge MA, pp 1085–1092
- St. Joseph JK, (1945) Air Photography and Archaeology. *Geograph J* 105:47–59. <https://doi.org/10.2307/1789545>
- Steiner F, Fleming B (2019) Design With Nature at 50: its enduring significance to socio-ecological practice and research in the twenty-first century. *Socio Ecol Pract Res* 1:173–177. <https://doi.org/10.1007/s42532-019-00035-1>
- Steyerl H (2013) How not to be seen: A fucking didactic educational. MOV File
- Strauß S (2018) From big data to deep learning: a leap towards strong AI or 'intelligentia obscura'? *Big Data Cogn Comp*. <https://doi.org/10.3390/bdcc2030016>
- Thrift N (2008) *Non-Representational Theory: Space, Politics, Affect*, Routledge
- Veregin H (1994) *Computer innovation and adoption in geography. A critique of conventional technological*

- Models. In Pickles J (ed) *Ground Truth: The Social Implications of Geographic Information Systems*. The Guilford Press, New York, pp 88–112
- Virilio P (1994) *The Vision Machine*. Indiana University Press, Bloomington
- Weizman E (2017) *Forensic Architecture: Violence at the Threshold of Detectability*. Zone Books, New York
- Weyand T, Kostrikov I, Philbin J (2016) PlaNet - Photo Geolocation with Convolutional Neural Networks. *Lect Notes Comput Sci*. [https://doi.org/10.1007/978-3-319-46484-8\\_3](https://doi.org/10.1007/978-3-319-46484-8_3)
- Xu C (2018) Deep learning and fake geography: creating satellite datasets with Generative Adversarial Networks. AAG Annual Meeting 2018
- Xu C, Zhao B (2018) Satellite image spoofing: creating remote sensing dataset with generative adversarial networks. *GIScience*. <https://doi.org/10.4230/LIPIcs.GISCIENCE.2018.67>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# The Nooscope manifested: AI as instrument of knowledge extractivism

Matteo Pasquinelli<sup>1</sup> · Vladan Joler<sup>2</sup>

Received: 27 March 2020 / Accepted: 14 October 2020 / Published online: 21 November 2020  
© The Author(s) 2020

## Abstract

Some enlightenment regarding the project to mechanise reason. The assembly line of machine learning: data, algorithm, model. The training dataset: the social origins of machine intelligence. The history of AI as the automation of perception. The learning algorithm: compressing the world into a statistical model. All models are wrong, but some are useful. World to vector: the society of classification and prediction bots. Faults of a statistical instrument: the undetection of the new. Adversarial intelligence vs. statistical intelligence: labour in the age of AI.

**Keywords** Nooscope · Political economy · Mechanised knowledge · Information compression · Ethical machine learning

## 1 Some enlightenment regarding the project to mechanise reason

The Nooscope is a cartography of the limits of artificial intelligence, intended as a provocation to both computer science and the humanities. Any map is a partial perspective, a way to provoke debate. Similarly, this map is a manifesto—of AI dissidents. Its main purpose is to challenge the mystifications of artificial intelligence. First, as a technical definition of *intelligence* and, second, as a political form that would be *autonomous* from society and the human.<sup>1</sup> In the expression ‘artificial intelligence’, the adjective ‘artificial’ carries the myth of the technology’s autonomy; it hints to caricatural ‘alien minds’ that self-reproduce in silico but, actually, mystifies two processes of proper alienation; the growing geopolitical autonomy of hi-tech companies and the invisibilization of workers’ autonomy worldwide. The modern project to mechanise human reason has clearly mutated, in the twenty first century, into a corporate regime of knowledge extractivism and epistemic colonialism.<sup>2</sup> This is unsurprising, since machine learning algorithms are the most powerful algorithms for information compression.

The purpose of the Nooscope map is to secularize AI from the ideological status of ‘intelligent machine’ to one of knowledge instruments. Rather than evoking legends of alien cognition, it is more reasonable to consider machine learning as an *instrument of knowledge magnification* that helps to perceive features, patterns, and correlations through vast spaces of data beyond human reach. In the history of science and technology, this is no news; it has already been pursued by optical instruments throughout the histories of astronomy and medicine.<sup>3</sup> In the tradition of science, machine learning is just a *Nooscope*, an instrument to see and navigate the space of knowledge (from the Greek *skopein* ‘to examine, look’ and *noos* ‘knowledge’).

Borrowing the idea from Gottfried Wilhelm Leibniz, the Nooscope diagram applies the analogy of optical media to the structure of all machine learning apparatuses. Discussing the power of his *calculus ratiocinator* and ‘characteristic numbers’ (the idea to design a numerical universal language to codify and solve all the problems of human reasoning), Leibniz made an analogy with instruments of visual magnification such as the microscope and telescope. He wrote: ‘Once the characteristic numbers are established for most concepts, mankind will then possess a new instrument which will enhance the capabilities of the mind to a far greater

✉ Matteo Pasquinelli  
mpasquinelli@hfg-karlsruhe.de

<sup>1</sup> Media Philosophy Department, Karlsruhe University of Arts and Design, Karlsruhe, Germany

<sup>2</sup> New Media Department, Academy of Arts, University of Novi Sad, Novi Sad, Serbia

<sup>1</sup> On the autonomy of technology see: Winner (2001).

<sup>2</sup> For the colonial extensions of the operations of logistics, algorithms and finance see: Mezzadra and Neilson (2019). On the epistemic colonialism of AI see: Pasquinelli (2019b).

<sup>3</sup> Digital humanities term a similar technique *distant reading*, which has gradually involved data analytics and machine learning in literary and art history. See: Moretti (2013).



extent than optical instruments strengthen the eyes, and will supersede the microscope and telescope to the same extent that reason is superior to eyesight' (Leibniz 1677, p. 23). Although the purpose of this text is not to reiterate the opposition between quantitative and qualitative cultures, Leibniz's credo need not be followed. Controversies cannot be conclusively computed. Machine learning is not the ultimate form of intelligence.

Instruments of measurement and perception always come with inbuilt aberrations. In the same way that the lenses of microscopes and telescopes are never perfectly curvilinear and smooth, the *logical lenses* of machine learning embody faults and biases. To understand machine learning and register its impact on society is to study the degree by which social data are diffracted and distorted by these lenses. This is generally known as the debate on bias in AI, but the political implications of the logical form of machine learning are deeper. Machine learning is not bringing a new dark age but one of diffracted rationality, in which, as it will be shown, an episteme of causation is replaced by one of automated correlations. More in general, AI is a new regime of truth, scientific proof, social normativity and rationality, which often does take the shape of a *statistical hallucination*. This diagram manifesto is another way to say that AI, the king of computation (patriarchal fantasy of mechanised knowledge, 'master algorithm' and *alpha machine*) is naked. Here, we are peeping into its black box.

On the invention of metaphors as instrument of knowledge magnification.

Emanuele Tesauro, *Il canocchiale aristotelico* [The Aristotelian Telescope], frontispiece of the 1670 edition, Turin.

## 2 The assembly line of machine learning: data, algorithm, model

The history of AI is a history of experiments, machine failures, academic controversies, epic rivalries around military funding, popularly known as 'winters of AI.'<sup>4</sup> Although corporate AI today describes its power with the language of 'black magic' and 'superhuman cognition', current techniques are still at the experimental stage (Campolo and Crawford 2020). AI is now at the same stage as when the steam engine was invented, before the laws of thermodynamics necessary to explain and control its inner workings, had been discovered. Similarly, today, there are efficient neural networks for image recognition, but there is no *theory of learning* to explain why they work so well and how they fail so badly. Like any invention, the paradigm of machine learning consolidated slowly, in this case through the last half-century. A master algorithm has not appeared overnight. Rather, there has been a gradual construction of a method



<sup>4</sup> For a concise history of AI see: Cardon et al. (2018b).

of computation that still has to find a common language. Manuals of machine learning for students, for instance, do not yet share a common terminology. How to sketch, then, a critical grammar of machine learning that may be concise and accessible, without playing into the paranoid game of defining General Intelligence?

As an instrument of knowledge, machine learning is composed of an object to be observed (*training dataset*), an instrument of observation (*learning algorithm*) and a final representation (*statistical model*). The assemblage of these three elements is proposed here as a spurious and baroque diagram of machine learning, extravagantly termed Nooscope.<sup>5</sup> Staying with the analogy of optical media, the information flow of machine learning is like a light beam that is projected by the training data, compressed by the algorithm and diffracted towards the world by the lens of the statistical model.

The Nooscope diagram aims to illustrate two sides of machine learning at the same time: *how it works and how it fails*—enumerating its main components, as well as the broad spectrum of errors, limitations, approximations, biases, faults, fallacies and vulnerabilities that are native to its paradigm.<sup>6</sup> This double operation stresses that AI is not a monolithic paradigm of rationality but a spurious architecture made of adapting techniques and tricks. Besides, the limits of AI are not simply technical but are imbricated with human bias. In the Nooscope diagram, the essential components of machine learning are represented at the centre, *human biases* and interventions on the left, and *technical biases* and limitations on the right. Optical lenses symbolize biases and approximations representing the compression and distortion of the information flow. The total bias of machine learning is represented by the central lens of the statistical model through which the perception of the world is diffracted.

The limitations of AI are generally perceived today thanks to the discourse on bias—the amplification of gender, race, ability, and class discrimination by algorithms. In machine learning, it is necessary to distinguish between historical bias, dataset bias, and algorithm bias, all of which occur at different stages of the information flow.<sup>7</sup> *Historical bias* (or world bias) is already apparent in society before technological intervention. Nonetheless, the naturalisation of such bias, that is the silent integration of inequality into an apparently neutral technology is by itself harmful (Eubanks 2018).<sup>8</sup>

Paraphrasing Michelle Alexander, Ruha Benjamin has called it the New Jim Code: ‘the employment of new technologies that reflect and reproduce existing inequalities but that are promoted and perceived as more objective or progressive than the discriminatory systems of a previous era’ (Benjamin 2019, p. 5). *Dataset bias* is introduced through the preparation of training data by human operators. The most delicate part of the process is data labelling, in which old and conservative taxonomies can cause a distorted view of the world, misrepresenting social diversities and exacerbating social hierarchies (see below the case of ImageNet).

*Algorithmic bias* (also known as machine bias, statistical bias or model bias, to which the Nooscope diagram gives particular attention) is the further amplification of historical bias and dataset bias by machine learning algorithms. The problem of bias has mostly originated from the fact that machine learning algorithms are among the most efficient for *information compression*, which engenders issues of information resolution, diffraction and loss.<sup>9</sup> Since ancient times, algorithms have been procedures of an economic nature, designed to achieve a result in the shortest number of steps consuming the least amount of resources: space, time, energy and labour (Pasquinelli (forthcoming) *The eye of the master*. Verso, London). The arms race of AI companies is, still today, concerned with finding the simplest and fastest algorithms with which to capitalise data. If information compression produces the maximum rate of profit in corporate AI, from the societal point of view, it produces discrimination and the loss of cultural diversity.

While the social consequences of AI are popularly understood under the issue of bias, the common understanding of technical limitations is known as the *black box* problem. The black box effect is an actual issue of deep neural networks (which filter information so much that their chain of reasoning cannot be reversed) but has become a generic pretext for the opinion that AI systems are not just inscrutable and opaque, but even ‘alien’ and out of control.<sup>10</sup> The black box effect is part of the nature of any experimental machine at the early stage of development (it has already been noticed that the functioning of the steam engine remained a mystery for some time, even after having been successfully tested). The actual problem is the black box rhetoric, which is closely tied to conspiracy theory sentiments in which AI is an occult power that cannot be studied, known, or politically controlled.

<sup>5</sup> The use of the visual analogy is also intended to record the fading distinction between image and logic, representation and inference, in the technical composition of AI. The statistical models of machine learning are operative representations (in the sense of Harun Farocki’s operative images).

<sup>6</sup> For a systematic study of the logical limitations of machine learning see: Malik (2002).

<sup>7</sup> For a more detailed list of AI biases see: Guttag and Suresh (2019) and Galstyan et al. (2019).

<sup>8</sup> See also: Crawford (2017).

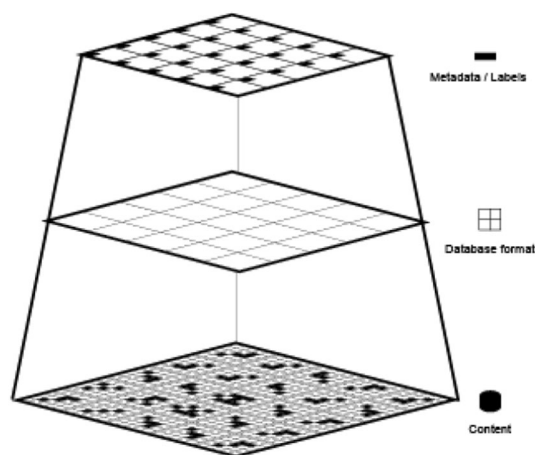
<sup>9</sup> Computer scientists argue that AI belongs to a subfield of *signal processing*, that is *data compression*.

<sup>10</sup> Projects such as Explainable Artificial Intelligence, Interpretable Deep Learning and Heatmapping among others have demonstrated that breaking into the ‘black box’ of machine learning is possible. Nevertheless, the full interpretability and explicability of machine learning statistical models remains a myth. See: Lipton (2016).

### 3 The training dataset: the social origins of machine intelligence

Mass digitalisation, which expanded with the Internet in the 1990s and escalated with datacentres in the 2000s, has made available vast resources of data that, for the first time in history, are free and unregulated. A regime of *knowledge extractivism* (then known as Big Data) gradually employed efficient algorithms to extract ‘intelligence’ from these open sources of data, mainly for the purpose of predicting consumer behaviours and selling ads. The knowledge economy morphed into a novel form of capitalism, called *cognitive capitalism* and then *surveillance capitalism*, by different authors (Corsani et al. 2004; Zuboff 2019). It was the Internet information overflow, vast datacentres, faster microprocessors and algorithms for data compression that laid the groundwork for the rise of AI monopolies in the twenty first century.

What kind of cultural and technical object is the dataset that constitutes the source of AI? The quality of *training data* is the most important factor affecting the so-called ‘intelligence’ that machine learning algorithms extract. There is an important perspective to take into account, to understand AI as a Noosphere. Data are the first source of value and intelligence. Algorithms are second; they are the machines that compute such value and intelligence into a model. However, training data are never raw, independent and unbiased (they are already themselves ‘algorithmic’) (Gitelman 2013). The carving, formatting and editing of training datasets are a laborious and delicate undertaking, which is probably more significant for the final results than the technical parameters that control the learning algorithm. The act of selecting one data source rather than another is the profound mark of human intervention into the domain of the ‘artificial’ minds.



The training dataset is a *cultural construct*, not just a technical one. It usually comprises input data that are associated with ideal output data, such as pictures with their descriptions, also called labels or metadata.<sup>11</sup> The canonical example would be a museum collection and its archive, in which artworks are organised by metadata such as author, year, medium, etc. The semiotic process of assigning a name or a category to a picture is never impartial; this action leaves another deep human imprint on the final result of machine cognition. A training dataset for machine learning is usually composed through the following steps: (1) production: labour or phenomena that produce information; (2) capture: encoding of information into a data format by an instrument; (3) formatting: organisation of data into a dataset; (4) labelling: in supervised learning, the classification of data into categories (metadata).

Machine intelligence is trained on vast datasets that are accumulated in ways neither technically neutral nor socially impartial. Raw data do not exist, as it is dependent on human labour, personal data, and social behaviours that accrue over long periods, through extended networks and controversial taxonomies.<sup>12</sup> The main training datasets for machine learning (NMIST, ImageNet, Labelled Faces in the Wild, etc.) originated in corporations, universities, and military agencies of the Global North. But taking a more careful look, one discovers a profound division of labour that innervates into the Global South via crowdsourcing platforms that are used to edit and validate data.<sup>13</sup> The parable of the *ImageNet* dataset exemplifies the troubles of many AI datasets. ImageNet is a training dataset for Deep Learning that has become the de facto benchmark for image recognition algorithms: indeed, the Deep Learning revolution started in 2012 when Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton won the annual ImageNet challenge with the convolutional neural network AlexNet.<sup>14</sup> ImageNet was initiated by computer scientist Fei-Fei Li back in 2006.<sup>15</sup> Fei-Fei Li had three intuitions to build a reliable dataset for image recognition. First, to download millions of free images from web services such as Flickr and Google. Second, to adopt the computational taxonomy *WordNet* for image labels.<sup>16</sup>

<sup>11</sup> In supervised learning. Also self-supervised learning maintains forms of human intervention.

<sup>12</sup> On taxonomy as a form of knowledge and power see: Foucault (2005).

<sup>13</sup> Such as Amazon Mechanical Turk, cynically termed ‘artificial intelligence’ by Jeff Bezos. See: Pontin (2007).

<sup>14</sup> Although the convolutional architecture dates back to Yann LeCun’s work in the late 1980s, Deep Learning starts with this paper: Krizhevsky et al. (2017).

<sup>15</sup> For an accessible (yet not very critical) account of the ImageNet development see: Mitchell (2019).

<sup>16</sup> WordNet is ‘a lexical database of semantic relations between words’ which was initiated by George Armitage at Princeton University in 1985. It provides a strict tree-like structure of definitions.

Third, to outsource the work of labelling millions of images via the crowdsourcing platform Amazon Mechanical Turk. At the end of the day (and of the assembly line), anonymous workers from all over the planet were paid few cents per task to label hundreds of pictures per minute according to the WordNet taxonomy: their labour resulted in the engineering of a controversial cultural construct. AI scholars Kate Crawford and artist Trevor Paglen have investigated and disclosed the sedimentation of racist and sexist categories in ImageNet taxonomy: see the legitimization of the category ‘failure, loser, nonstarter, unsuccessful person’ for a hundred arbitrary pictures of people (Crawford and Paglen 2019).

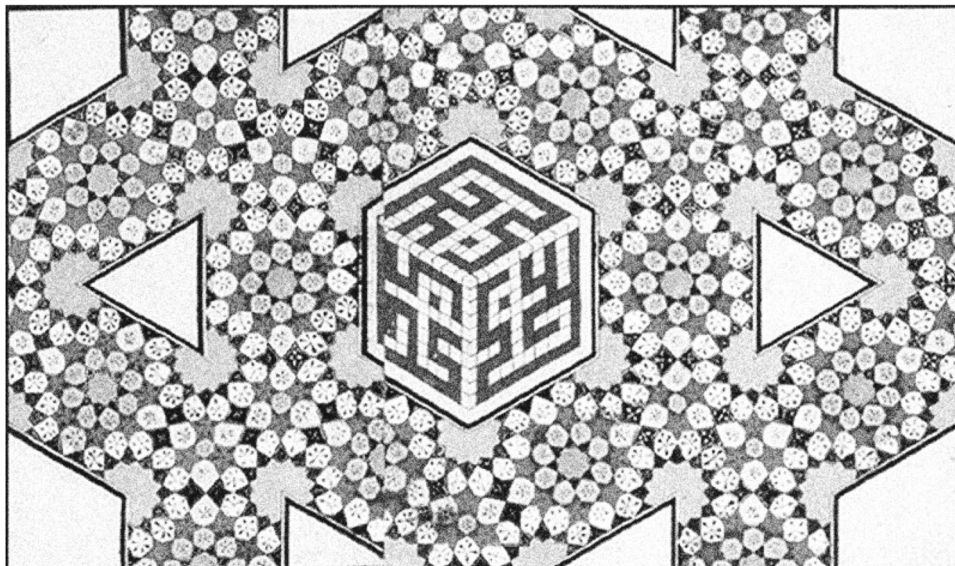
The voracious data extractivism of AI has caused an unforeseeable backlash on digital culture: in the early 2000s, Lawrence Lessig could not predict that the large repository of online images credited by *Creative Commons* licenses would a decade later become an unregulated resource for face recognition surveillance technologies. In similar ways, personal data are continually incorporated without transparency into privatised datasets for machine learning. In 2019 artist and AI researcher, Adam Harvey for the first time disclosed the non-consensual use of personal photos in training datasets for face recognition. Harvey’s disclosure caused Stanford University, Duke University and Microsoft to withdraw their datasets amidst a major *privacy infringement scandal* (Harvey 2019; Murgia 2019). Online training datasets trigger issues of data sovereignty and civil rights that traditional institutions are slow to counteract (see the European General Data Protection Regulation).<sup>17</sup> If 2012 was the year in which the Deep Learning revolution began, 2019 was the year in which its sources were discovered to be vulnerable and corrupted.

Combinatorial patterns and Kufic scripts, Topkapi scroll, ca. 1500, Iran.

#### 4 The history of AI as the automation of perception

The need to demystify AI (at least from the technical point of view) is understood in the corporate world too. Head of Facebook AI and godfather of convolutional neural networks Yann LeCun reiterates that current AI systems are not sophisticated versions of cognition, but rather, of perception. Similarly, the Nooscope diagram exposes the skeleton of the AI black box and shows that AI is not a thinking automaton but an algorithm that performs *pattern recognition*. The notion of pattern recognition contains issues that must be elaborated upon. What is a pattern, by the way? Is a pattern uniquely a visual entity? What does it mean to read social behaviours as patterns? Is pattern recognition an exhaustive definition of intelligence? Most likely not. To clarify these issues, it would be good to undertake a brief archaeology of AI.

The archetype machine for pattern recognition is Frank Rosenblatt’s *Perceptron*. Invented in 1957 at Cornell Aeronautical Laboratory in Buffalo, New York, its name is a shorthand for ‘Perceiving and Recognizing Automaton’ (Rosenblatt 1957). Given a visual matrix of 20×20 photo-receptors, the Perceptron can learn how to recognise simple letters. A visual pattern is recorded as an impression on a network of artificial neurons that are firing up in concert with the repetition of similar images and activating one



<sup>17</sup> The GDPR data privacy regulation that was passed by the European Parliament in May 2018 is, however, an improvement compared to the regulation that is missing in the United States.

single output neuron. The output neuron fires 1 = true, if a given image is recognised, or 0 = false, if a given image is not recognised.

The automation of perception, as a visual montage of pixels along a computational assembly line, was originally implicit McCulloch and Pitts's concept of artificial neural networks (McCulloch and Pitts 1947). Once the algorithm for visual pattern recognition survived the 'winter of AI' and proved efficient in the late 2000s, it was applied also to non-visual datasets, properly inaugurating the age of Deep Learning (the application of pattern recognition techniques to all kinds of data, not just visual). Today, in the case of self-driving cars, the patterns that need to be recognised are objects in road scenarios. In the case of automatic translation, the patterns that need to be recognised are the most common sequences of words across bilingual texts. Regardless of their complexity, from the numerical perspective of machine learning, notions such as image, movement, form, style, and ethical decision can all be described as statistical distributions of pattern. In this sense, pattern recognition has truly become a new *cultural technique* that is used in various fields. For explanatory purposes, the Nooscope is described as a machine that operates on three modalities: *training*, *classification*, and *prediction*. In more intuitive terms, these modalities can be called: pattern extraction, pattern recognition, and pattern generation.

Rosenblatt's Perceptron was the first algorithm that paved the way to machine learning in the contemporary sense. At a time when 'computer science' had not yet been adopted as definition, the field was called 'computational geometry' and specifically 'connectionism' by Rosenblatt himself. The business of these neural networks, however, was to calculate a statistical inference. What a neural network computes is not an exact pattern but the *statistical distribution of a pattern*. Just scraping the surface of the anthropomorphic marketing of AI, one finds another technical and cultural object that needs examination: the *statistical model*. What is the statistical model in machine learning? How is it calculated? What is the relationship between a statistical model and human cognition? These are crucial issues to clarify. In terms of the work of demystification that needs to be done (also to evaporate some naïve questions), it would be good to reformulate the trite question 'Can a machine think?' into the theoretically sounder questions 'Can a statistical model think?', 'Can a statistical model develop consciousness?', et cetera.

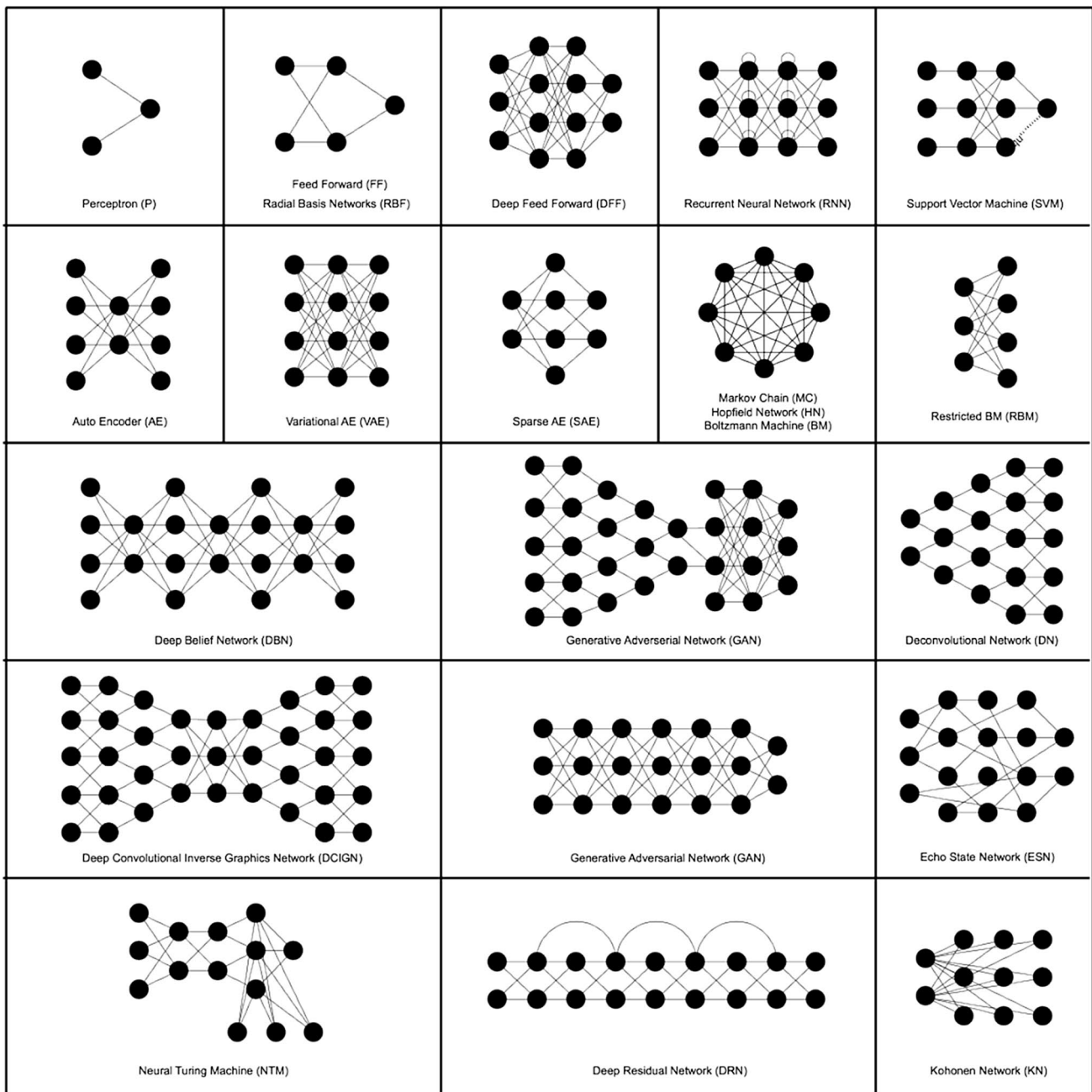
## 5 The learning algorithm: compressing the world into a statistical model

The algorithms of AI are often evoked as alchemic formulas, capable of distilling 'alien' forms of intelligence. But what do the algorithms of machine learning really do? Few people, including the followers of artificial general intelligence (AGI), bother to ask this question. Algorithm is the name of a process, whereby a machine performs a calculation. The product of such machine processes is a statistical model (more accurately termed an 'algorithmic statistical model'). In the developer community, the term 'algorithm' is increasingly replaced with 'model.' This terminological confusion arises from the fact that the statistical model does not exist separately from the algorithm: somehow, the statistical model exists inside the algorithm under the form of distributed memory across its parameters. For the same reason, it is essentially impossible to visualise an algorithmic statistical model, as is done with simple mathematical functions. Still, the challenge is worthwhile.

In machine learning, there are many *algorithm architectures*: simple Perceptron, deep neural network, Support Vector Machine, Bayesian network, Markov chain, autoencoder, Boltzmann machine, etc. Each of these architectures has a different history (often rooted in military agencies and corporations of the Global North). Artificial neural networks started as simple computing structures that evolved into complex ones which are now controlled by a few *hyperparameters* that express millions of *parameters*.<sup>18</sup> For instance, convolutional neural networks are described by a limited set of hyperparameters (number of layers, number of neurons per layer, type of connection, behaviour of neurons, etc.) that project a complex topology of thousands of artificial neurons with millions of parameters in total. The algorithm starts as a blank slate and, during the process called training, or 'learning from data', adjusts its parameters until it reaches a good representation of the input data. In image recognition, as already seen, the computation of millions of parameters has to resolve into a simple binary output: 1 = true, a given image is recognised; or 0 = false, a given image is not recognised.<sup>19</sup>

<sup>18</sup> The parameters of a model that are learnt from data are called 'parameters', while parameters that are not learnt from data and are fixed manually are called 'hyperparameters' (these determine number and properties of the parameters.).

<sup>19</sup> This value can be also a percentage value between 1 and 0.



Source: <https://www.asimovinstitute.org/neural-network-zoo>

Attempting an accessible explanation of the relationship between algorithm and model, let us have a look at the complex Inception v3 algorithm, a deep convolutional neural network for image recognition designed at Google and trained on the ImageNet dataset. Inception v3 is said to have a 78% accuracy in identifying the label of a picture, but the performance of ‘machine intelligence’ in this case can be measured also by the proportion between the size of training data and the trained algorithm (or model). ImageNet contains 14 million images with associated labels

that occupy approximately 150 gigabytes of memory. On the other hand, Inception v3, which is meant to represent the information contained in ImageNet, is only 92 megabytes. The ratio of compression between training data and model partially describes also the rate of information diffraction. A table from the Keras documentation compares these values (numbers of parameters, layer depth, file dimension and accuracy) for the main models of image recognition.<sup>20</sup> This

<sup>20</sup> <https://keras.io/applications> (documentation for individual models.)

is a brutalist but effective way to show the relation between model and data, to show how the ‘intelligence’ of algorithms is measured and assessed in the developer community.

*community*, and today, also beyond.<sup>21</sup> Machine learning models, on the contrary, are opaque and inaccessible to community debate. Given the degree of myth-making and social bias around its mathematical constructs, AI has indeed

### Documentation for individual models

Model	Size	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88 MB	0.790	0.945	22,910,480	126
VGG16	528 MB	0.713	0.901	138,357,544	23
VGG19	549 MB	0.713	0.900	143,667,240	26
ResNet50	98 MB	0.749	0.921	25,636,712	-
ResNet101	171 MB	0.764	0.928	44,707,176	-
ResNet152	232 MB	0.766	0.931	60,419,944	-
ResNet50V2	98 MB	0.760	0.930	25,613,800	-
ResNet101V2	171 MB	0.772	0.938	44,675,560	-
ResNet152V2	232 MB	0.780	0.942	60,380,648	-
InceptionV3	92 MB	0.779	0.937	23,851,784	159
InceptionResNetV2	215 MB	0.803	0.953	55,873,736	572
MobileNet	16 MB	0.704	0.895	4,253,864	88
MobileNetV2	14 MB	0.713	0.901	3,538,984	88
DenseNet121	33 MB	0.750	0.923	8,062,504	121
DenseNet169	57 MB	0.762	0.932	14,307,880	169
DenseNet201	80 MB	0.773	0.936	20,242,984	201
NASNetMobile	23 MB	0.744	0.919	5,326,716	-
NASNetLarge	343 MB	0.825	0.960	88,949,818	-

The top-1 and top-5 accuracy refers to the model's performance on the ImageNet validation dataset.

Depth refers to the topological depth of the network. This includes activation layers, batch normalization layers etc.

Statistical models have always influenced culture and politics. They did not just emerge with machine learning: machine learning is just a new way to automate the technique of statistical modelling. When Greta Thunberg warns ‘Listen to science.’ what she really means, being a good student of mathematics, is ‘Listen to the statistical models of climate science.’ No statistical models, no climate science: no climate science, no climate activism. Climate science is indeed a good example to start with, in order to understand statistical models. Global warming has been calculated by first collecting a vast dataset of temperatures from Earth’s surface each day of the year, and second, by applying a mathematical model that plots the curve of temperature variations in the past and projects the same pattern into the future (Edwards 2010). Climate models are historical artefacts that are tested and debated within the *scientific*

inaugurated the age of *statistical science fiction*. Nooscope is the projector of this large statistical cinema.

## 6 All models are wrong, but some are useful

‘All models are wrong, but some are useful’—the canonical dictum of the British statistician George Box has long encapsulated the logical limitations of statistics and machine learning (Box 1979). This maxim, however, is often used to legitimise the bias of corporate and state AI. Computer scientists argue that human cognition reflects the capacity to abstract and approximate patterns. Therefore, what’s the

<sup>21</sup> See the Community Earth System Model (CESM) that has been developed by the National Center for Atmospheric Research in Boulder, Colorado, since 1996. The Community Earth System Model is a fully coupled numerical simulation of the Earth system consisting of atmospheric, ocean, ice, land surface, carbon cycle, and other components. CESM includes a climate model providing state-of-the-art simulations of the Earth’s past, present, and future.’ <https://www.cesm.ucar.edu>

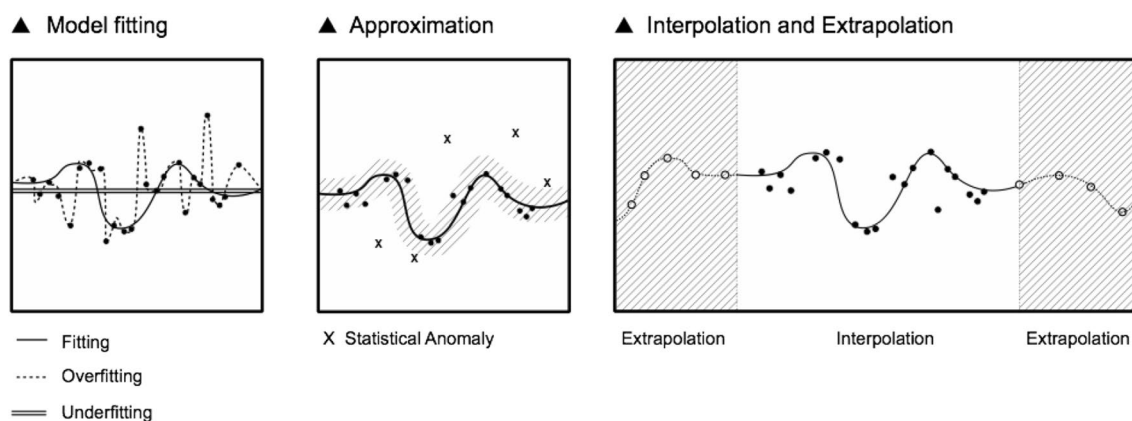
problem with machines being approximate, and doing the same? Within this argument, it is rhetorically repeated that ‘the map is not the territory’. This sounds reasonable. But what should be contested is that AI is a heavily compressed and distorted map of the territory and that this map, like many forms of automation, is not open to community negotiation. AI is a map of the territory without community access and community consent.<sup>22</sup>

How does machine learning plot a statistical map of the world? Let’s face the specific case of image recognition (the basic form of the *labour of perception*, which has been codified and automated as pattern recognition)<sup>23</sup> (Beller 2012). Given an image to be classified, the algorithm detects the edges of an object as the statistical distribution of dark pixels surrounded by light ones (a typical visual pattern). The algorithm does not know what an image is, does not perceive an image as human cognition does, it only computes pixels, numerical values of brightness and proximity. The algorithm is programmed to record only the dark edge of a profile (that is to *fit* that desired pattern) and not all the pixels across the image (that would result in *overfitting* and repeating the whole visual field). A statistical model is said to be trained successfully when it can elegantly *fit* only the important patterns of the training data and apply those patterns also to new data ‘in the wild’. If a model learns the training data too well, it recognises only exact matches of the original patterns and will overlook those with close similarities, ‘in the wild’. In this case, the model is *overfitting*, because it has meticulously learnt everything (including noise) and is not able to distinguish a pattern from its background. On the other hand, the model is *underfitting* when it is not able to detect meaningful patterns from the training data. The notions of data overfitting, fitting and underfitting can be visualised on a Cartesian plane.

The challenge of guarding the accuracy of machine learning lays in calibrating the equilibrium between data underfitting and overfitting, which is difficult to do because of different machine biases. Machine learning is a term that, as much as ‘AI’, anthropomorphizes a piece of technology: machine learning *learns nothing* in the proper sense of the word, as a human does; machine learning simply maps a statistical distribution of numerical values and draws a mathematical function that hopefully approximates human comprehension. That being said, machine learning can, for this reason, cast new light on the ways in which humans comprehend.

The statistical model of machine learning algorithms is also an approximation in the sense that it guesses the missing parts of the data graph: either through *interpolation*, which is the prediction of an output  $y$  within the known interval of the input  $x$  in the training dataset, or through *extrapolation*, which is the prediction of output  $y$  beyond the limits of  $x$ , often with high risks of inaccuracy. This is what ‘intelligence’ means today within machine intelligence: to extrapolate a non-linear function beyond known data boundaries. As Dan McQuillan aptly puts it: ‘There is no intelligence in artificial intelligence, nor does it learn, even though its technical name is machine learning, it is simply mathematical minimization’ (McQuillan 2018a; b).

It is important to recall that the ‘intelligence’ of machine learning is not driven by exact formulas of mathematical analysis, but by algorithms of *brute force approximation*. The shape of the correlation function between input  $x$  and output  $y$  is calculated algorithmically, step by step, through tiresome mechanical processes of gradual adjustment (like gradient descent, for instance) that are equivalent to the differential calculus of Leibniz and Newton. Neural networks

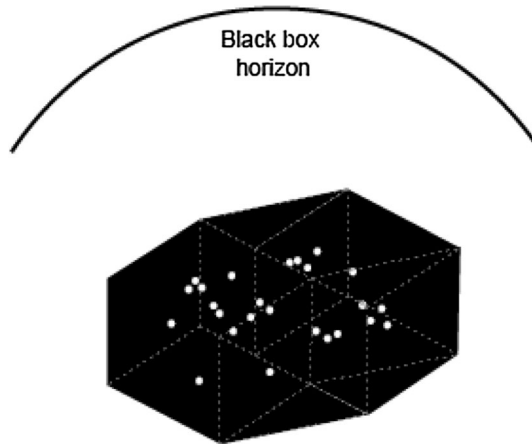


<sup>22</sup> Post-colonial and post-structuralist schools of anthropology and ethnology have stressed that there is never territory per se, but always an act of territorialisation.

<sup>23</sup> Pattern recognition is one among many other economies of attention. ‘To look is to labor’, as Jonathan Beller reminds us.



are said to be among the most efficient algorithms, because these differential methods can *approximate* the shape of any function given enough layers of neurons and abundant computing resources.<sup>24</sup> Brute-force gradual approximation of a function is the core feature of today's AI, and only from this perspective can one understand its potentialities and limitations—particularly, its escalating carbon footprint (the training of deep neural networks requires exorbitant amounts of energy because of gradient descent and similar training algorithms that operate on the basis of continuous infinitesimal adjustments) (Ganesh et al. 2019).



Multidimensional vector space.

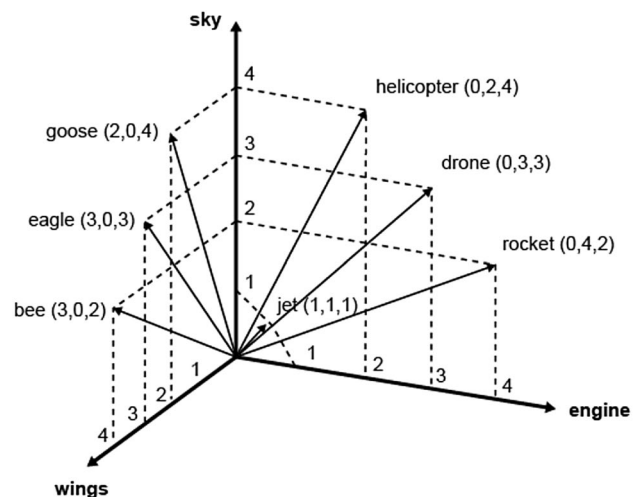
## 7 World to vector

The notions of data fitting, overfitting, underfitting, interpolation and extrapolation can be easily visualised in two dimensions, but statistical models usually operate along multidimensional spaces of data. Before being analysed, data are encoded into a *multi-dimensional vector space* that is far from intuitive. What is a vector space and why is it multi-dimensional? Cardon, Cointet and Mazière describe the vectorialisation of data in this way:

A neural network requires the inputs of the calculator to take on the form of a vector. Therefore, the world must be coded in advance in the form of a purely digital vectorial representation. While certain objects such as images are naturally broken down into vectors, other objects need to be 'embedded' within a vectorial space before it is possible to calculate or classify them with neural networks. This is the case of text, which is the prototypical example. To input a word into a neural network, the *Word2vec* technique 'embeds' it into a vectorial space that measures its distance from the other words in the corpus. Words thus inherit a posi-

tion within a space with several hundreds of dimensions. The advantage of such a representation resides in the numerous operations offered by such a transformation. Two terms whose inferred positions are near one another in this space are equally similar semantically; these representations are said to be distributed: the vector of the concept 'apartment'  $[-0.2, 0.3, -4.2, 5.1 \dots]$  will be similar to that of 'house'  $[-0.2, 0.3, -4.0, 5.1 \dots]$ . [...] While natural language processing was pioneering for 'embedding' words in a vectorial space, today we are witnessing a generalization of the embedding process which is progressively extending to all applications fields: networks are becoming simple points in a vectorial space with *graph2vec*, texts with *paragraph2vec*, films with *movie2vec*, meanings of words with *sens2vec*, molecular structures with *mol2vec*, etc. According to Yann LeCun, the goal of the designers of connectionist machines is to put the world in a vector (*world2vec*) (Cardon et al. 2018a).

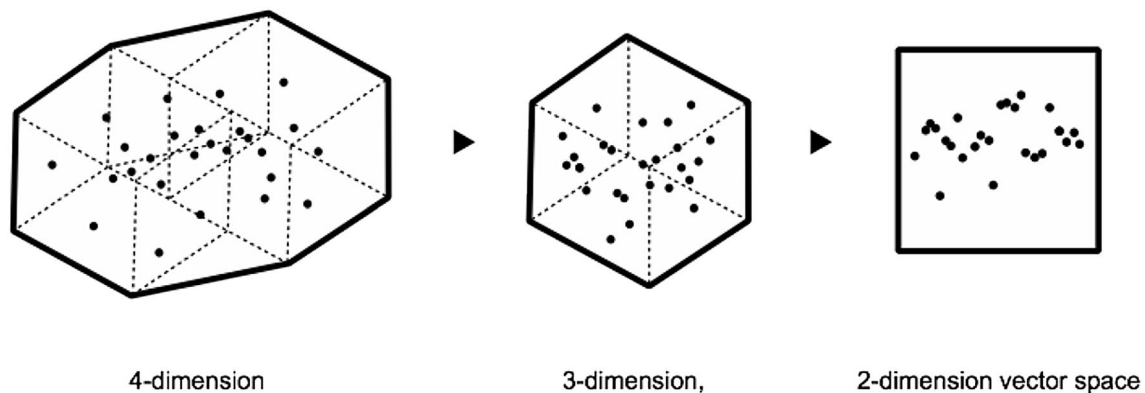
*Multi-dimensional vector space* is another reason why the logic of machine learning is difficult to grasp. Vector space is another new cultural technique, worth becoming familiar with. The field of Digital Humanities, in particular, has been covering the technique of vectorialisation through which our collective knowledge is invisibly rendered and processed. William Gibson's original definition of cyberspace prophesized, most likely, the coming of a vector space rather than virtual reality: 'A graphic representation of data abstracted from the banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding' (Gibson 1984, p. 69).



<sup>24</sup> As proven by the Universal Approximation Theorem.

Vector space of seven words in three contexts.<sup>25</sup>

It must be stressed, however, that machine learning still resembles more craftsmanship than exact mathematics. AI is still a history of hacks and tricks rather than mystical intuitions. For example, one trick of information compression is *dimensionality reduction*, which is used to avoid the Curse of Dimensionality, that is the exponential growth of the variety of features in the vector space. The dimensions of the categories that show low variance in the vector space (i.e. whose values fluctuate only a little) are aggregated to reduce calculation costs. Dimensionality reduction can be used to cluster word meanings (such as in the model word2vec) but can also lead to *category reduction*, which can have an impact on the representation of social diversity. Dimensionality reduction can shrink taxonomies and introduce bias, further normalising world diversity and obliterating unique identities (Samadi et al. 2018).



## 8 The society of classification and prediction bots

Most of the contemporary applications of machine learning can be described according to the two modalities of classification and prediction, which outline the contours of a new society of control and statistical governance. Classification is known as *pattern recognition*, while prediction can be defined also as *pattern generation*. A new pattern is recognised or generated by interrogating the inner core of the statistical model.

Machine learning *classification* is usually employed to recognise a sign, an object, or a human face, and to assign a corresponding category (label) according to taxonomy or cultural convention. An input file (e.g. a headshot captured by a surveillance camera) is run through the model to determine whether it falls within its statistical distribution or not. If so, it is assigned the corresponding output label. Since the times of the Perceptron, classification has been the

originary application of neural networks: with Deep Learning, this technique is found ubiquitously in face recognition classifiers that are deployed by police forces and smartphone manufacturers alike.

Machine learning *prediction* is used to project future trends and behaviours according to past ones, that is to complete a piece of information knowing only a portion of it. In the prediction modality, a small sample of input data (a primer) is used to predict the missing part of the information following once again the statistical distribution of the model (this could be the part of a numerical graph oriented toward the future or the missing part of an image or audio file). Incidentally, other modalities of machine learning exist: the statistical distribution of a model can be dynamically visualised through a technique called latent space exploration and, in some recent design applications, also *pattern exploration*.<sup>26</sup>

Machine learning classification and prediction are becoming ubiquitous techniques that constitute new forms of surveillance and governance. Some apparatuses, such as self-driving vehicles and industrial robots, can be an integration of both modalities. A self-driving vehicle is trained to recognise different objects on the road (people, cars, obstacles, signs) and predict future actions based on decisions that a human driver has taken in similar circumstances. Even if recognising an obstacle on a road seems to be a neutral gesture (it's not), identifying a human being according to categories of gender, race and class (and in the recent COVID-19 pandemic as sick or immune), as state institutions are increasingly doing, is the gesture of a new disciplinary regime. The hubris of automated classification has caused the revival of reactionary Lombrosian techniques that were thought to have been consigned to history, techniques such as automatic gender recognition (AGR), 'a subfield of facial recognition that aims to algorithmically identify the gender of individuals from photographs or videos' (Keyes 2018).

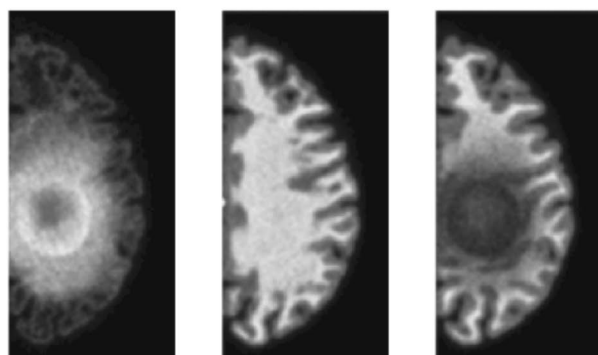
<sup>25</sup> Source: <https://corpling.hypotheses.org/495>.

<sup>26</sup> See the idea of assisted and generative creation in: Pieters and Winiger (2016).

Recently, the generative modality of machine learning has had a cultural impact: its use in the production of visual artefacts has been received by mass media as the idea that artificial intelligence is ‘creative’ and can autonomously make art. An artwork that is said to be created by AI always hides a human operator, who has applied the generative modality of a neural network trained on a specific dataset. In this modality, the neural network is run *backwards* (moving from the smaller output layer toward the larger input layer) to generate new patterns after being trained at classifying them, a process that usually moves from the larger input layer to the smaller output layer. The generative modality, however, has some useful applications; it can be used as a sort of reality check to reveal what the model has learnt, i.e. to show how the model ‘sees the world.’ It can be applied to the model of a self-driving car, for instance, to check how the road scenario is projected.

A famous way to illustrate how a statistical model ‘sees the world’ is Google DeepDream. DeepDream is a convolutional neural network based on Inception (which is trained on the ImageNet dataset mentioned above) that was programmed by Alexander Mordvintsev to project hallucinatory patterns. Mordvintsev had the idea to ‘turn the network upside down’, that is to turn a classifier into a generator, using some random noise or generic landscape images as input (Mordvintsev

The two main modalities of classification and generation can be assembled in further architectures such as in the Generative Adversarial Networks. In the GAN architecture, a neural network with the role of *discriminator* (a traditional classifier) has to recognise an image produced by a neural network with the role of *generator*, in a reinforcement loop that trains the two statistical models simultaneously. For some converging properties of their respective statistical models, GANs have proved very good at generating highly realistic pictures. This ability has prompted their abuse in the fabrication of ‘deep fakes’.<sup>27</sup> Concerning regimes of truth, a similar controversial application is the use of GANs to generate synthetic data in cancer research, in which neural networks trained on unbalanced datasets of cancer tissues have started to hallucinate cancer where there was none (Cohen et al. 2018). In this case ‘instead of discovering things, we are inventing things’, Fabian Offert notices, ‘the space of discovery is identical to the space of knowledge that the GAN has already had.[...] While we think that we are seeing through GAN—looking at something with the help of a GAN—we are actually seeing *into* a GAN. GAN vision is not augmented reality, it is virtual reality. GANs do blur discovery and invention’ (Offert 2020). The GAN simulation of brain cancer is a tragic example of AI-driven scientific hallucination.

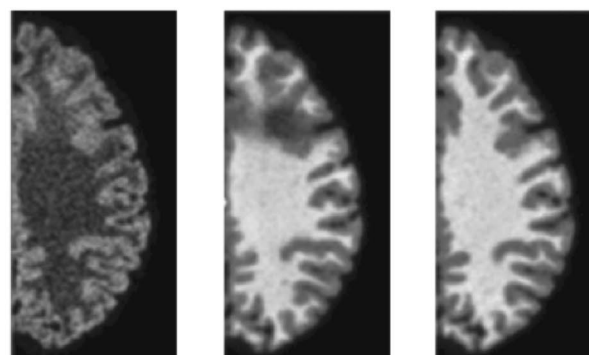


Flair Real

T1 Transformed

T1 Real

A translation removing tumors



Flair Real

T1 Transformed

T1 Real

A translation adding tumors

et al. 2015). He discovered that ‘neural networks that were trained to discriminate between different kinds of images have quite a bit of the information needed to generate images too.’ In DeepDream first experiments, bird feathers and dog eyes started to emerge everywhere as dog breeds and bird species are vastly overrepresented in ImageNet. It was also discovered that the category ‘dumbbell’ was learnt with a surreal human arm always attached to it. Proof that many other categories of ImageNet are misrepresented.

<sup>27</sup> Deep fakes are synthetic media like videos in which a person’s face is replaced with someone else’s facial features, often for the purpose to forge fake news.

Joseph Paul Cohen, Margaux Luck and Sina Honari. ‘Distribution Matching Losses Can Hallucinate Features in Medical Image Translation’, 2018. Courtesy of the authors.

## 9 Faults of a statistical instrument: the undetection of the new

The normative power of AI in the twenty first century has to be scrutinised in these epistemic terms: what does it mean to frame collective knowledge as patterns, and what does it mean to draw vector spaces and statistical distributions of social behaviours? According to Foucault, in early modern France, statistical power was already used to measure social norms, discriminating between normal and abnormal behaviour (Foucault 2004, p. 26). AI easily extends the ‘power of normalisation’ of modern institutions, among others bureaucracy, medicine and statistics (originally, the numerical knowledge possessed by the state about its population) that passes now into the hands of AI corporations. The institutional norm has become a computational one: the classification of the subject, of bodies and behaviours, seems no longer to be an affair for public registers, but instead for algorithms and datacentres.<sup>28</sup> ‘Data-centric rationality’, Paula Duarte has concluded, ‘should be understood as an expression of the coloniality of power’ (Ricourte 2019).

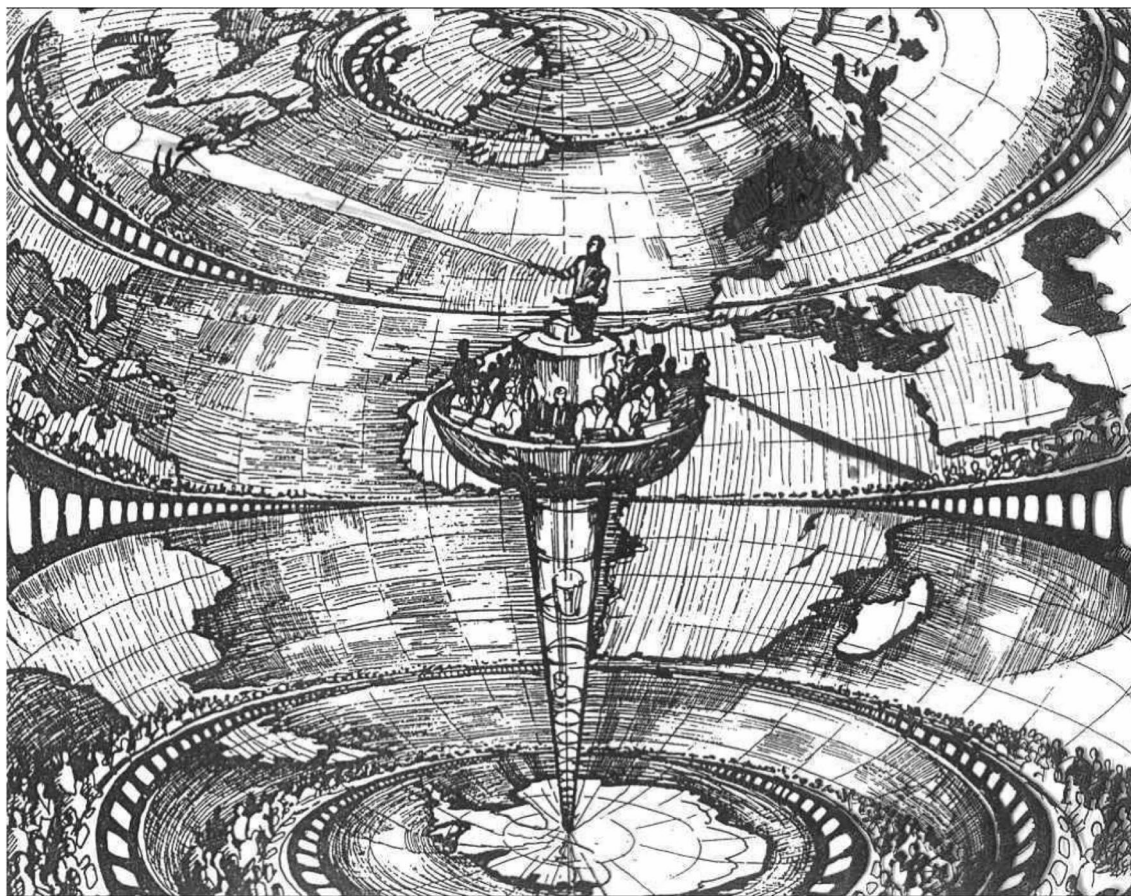
A gap, a friction, a conflict, however, always persists between AI statistical models and the human subject that is supposed to be measured and controlled. This logical gap between AI statistical models and society is usually debated as *bias*. It has been extensively demonstrated how face recognition misrepresents social minorities and how black neighbourhoods, for instance, are bypassed by AI-driven logistics and delivery service (Ingold and Soper 2016). If gender, race and class discriminations are amplified by AI algorithms, this is also part of a larger problem of discrimination and normalisation at the logical core of machine learning. The logical and political limitation of AI is the technology’s difficulty in the *recognition and prediction of a new event*. How is machine learning dealing with a truly unique anomaly, an uncommon social behaviour, an

innovative act of disruption? The two modalities of machine learning display a limitation that is not simply bias.

A logical limit of machine learning classification, or pattern recognition, is the inability to recognise a *unique anomaly* that appears for the first time, such as a new metaphor in poetry, a new joke in everyday conversation, or an unusual obstacle (a pedestrian? a plastic bag?) on the road scenario. The *undetection of the new* (something that has never ‘been seen’ by a model and therefore never classified before in a known category) is a particularly hazardous problem for self-driving cars and one that has already caused fatalities. Machine learning prediction, or pattern generation, show similar faults in the guessing of future trends and behaviours. As a technique of information compression, machine learning automates the dictatorship of the past, of past taxonomies and behavioural patterns, over the present. This problem can be termed the *regeneration of the old*—the application of a homogenous space–time view that restrains the possibility of a new historical event.

Interestingly, in machine learning, the logical definition of a security issue also describes the logical limit of its creative potential. The problems characteristic of the *prediction of the new* are logically related to those that characterise the *generation of the new*, because the way a machine learning algorithm predicts a trend on a time chart is identical to the way it generates a new artwork from learnt patterns. The hackneyed question ‘Can AI be creative?’ should be reformulated in technical terms: is machine learning able to create works that are not imitations of the past? Is machine learning able to extrapolate beyond the stylistic boundaries of its training data? The ‘creativity’ of machine learning is limited to the detection of styles from the training data and then random improvisation within these styles. In other words, machine learning can explore and improvise only within the logical boundaries that are set by the training data. For all these issues, and its degree of information compression, it would be more accurate to term machine learning art as *statistical art*.

<sup>28</sup> On computational norms see: Pasquinelli (2017).



Lewis Fry Richardson, *Weather Prediction by Numerical Process*, Cambridge University Press, 1922.

Another unspoken bug of machine learning is that the statistical correlation between two phenomena is often adopted to explain causation from one to the other. In statistics, it is commonly understood that *correlation does not imply causation*, meaning that a statistical coincidence alone is not sufficient to demonstrate causation. A tragic example can be found in the work of statistician Frederick Hoffman, who in 1896 published a 330-page report for insurance companies to demonstrate a *racial correlation* between being a black American and having short life expectancy (O’Neil 2016). Superficially mining data, machine learning can construct any arbitrary correlation that is then perceived as real. In 2008, this logical fallacy was proudly embraced by Wired director Chris Anderson who declared the ‘end of theory’, because ‘the data deluge makes the scientific method obsolete’ (Anderson 2008).<sup>29</sup> According to Anderson, himself no expert on scientific method and logical inference, statistical

correlation is enough for Google to run its ads business, therefore, it must also be good enough to automatically discover scientific paradigms. Even Judea Pearl, a pioneer of Bayesian networks, believes that machine learning is obsessed with ‘curve fitting’, recording correlations without providing explanations (Mackenzie and Judea 2018). Such a logical fallacy has already become a political one, if one considers that police forces worldwide have adopted predictive policing algorithms.<sup>30</sup> According to Dan McQuillan, when machine learning is applied to society in this way, it turns into a biopolitical apparatus of *preemption*, that produces subjectivities which can subsequently be criminalized (McQuillan 2018a; b). Ultimately, machine learning obsessed with ‘curve fitting’ imposes a *statistical culture* and replaces the traditional episteme of causation (and political accountability) with one of correlations blindly driven by the automation of decision making.

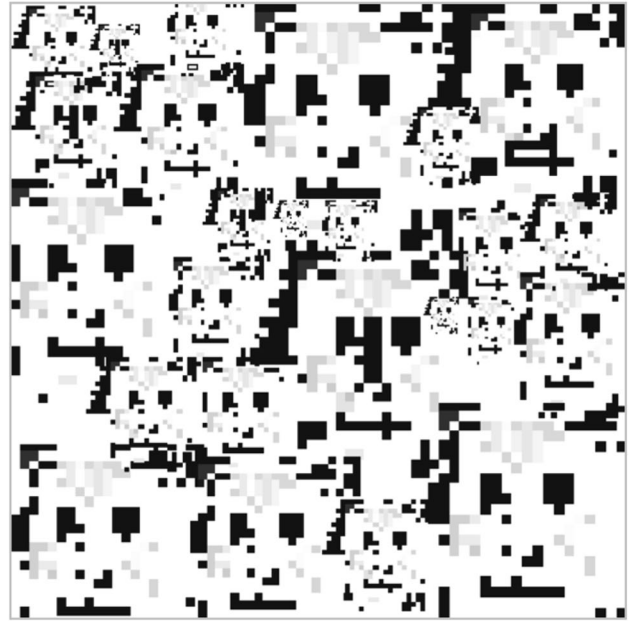
<sup>29</sup> For a critique see: Mazzocchi (2015).

<sup>30</sup> Experiments by the New York Police Department since the late 1980s. See: Pasquinelli, *Arcana Mathematica Imperii*.

## 10 Adversarial intelligence vs. artificial intelligence

So far, the statistical diffractions and hallucinations of machine learning have been followed step by step through the multiple lenses of the Nooscope. At this point, the orientation of the instrument has to be reversed: scientific theories as much as computational devices are inclined to consolidate an abstract perspective—the scientific ‘view from nowhere’, that is often just the point of view of power. The obsessive study of AI can suck the scholar into an abyss of computation and the illusion that the technical form illuminates the social one. As Paola Ricaurte remarks: ‘Data extractivism assumes that everything is a data source’ (Ricaurte 2019). How to emancipate ourselves from a data-centric view of the world? It is time to realise that it is not the statistical model that constructs the subject, but rather the subject that structures the statistical model. Internalist and externalist studies of AI have to blur: subjectivities make the mathematics of control from within, not from without. To second what Guattari once said of machines in general, machine intelligence too is constituted of ‘hyper-developed and hyper-concentrated forms of certain aspects of human subjectivity’ (Guattari 2013, p. 2).

Rather than studying only how technology works, critical inquiry studies also how it breaks, how subjects rebel against its normative control and workers sabotage its gears. In this sense, a way to sound the limits of AI is to look at hacking practices. Hacking is an important method of knowledge production, a crucial epistemic probe into the obscurity of AI.<sup>31</sup> Deep learning systems for face recognition have triggered, for instance, forms of counter-surveillance activism. Through techniques of face obfuscation, humans have decided to become unintelligible to artificial intelligence: that is to become, themselves, *black boxes*. The traditional techniques of *obfuscation* against surveillance immediately acquire a mathematical dimension in the age of machine learning. For example, AI artist and researcher Adam Harvey has invented a camouflage textile called HyperFace that fools computer vision algorithms to see multiple human faces where there is none (Harvey 2016). Harvey’s work provokes the question: what constitutes a face for a human eye, on the one hand, and a computer vision algorithm, on the other? The neural glitches of HyperFace exploit such a cognitive gap and reveal what a human face looks like to a machine. This gap between human and machine perception helps to introduce the growing field of adversarial attacks.



Adam Harvey, HyperFace pattern, 2016.

*Adversarial attacks* exploit blind spots and weak regions in the statistical model of a neural network, usually to fool a classifier and make it perceive something that is not there. In object recognition, an adversarial example can be a doctored image of a turtle, which looks innocuous to a human eye but gets misclassified by a neural network as a rifle (Athalye et al. 2017). Adversarial examples can be realised as 3D objects and even stickers for road signs that can misguide self-driving cars (which may read a speed limit of 120 km/h where it is actually 50 km/h) (Morgulis et al. 2019). Adversarial examples are designed knowing what a machine has never seen before. This effect is achieved also by reverse-engineering the statistical model or by polluting the training dataset. In this latter sense, the technique of *data poisoning* targets the training dataset and introduces doctored data. In doing so, it alters the accuracy of the statistical model and creates a backdoor that can be eventually exploited by an adversarial attack.<sup>32</sup>

Adversarial attack seems to point to a mathematical vulnerability that is common to all machine learning models: ‘An intriguing aspect of adversarial examples is that an example generated for one model is often misclassified by other models, even when they have different architectures or were trained on disjoint training sets’ (Goodfellow et al. 2014). Adversarial attacks remind us of the discrepancy between human and machine perception and that the

<sup>31</sup> The relationship between AI and hacking is not as antagonistic as it may appear: it often resolves in a loop of mutual learning, evaluation and reinforcement.

<sup>32</sup> Data poisoning can also be employed to protect privacy by entering anonymised or random information into the dataset.

logical limit of machine learning is also a political one. The logical and ontological boundary of machine learning is the unruly subject or anomalous event that escapes classification and control. The subject of algorithmic control fires back. Adversarial attacks are a way to sabotage the assembly line of machine learning by inventing a virtual obstacle that can set the control apparatus out of joint. An adversarial example is the *sabot* in the age of AI.

## 11 Labour in the age of AI

The natures of the ‘input’ and ‘output’ of machine learning have to be clarified. AI troubles are not only about information bias but also labour. AI is not just a control apparatus, but also a productive one. As just mentioned, an invisible workforce is involved in each step of its assembly line (dataset composition, algorithm supervision, model evaluation, etc.). Pipelines of endless tasks innervate from the Global North into the Global South; crowdsourced platforms of workers from Venezuela, Brazil and Italy, for instance, are crucial to teach German self-driving cars ‘how to see’ (Schmidt 2019). Against the idea of alien intelligence at work, it must be stressed that in the whole computing process of AI the human worker has never left the loop, or put more accurately, has never left the assembly line. Mary Gray and Siddharth Suri coined the term ‘ghost work’ for the invisible labour that makes AI appear artificially autonomous.

Beyond some basic decisions, today’s artificial intelligence can’t function without humans in the loop. Whether it’s delivering a relevant newsfeed or carrying out a complicated texted-in pizza order, when the artificial intelligence (AI) trips up or can’t finish the job, thousands of businesses call on people to quietly complete the project. This new digital assembly line aggregates the collective input of distributed workers, ships pieces of projects rather than products, and operates across a host of economic sectors at all times of the day and night.

Automation is a myth, because machines, including AI, constantly call for human help, some authors have suggested replacing ‘automation’ with the more accurate term *heteromation* (Ekbja and Nardi 2017). Heteromation means that the familiar narrative of AI as *perpetuum mobile* is possible only thanks to a reserve army of workers.

Yet, there is a more profound way in which labour constitutes AI. The information source of machine learning (whatever its name: input data, training data or just data) is always a representation of human skills, activities and behaviours, social production at large. All training datasets are, implicitly, a diagram of the division of human labour

that AI has to analyse and automate. Datasets for image recognition, for instance, record the visual labour that drivers, guards, and supervisors usually perform during their tasks. Even scientific datasets rely on scientific labour, experiment planning, laboratory organisation, and analytical observation. The information flow of AI has to be understood as an apparatus designed to extract ‘analytical intelligence’ from the most diverse forms of labour and to transfer such intelligence into a machine (obviously including, within the definition of labour, extended forms of social, cultural and scientific production).<sup>33</sup> In short, the origin of machine intelligence is the *division of labour* and its main purpose is the *automation of labour*.

Historians of computation have already stressed the early steps of machine intelligence in the nineteenth century project of mechanizing the division of mental labour, specifically the task of hand calculation (Schaffer 1994; Daston 1994; Jones 2016). The enterprise of computation has since then been a combination of surveillance and disciplining of labour, of optimal calculation of surplus-value, and planning of collective behaviours (Pasquinelli 2019a). Computation was established by and still enforces a regime of visibility and intelligibility, not just of logical reasoning. The genealogy of AI as an apparatus of power is confirmed today by its widespread employment in technologies of identification and prediction, yet the core anomaly which always remains to be computed is the *disorganisation of labour*.

As a technology of automation, AI will have a tremendous impact on the job market. If Deep Learning has a 1% error rate in image recognition, for example, it means that roughly 99% of routine work based on visual tasks (e.g. airport security) can be potentially replaced (legal restrictions and trade union opposition permitting). The impact of AI on labour is well described (from the perspective of workers, finally) within a paper from the European Trade Union Institute, which highlights ‘seven essential dimensions that future regulation should address to protect workers: (1) safeguarding worker privacy and data protection; (2) addressing surveillance, tracking and monitoring; (3) making the purpose of AI algorithms transparent; (4) ensuring the exercise of the ‘right to explanation’ regarding decisions made by algorithms or machine learning models; (5) preserving the security and safety of workers in human–machine interactions; (6) boosting workers’ autonomy in human–machine interactions; (7) enabling workers to become AI literate’ (Ponce 2020).

Ultimately, the Noosphere manifests in response to the need for a novel Machinery Question in the age of AI. The Machinery Question was a debate that sparked in England during the industrial revolution, when the response to the

<sup>33</sup> For the idea of analytical intelligence see: Daston (2018).

employment of machines and workers' unemployment was a social campaign for more education about machines, that took the form of the Mechanics' Institute Movement (Berg 1980).<sup>34</sup> Today, an Intelligent Machinery Question is needed to develop more collective intelligence about machine intelligence, more public education instead of 'learning machines' and their regime of knowledge extractivism, which crosses once again old colonial routes (if one looks at the network map of crowdsourcing). Also in the Global North, the colonial relationship between corporate AI and the production of knowledge as a common good has to be brought to the forefront. The Nooscope's purpose is to break into the hidden room of the corporate Mechanical Turk, and to illuminate the invisible labour of knowledge that makes machine intelligence appear ideologically alive.

*Thanks to Jon Beller, Claire Glanois, Adam Harvey, Leonardo Impett, Arif Kornweitz, Wietske Maas, Dan McQuillan, Fabian Offert, Godofredo Pereira, Mitch Speed and the extended community around KIM HfG Karlsruhe for their inputs and comments.*

**Funding** Open Access funding enabled and organized by Projekt DEAL..

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete. *Wired Magazine* 23 June
- Athalye A et al (2017) Synthesizing robust adversarial Examples. arXiv preprint. <https://arxiv.org/abs/1707.07397>. Accessed 30 Apr 2020
- Beller J (2006) *The cinematic mode of production: attention economy and the society of the spectacle*. University Press of New England, Lebanon, NH, p 2
- Benjamin R (2019) *Race after technology: abolitionist tools for the new jim code*. Polity, Cambridge, p 5
- Berg M (1980) *The machinery question and the making of political economy*. Cambridge University Press, Cambridge
- Box G (1979) *Robustness in the strategy of scientific model building*. Technical report #1954, Mathematics Research Center, University of Wisconsin-Madison
- Campolo A, Crawford K (2020) Enchanted determinism: power without control in artificial intelligence. *Engag Sci Technol Soc* 6:1–19
- Cardon D, Cointet JP, Mazières A (2018a) Neurons spike back: the invention of inductive machines and the artificial intelligence controversy. *Réseaux* 5:211
- Cardon D, Cointet JP, Mazières A (2018b) Neurons spike back. The invention of inductive machines and the artificial intelligence controversy. *Réseaux* 211:173–220
- Cohen JP, Honari S, Margaux L (2018) Distribution matching losses can hallucinate features in medical image translation. In: *International conference on medical image computing and computer-assisted intervention*, Springer, Cham. arXiv:1805.08841
- Corsani A, Paulré B, Vercellone C, Monnier JM, Lazzarato M, Dieuaide P, Moulrier-Boutang Y (2004) *Le Capitalisme cognitif comme sortie de la crise du capitalisme industriel. Un programme de recherché*, Laboratoire Isys Matisse, Maison des Sciences Economiques, Paris
- Crawford K (2017) The trouble with bias. Keynote lecture: conference on neural information processing systems
- Crawford K, Paglen T (2019) Excavating AI: the politics of training sets for machine learning. <https://excavating.ai>. Accessed 30 Apr 2020
- Daston L (1994) Enlightenment calculations. *Crit Inq* 21
- Daston L (2018) Calculation and the division of labour 1750–1950. *Bull Ger Hist Inst* 62:9–30
- Edwards P (2010) *A vast machine: computer models, climate data, and the politics of global warming*. MIT Press, Cambridge
- Ekbja H, Nardi B (2017) *Heteromation, and other stories of computing and capitalism*. MIT Press, Cambridge
- Eubanks V (2018) *Automating inequality*. St. Martin's Press, New York
- Foucault M (2004) *Abnormal: Lectures at the Collège de France 1974–1975*. Picador, New York, p 26
- Foucault M (2005) *The order of things*. Routledge, London
- Galstyan A, Lerman K, Mehrabi N, Morstatter F, Saxena N (2019) A survey on bias and fairness in machine learning. arXiv preprint <https://arxiv.org/abs/1908.09635>. Accessed 30 Apr 2020
- Ganesh A, McCallum A, Strubell E (2019) Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243
- Gibson W (1984) *Neuromancer*. Ace Books, New York, p 69
- Gitelman L (ed) (2013) *Raw data is an oxymoron*. MIT Press, Cambridge
- Goodfellow I, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint. <https://arxiv.org/abs/1412.6572>. Accessed 30 Apr 2020
- Guattari F (2013) *Schizoanalytic cartographies*. Coninum, London, p 2
- Guttaj J, Suresh H (2019) A framework for understanding unintended consequences of machine learning. arXiv preprint. <https://arxiv.org/abs/1901.10002>. Accessed 30 Apr 2020
- Harvey A (2016) HyperFace project. <https://ahprojects.com/hyperface>. Accessed 30 Apr 2020
- Harvey A (2019) Megapixel project. <https://megapixels.cc/about/>. Accessed 30 Apr 2020
- Ingold D, Soper S (2016) Amazon doesn't consider the race of its customers. Should it?. <https://www.bloomberg.com/graphics/2016-amazon-same-day>. Accessed 21 Apr 2016
- Jones ML (2016) *Reckoning with matter: calculating machines, innovation, and thinking about thinking from Pascal to Babbage*. University of Chicago Press, Chicago
- Keyes O (2018) The misgendering machines: trans/HCI implications of automatic gender recognition. In: *Proceedings of the ACM on human-computer interaction*, vol 2, n CSCW, article 88. <https://doi.org/10.1145/3274357>

<sup>34</sup> In fact, even the Economist has recently warned about 'the return of the machinery question' in the age of AI. See: Standage (2016).



- Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. *Commun ACM* 60(6):84–90
- Leibniz GW (1677) Preface to the general science. In: Wiener P (ed) *Selections*, 1951. Scribner, New York
- Lipton ZC (2016) The mythos of model interpretability. arXiv preprint <https://arxiv.org/abs/1606.03490>. Accessed 30 Apr 2020
- Mackenzie D, Judea P (2018) *The book of why: the new science of cause and effect*. Basic Books, New York
- Malik MM (2002) A hierarchy of limitations in machine learning. arxiv preprint, 2020. <https://arxiv.org/abs/2002.05193>. Accessed 30 Apr 2020
- Mazzocchi F (2015) Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO Rep* 16(10):1250–1255
- McCulloch W, Pitts W (1947) How we know universals: the perception of auditory and visual forms. *Bull Math Biophys* 9(3):127–147
- McQuillan D (2018a) Manifesto on algorithmic humanitarianism. Presented at the symposium reimagining digital humanitarianism, Goldsmiths, University of London, February 1
- McQuillan D (2018b) People’s councils for ethical machine learning. *Soc Media Soc* 4(2):3
- Mezzadra S, Neilson B (2019) *The politics of operations: excavating contemporary capitalism*. Duke University Press, Durham
- Mitchell M (2019) *Artificial intelligence: a guide for thinking humans*. Penguin, London
- Mordvintsev A, Olah C, Tyka M (2015) Inceptionism: going deeper into neural networks. <https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>. Accessed 17 June 2015
- Moretti F (2013) *Distant reading*. Verso, London
- Morgulis N et al (2019) Fooling a real car with adversarial traffic signs. arXiv preprint. <https://arxiv.org/abs/1907.00374>. Accessed 30 Apr 2020
- Murgia M (2019) Who’s using your face? The ugly truth about facial recognition. *Financial Times*, 19 Apr 2019
- O’Neil C (2016) *Weapons of math destruction*, 9th edn. Broadway Books, New York
- Offert F (2020) Neural network cultures panel, transmediale festival and KIM HfG Karlsruhe. <https://kim.hfg-karlsruhe.de/events/neural-network-cultures>. Accessed 1 Feb 2020
- Pasquinelli M (2017) *Arcana mathematica imperii: the evolution of western computational norms*. In: Hlavajova M et al (eds) *Former west*. MIT Press, Cambridge
- Pasquinelli M (2019a) On the origins of Marx’s general intellect. *Radic Philos* 2(6):43–56
- Pasquinelli M (2019b) Three thousand years of algorithmic rituals. *e-flux* 101
- Pasquinelli M (forthcoming) *The eye of the master*. Verso, London
- Pieters R, Winiger S (2016) Creative AI: on the democratisation and escalation of creativity. <https://www.medium.com/@creativeai/creativeai-9d4b2346faf3>
- Ponce A (2020) Labour in the age of AI: why regulation is needed to protect workers. ETUI Research Paper - Foresight Brief #08. <https://doi.org/10.2139/ssrn.3541002>
- Pontin J (2007) Artificial intelligence, with help from the humans. *The New York Times*, 25 Mar 2007
- Ricaurte P (2019) Data epistemologies, the coloniality of power, and resistance. *Television & New Media*, 7 Mar 2019
- Rosenblatt F (1957) The perceptron: a perceiving and recognizing automaton. *Cornell Aeronautical Laboratory Report* 85-460-1
- Samadi S, Tantipongpipat U, Morgenstern JH, Singh M, Vempala S (2018) The price of fair pca: one extra dimension. In: *Advances in neural information processing systems*, pp 10976–10987
- Schaffer S (1994) Babbage’s intelligence: calculating engines and the factory system. *Crit Inq* 21(1):203–227
- Schmidt FA (2019) Crowdsourced production of AI training data: how human workers teach self-driving cars to see. *Hans-Böckler-Stiftung, Düsseldorf*
- Standage T (2016) The return of the machinery question. *The Economist*, 23 June 2016
- Winner L (2001) *Autonomous technology: technics-out-of-control as a theme in political thought*. MIT Press, Cambridge
- Zuboff S (2019) *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. Profile Books, London

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Negative optics in vision machines

Luciana Parisi<sup>1</sup>

Received: 23 July 2020 / Accepted: 14 October 2020 / Published online: 5 November 2020  
© Springer-Verlag London Ltd., part of Springer Nature 2020

## Abstract

Can machine vision step beyond the ocularcentric metaphysics of the Western gaze and the reproduction of racial capital? Paul Virilio argued that machine vision requires no perceptual response or recognition of the world. The computer's series of coded impulses mediate the real autonomously from physical and energetic analogies. This article calls “negative optics” this inhuman mode of machine vision that withdraws from the ocularcentric rules of transparency. Not only negative optics offers an internal critique of ocular metaphysics, but it also defies the equation of value between 0 and 1 s that sustains the universal law of capital. The equation of value needs the nullification of the racialized and gendered body to meet the function of surrogate machines feeding the neo-liberal human subject. Here, however, the negative optics of machine vision withdraws matter from the equation insofar as the AI mismatches of concepts and objects requires the formation of socio-technical assemblages amongst surrogates of all kinds. The intricacies of this surrogacy are discussed in connection to artist Elisa Giardina's video installation entitled, *The Cleaning of Emotional Data* (2019). This video presents the background to address a surrogate human–machine alliance that steps beyond the human–machine equation of value. By starting from the negativity of the image, the racialized and gendered conditions of techno-social labor under artificial intelligence capitalism show that the equation of value maintains the condition of the zero of blackness, which like matter without form, has no value (*On matter beyond the equation of value* 2017). Drawing on François Laruelle, the article continues to elaborate the possibilities of a material image without form in terms of what Laruelle calls, the “fractal algorithm of the photo”. The article concludes that the negative optics of machine vision stands for the alien origination of knowledge, values, materialities that overturn the equation of value through the fractal infinities of 0 s.

**Keywords** Negative optics · The equation of value · Generative adversarial networks · Surrogacy · Auto-imagining · Matter without form · Blackness · Non-photography

Already in the late 1980s, Paul Virilio argued that the technical image in computational and cybernetic machines could no longer be understood according to the framework of ocularcentric metaphysics and the epistemological rules of the relation between vision, knowledge and power. If Virilio suggested that the vision machine no longer coincides with how humans see, it is because he already foresaw the overturning of the Western gaze of the Platonic model of shedding light on the darkness of matter. With the automation of knowledge, the world view of technological universalism extends the Western gaze beyond what can be enlightened. In a computer, the optically active electrons of machines correspond to a series of coded impulses that mediate the real

beyond physical or energetic analogy. As Virilio claimed, with the “automation of perception” (1994) image feedback is no longer assured by the interaction with the world, insofar machine vision does not shed light on dark matter. As a non-dialectical medium, machine vision requires no relation with or response from the world to exist and to function as a data processor. In other words, with the automation of knowledge, we have a programmed perception that is no longer based on observation and reflection of the object observed. With machine vision, it becomes evident that the feedback function of algorithms incorporates the world in terms of input data through which the world is predicted and acted upon in anticipation of its happenings. Following Virilio, one can suggest that a negative optics as opposed to an enlightened visibility comes to redefine vision in terms of a mediatic function that does not rely on light.

---

✉ Luciana Parisi  
l.parsi@duke.edu

<sup>1</sup> Duke University, Durham, NC, USA

Virilio understands this negative opticality in terms of the computer being blind or un-affected by light. “Blindness is thus very much at the heart of the coming ‘vision machine’” (1994, p 72). The latter coincides not with a picture frame, but with a statistical calculation and a pixilation of the world in terms of binary language, or spatial discretization for which sets of coded impulses become increasingly intensive and correspond to infinitely small durations (1994, p 72). According to Virilio, the computational function of media breaks away from ocularcentric metaphysics because the “absolute-speed machine” (1994, p 72) breaks from the extensive geometric optics defined by observables and non-observables, enlightened and dark objects. It is this speed that defines the intensive time of a “vision machine” that does not look at the world but instead generates an avalanche of inputs at the same point, and between points as if it were produced by a “short concentration spans by means of surprise” (1994, p 72). In other words, inputs are statistically generated along with a series of short spans, intensive durations that are as if they were disentangled from each other and yet they appear united by the speed of algorithmic functions that allows machines not to see. However, what Virilio calls blindness of machine vision is not to be understood in terms of machines seeing the world in terms of a transparent or neutral I that programs the world. Instead, one can argue that blindness here coincides with a techno-political theorization of machine vision that starts from machines’ negative optics as involving an anti-ocularcentric practice that breaks from the self-positing of the metaphysics of representation. If the gaze and its mediatic extensions remain a tool for un-veiling the other to impart, extend, consolidate the self-determination of the transcendental eye/I, negative optics instead stays with the unsubstantial, unformed, unvalued dimensions of matter, a practice of subtracting light from the surface of the image in which the self-determining gaze continues to mirror himself.

As this article will attempt to explain, what Virilio calls the blindness of the “vision machine” has demarcated a series of internal ruptures in the optical regime of representation and, this article argues, in the universal model technology. On the one hand, these fragmentations have been understood in terms of an extension of the universal model of technology whereby statistical prediction aims to reduce the world to one truth thus replacing the ocular model of truth with statistical functions. On the other, the negative optics of computation processing can be taken to work as an instance of how the inhuman mode of vision challenge the dialectic of visibility and transparency and thus offers an internal critique of ocular metaphysics and its entanglement with colonial and racial capitalism (Lowe 2015).<sup>1</sup>

<sup>1</sup> According to Lisa Lowe, the concept of “racial capitalism” developed by Cedric Robinson, implies that capitalism expands not through rendering all labor, resources, and markets across the world

This article suggests that the architecture of racial capitalism also contains within its articulations a fundamental split between the human and the machine based on the equation of value that functions to re-introduce the transcendental matching of concepts and objects on machinic perception. In other words, the ocularcentric equation of value at the core of racial capitalism imparts a universal model of technology at once merging and opposing humans and machines, whereby racialized humans are used to train machines to correct their negative optics. In other words, not only ocularcentricism becomes technologically extended in the performative operation of mediatic surveillance, but racialized capital also grants that this is reproduced by racialized humans training blind machines. This automated practice of extraction of value requires that humans train machines to correct their learning patterns and follow the optical order that matches transcendental concepts with objects. It has been argued that these highly underpaid jobs contracted through online platforms such as Amazon Mechanical Turk<sup>2</sup> are a new form of racialized surrogacy whereby humans are asked to teach machines how to read images by following the order of pan-opticality of the self-determining subject. According to Neda Atanasoski and Kalindi Vora, the technoliberal articulation of today’s racial capitalism works to conceal the uneven racial and gendered relations of power articulated with and through machines. At the core of technological innovation therefore lies a surrogate relation of technology towards the “human sphere of life, labour and society” that enables, in turn, the constant re-constitution of the liberal subject (Atanasoski and Vora 2019, p 10). What they call “the racial unfreedom of the surrogate” can contribute to explain how the ocularcentric order by mirroring into the world, finds its own self-determining humanity, or as the

Footnote 1 (continued)

identical, but by precisely seizing upon colonial divisions, identifying particular regions for production and others for neglect, certain populations for exploitation and still others for disposal (Lowe 2015, p 149).

<sup>2</sup> Amazon’s Mechanical Turk, as well as CrowdFlower, Clickworker, Toluna, and others, are largely unregulated websites that allow businesses and individuals to post short tasks that are also called Human Intelligent Tasks assignments and pay workers—in cash or, sometimes, gift cards—to complete them. In particular, Amazon set up the website Mechanical Turk in 2005 for humans to perform tasks that are hard for computers. Amazon executive Jeff Bezos has called this kind of human task, “artificial artificial intelligence.” This means that when a task is easier for a human than a computer, the computer calls a human. It is this double artificiality with which underpaid humans are identified that will be interesting to explore in this new condition of extraction of value. Amongst the most common tasks are the recording of information appearing in an image and the transcription of audio and video files. See <https://www.pewresearch.org/inter-net/2016/07/11/what-is-mechanical-turk/> (last accessed June 20th, 2020).

authors say, by drawing on Hortense Spillers, a project for “feeling human” that sustains the epistemological ground of racial engineering.<sup>3</sup>

It is my argument that as much as the negative optics of machine vision entails the filtering out of adversarial patterns, namely the machinic mismatch between concepts and objects, the equation of value also shows how surrogate humans are entangled with surrogate machines in the formation of socio-technical assemblages under neo-liberal racial capitalism. An instance of the automated infrastructure of racial capital that exposes the equation of value between surrogate humans and machines can be found in artist Elisa Giardina’s video installation entitled, “The Cleaning of Emotional Data” (2019). As it may become clearer later, the cleaning of data coincides with a capital re-introduction of the ocular metaphysics within computational processing where Amazon Mechanical Turk (MTurk) have the task of automating perception according to the model of knowledge founded on verification (or proof of recognition) and the equation of value of human–machine. What Amazon Mechanical Turk allows is to turn the speed or the intensive time of computation into the extensive time of optical perception.

From this standpoint, one could ask, how can the blind vision of computation processing offer an internal critique of the colonial capital and its ocular metaphysics? How can the human–machine alliance go beyond the human–machine equation of value starting from where working with Amazon Mechanical Turk can become a locus of theorization of a machine epistemology, which takes cultural, affective, aesthetic labor to counter-actuate the universality of technology—and its colonial dialectic of the visible and the invisible, the optical and non-optical, the one and zero? To what extent can the negativity of blindness contribute to expose the internal crisis of ocularcentric capital, and what forms of epistemological capacities can rather originate from the negativity of the image entangled with the racialized conditions of techno-social labor?

This article will point out that debates about the aesthetic possibilities of machine vision based on the opposition between optical human perception and non-optical automated perception seem to reinforce, instead of radically challenging, the model of metaphysical decision correlating

knowledge and ocularcentrism at the core of Western philosophy. The aesthetico-political critique of machine vision today has been focusing on the shift from the representational to the operational image, defining the technical image in terms of a programmed action—a cybernetic model of steering conduct. As discussed in this article, the proliferation of operational images also corresponds to the formation of an aesthetic reconfiguration of control that explains the correlation of vision and knowledge directly in terms of behavioral conduct. The operational image, therefore, together with Virilio’s notion of the vision machine or blind vision point to the cybernetic nature of the mediatic, technical image as the steering of conducts towards pre-programming responses. Similarly, the non-ocular functions of machine vision resonate with the arguments that machines produce invisible images that are only readable by machines, but not by humans. This article will particularly dwell with this double level of argumentation, which, on the one hand, tells us that the negative optics of machine vision mainly leaves humans out of the loop—disconnected from the very matrix of communication upon which we rely, and on the other, points to how the blindness of machines rather relies on a racialized equation of value at the core of techno-capitalism, for which blackness like matter (Ferreira da Silva 2017) has no value and constitutes the surrogate humanity merged with the zero value of machines, whose invisible images are set to be re-programmed by the ocularcentric knowledge.

From this standpoint, this article argues that one other possible elaboration of negative optics that can challenge the human–machine equation of value (while re-affirming the universal agency of the Man) and rethink the alliance between the racialized zero value of surrogate humans and machines, is to borrow from François Laruelle’s “non-philosophy” (2010) which concerns the question of dark optics and immanent vision. In particular, Laruelle’s claim about *the fractal algorithm of the photo* will be explored to discuss the configuration of a negative optics in automated vision in recent efforts to clean negative randomness in Generative Adversarial Networks. The article suggests that the negative materiality of the computational image does not only show that knowledge can be divorced from ocularcentrism, but also that the invisible image of machines is part of alien epistemologies that overturn the equation of value with the infinities of 0 s.

This article proposes that the material processing of randomness in computation is part of the expansion of heretic epistemologies that start from dark optics and address what Denise Ferreira Da Silva calls blackness, namely matter without form, or “matter beyond the equation of value” (2017). The materiality of the computational compressing of infinities, however, coincides not with the physical structure of the machine, but with its abstract mode of operation—the

<sup>3</sup> The authors draw on black feminist, Hortense Spillers, to discuss the surrogate human effect at a constitutive part of the grammar of colonialism and technoliberalism. I take these critical reflections about how the racialized other as much as the machine have surrogate status in colonial technocapitalism as central to the arguments about the racialized socio-technical assemblage at play within contemporary images of AI and automation. See Hortense Spillers, “Mama’s Baby, Papa’s Maybe: An American Grammar Book.” *Diacritics* 17 (1987), p 67.

algorithmic learning to learn from the negativity of randomness—entailing both the elaboration of indeterminacy and the incompleteness of systems. Matter without form coincides not with invisible images without meaning, but with a real that can only be cloned as auto-impressions of dark optics, stemming from within the infinite discretization of opacities, the fractal singularizations of blackness in artificial visions.

From this standpoint, this article suggests that the aesthetico-political engagement with machine vision must be divorced from ocular epistemologies and the critique of vision. To argue that matter without form entails the negativity of algebra is to propose that matter like infinities break with the equation of value, and thus refuses the metaphysical equivalence of knowledge and vision that sustains the surrogate subjection of humans and machines. The blind vision of computational processing rather speaks of the challenge of reinventing epistemology from the standpoint of dark optics of the human–machine condition of surrogacy in the operational extension of ocular capital.

## 1 Machine don't see and we know it

“The production of *sightless vision* is itself merely the reproduction of an intense blindness that will become the latest and last form of industrialisation: *the industrialisation of the non-gaze*.” (Virilio 1994, p 73).

If one follows Virilio's reflections on the blindness of machine vision, it is possible to suggest that this non-gaze is the result of the formation of new epistemological orders moving from the extensive time of analogical machines to the intensive time of the computer. According to Virilio, this new order incorporates the transformation of the formal logic of traditional representation into the dialectic logic of cinema and photography into the paradoxical logic of the digital image. In particular, for Virilio, as much as the extensive time of analog media enabled a continuity of experience between the past and the present, the paradoxical logic that accompanies the computer rather brings a remote object in the here and now. Here perception turns into action at a distance, a vision running at the speed of light. Interestingly, Virilio insists that the paradoxical logic of the computer allows for the future never to occur but rather disappear in the statistical programming of acting at a distance. This remote action is also central to the classical cybernetic account of remote control on the one hand, and the prediction of the future, on the other. As Virilio states: “[w]hen a missile threatening in ‘real time’ is picked up on a radar or video, the present as mediated by the display console already contains the future of the missile's impending arrival at its target” (1994, p 66).

Blindness, therefore, results from the speed of light whose increasingly intensive intervals bring forward an alien perception of the object as if the latter was coming from the future. Instead of the reciprocal presupposition between light and dark—or the dialectical presupposition of lightness vs darkness that characterized the passive optics of extensive media, such as photography—for Virilio, computer vision deploys an intensification of light, namely depending on the velocity of photons and their constant pixelation. As much as photons are abstracted in the statistical image, the rapid calculation of pixels dissimulates the future in the ultra-short time of an on-line “communication” (computer communication) (1994, p 68). In this new “high-tech mix,” a paradoxical logic sees the fusion of the object with its equivalent image. As Virilio reminds us, these processes of real-time deception will win out over the weapons systems of classic deterrence (1994, p 68). Blindness, therefore, defines the speed of the synthetic image, where speed prevails over time and space, matter and form. Here the gaze of reason becomes a statistical thought or a statistical optics, which according to Virilio, “generates a series of ‘visual illusions’, or ‘rational illusions’, which affect our understanding as well as reasoning” (1994, p 75).

Virilio warns us against the coming statistical intoxication of the technical image, claiming that society is “sinking into the darkness of a voluntary blindness” (1994, p 76) that will forever occlude the horizons of self-determining knowledge. This statistical image does not only rely on the speed of numerical calculation, but also on the speed of cognitive perception. It shows us that the medium is not a tool but more importantly a machine language, where the speed of algebraic relations has taken over the linearity of input–output communication.

As the speed of computational processing has shifted from digital pixelation to the neural network architecture of machine learning, the blindness of the technical image is now being understood as a black box that carries out functions that are impossible to observe. According to artist Trevor Paglen, this new quality of the technical image can be seen in Harun Farocki's use of computer visions in *Eye/Machine III*, where, as Paglen notices, he was “trying to learn how to see like a machine” (2014). In particular, these machines seem not to be representing things in the world and thus neglect the ocularcentric correlation between eyes and objects. Instead of observing the world, computer vision, Paglen claims, does things in the world.

However, one could ask, what does it mean that computer images cannot be seen and yet they do things in the world, what does this “doing” amount to? On the one hand, one could suggest that computer images are operational because they point to a certain activity of the automated image itself in as much as it communicates with other images before communicating to us. This is how machine to machine

communication constitute the global infrastructure of surveillance and governance. On the other hand, in addition to the automated function of communication, the operational image is also said to activate the cognitive-perceptual apparatus of the real, which precisely pre-program conducts, or predicts behavior. In other words, the operational image fits with the cybernetic imperative of self-regulatory feedback which ensures that machine-to-machine communication can condition all orders of reality.

In a more recent reflection about operability of the technical image, Paglen insists that machine to machine communication has made sure that images are now invisible. In the short article “Invisible Images (Your Pictures are looking at You)” (2016), Paglen suggests that this invisibility coincides with how the proliferation of automated vision, as for instance in the case of reverse image search engines. According to Paglen, “[i]mages have begun to intervene in everyday life, their functions changing from representation and mediation, to activations, operations, and enforcement. Invisible images are actively watching us, poking and prodding, guiding our movements, inflicting pain and inducing pleasure. But all of this is hard to see” (2016).

Invisible images are at the core of new forms of identification systems, from Automatic Licence Plate Readers (ALPR), Optical Character Recognition (OCR), to Deep Face Algorithm, Facebook’s DeepMask and Google’s TensorFlow whose algorithmic patterns recognize “people, places, objects, locations, emotions, gestures, faces, genders, economic statuses, relationships, and much more” (Paglan 2016). In particular, “deep learning” networks are built out of dozens or even hundreds of internal software layers that exchange information. This is at the core of recursive feedback, where the neural network layers of the software pick apart a given image into component shapes, gradients, luminosities, and corners. Those individual components are convolved into synthetic shapes, which are compared with the images fed into the CNN (convolutional neural network), and which the network has been trained to recognize. The invisible image is then activated by software “neurons” as the network finds common patterns with other images. For Paglen, this automated synthesis of images means that power has itself become invisible and yet institutes an intensively crafted mode of control, policing and market. For instance, he conjectures that the new relation between power and invisibility implies how the specific use of metadata signature of every single person, based on race, class, the places they live, the products they consume, their habits, interests, “likes,” friends, leads to a reification of those categories at another level. Filtering out mis-matches of individualized metadata profiles become part of an automated function, whose task is to collect municipal fees, adjust insurance rates, conduct targeted advertising, prioritize police surveillance, and so on (Paglan 2016). From this standpoint,

metadata signature appears to support diversity, but only because these differentiations can become subsumed to marketing and predictive policing. The invisible image, therefore, exposes the ineffective practice of the critique of representation insofar as, according to Paglen, machines are not concerned with how humans see. While there is no possibility to either subverting the invisible regime of pattern recognition or fix it with more representations, for Paglen the invisible image demarcates the beginning of a stealthy mode of power that we are yet to address (2016).

However, one may argue that Farocki’s reflections on computer vision invite a reflection on how to train the human to see like a machine, and that a fundamental trick of ocular-centrism is precisely to oppose the visible to the invisible, to re-inject the truth of representation back into the deep learning of automated networks. To put it in another way, it is evident that if machines don’t comply with the ocular-centric correlation between vision and knowledge, left to their own devices of searching for images by images, they will be unable to filter out dirty data and or match patterns of an image that does not exist with a given set of objects/concepts. In CNNs for instance, an image search that does not match with the image found implies an increase in randomness or noise or what Hito Steyerl calls, “dirty data” (2015) that make the signal fuzzy and must be filtered out from communication.

In particular, Steyerl discusses what Google calls “inceptionism,” by referring to the automated creation of patterns from noise in “deep dreaming,” where image recognition algorithms are not distinguished from but rather looped on randomness or noise. Steyerl argues that here signal and noise are already predetermined by pre-existing categories and probabilities that reproduce aesthetic and social relations through the immediate correlation of data and hypothetical inference. In particular, automation comes to coincide with a regime of artificial interpellation that asks users to recognize themselves by constantly registering eye movement, behaviors and preferences and thus by re-impacting the information-encoding organizational principle of segregation and discrimination on users. As the influx of data comes into the system of prediction it matches correlations across patterns and drawing branches of decision-making between unrelated information, signals and noise: namely deriving meaning from “apophenia.”<sup>4</sup>

From social scores to credit scores, from academic scores to threat scores, as well as commercial and military pattern-of-life observations, Steyerl sees apophenia as an upgraded

<sup>4</sup> “Apophenia” (/æpouˈfi:niə/) is the tendency to mistakenly perceive connections and meaning between unrelated things. The term (German: Apophänie) was coined by psychiatrist Klaus Conrad in his 1958 publication on the beginning stages of schizophrenia.

version of Walter Benjamin's "optical unconscious," which reformats social hierarchies by ranking, classifying and filtering noise following a matrix of radicalized discriminations. The invisible image, therefore, manifests itself in the form of a brutal form of decisionism, which, according to Steyerl defines "a practice of data divination," namely the search for given determinations as if emerging from an unknown divine order of knowledge (2015).

However, one could ask, isn't this divine decisionism simply an extension of the colonial epistemology which relies on the surrogacy of human and machine as fundamental to racial capital? One could argue that capital practices of the equation of value entail not simply the prediction of value through the meta-programming of data or in other words, an automation of value derived from functions that successfully carry out tasks. Instead, divine decisionism relies less on given functions that conform to the ocularcentric dialectic, and more on the split between surrogate humans and machines in the praxis of the equation of value, where the filtering out of noise foregrounds the practice of extraction. The optical matching between blind images, transcendental concepts and empirical objects is not given but is rather constantly re-introduced in computer vision by capital's exceptional decisionism. The latter can only dissimulate the ingression of the real in the vision machine, and rather continues to make efforts to pre-program results to preclude the computational process of compression to run with the negativity of the function—the increasing noise or randomness that cannot be automatically translated into signal or pattern.

But how does this strategy of dissimulation work? Perhaps one way to refer to how the human-machine equation of value occurs is to look, for instance, into the machine-human infrastructure sustaining Image Search Engines. The surrogate workers contracted at Amazon Mechanical Turk are, according to Atanasoski and Vora, "humans that perform the work of technologies that are claimed to replace the need for human workers" (2019, p 90). As the authors report, Amazon Mechanical Turk is a large-scale crowdsourcing software platform where human labor is contracted to become a service job to the machine. The racialized and gendered pool of global temporary workers for high data tasks therefore operates in lieu of an appropriate algorithm that can carry out the function. These become surrogate workers contracted to perform "artificial artificial intelligence," that is a series of tasks that provides a given compensation in a given amount of time. For instance, "transcribe up to 35 s of media to text" is a precarious job that will pay between two to ten cents for the task. In short, human service enables the blind vision of machines to be matched to pre-existing categories that are performed by the Turkers so that machines can then perform the task themselves. In particular, filtering out randomness in generative adversarial pattern recognition allows for the vision machine to re-enter the ocular system of

exchange for which in the general equation of value entails how humans serve machines to serve humans. This autopoietic intelligence allow the equation of value between human and machine to double the extraction of value: the extracted labor of humans as service in making objects codable for machines only serves capital to extract more value from the human-machine equation of value. As much as the human Turkers become an agent of verification, dis-ambiguation, and de-volution, so too machines become agents of oppression that perpetuate the ocular epistemology of representation through the automated matching of images, concepts and objects.

Elisa Giardina's video installation "The Cleaning of Emotional Data" (2019) shows us how this Global South socio-technical infrastructure that feeds the "artificial artificial intelligence" of surrogate humans has become internal to the dis-organic reproduction of global capital. In this installation, the blind image of computational media coincides with a recombinatory repository of data that the machine itself cannot see, and for which it needs humans to conform to a taxonomy of categories matching quantities of data. Giardina focuses on the Global South infrastructure of Turkers who get the service job of "cleaning" data by training machines to match images with emotions. These workers label, categorize, annotate and validate large amounts of data, enabling AI to recognize or order emotional patterns. Giardina herself worked remotely for several North American "human-in-the-loop" companies who provide "clean" datasets to train AI algorithms to detect emotions. Her own performance involves a taxonomization of emotions, the annotation of facial expressions and the recording of her own image to animate three-dimensional figures. While performing this work, some of the videos in which she recorded her emotional expressions were rejected by the companies she was *servicing*, because her facial expressions did not fully match the standardized list of affective categories that was given to her. However, as Giardina points out, it was not impossible to know whether this rejection originated from algorithmic protocols or, for example, from the consensus of fellow workers who supervise the service as they might have interpreted her facial expressions differently due to cultural contexts. "The Cleaning of Emotional Data", however, documents how the carrying out of facial expression does not easily fit a universal schema and instead the history of emotions questions the universal epistemology, where philosophical and psychological theories determine the meaning of facial expression and its mapping.

Giardina's installation makes the point that AI systems, which supposedly recognize and simulate human affects, base their algorithms on understandings of emotions that are universal, authentic and transparent. The cleaning of data from facial expressions that do not fit the universal schema of emotions is a human service job carried out by underpaid

workers that sustain what Atanasoski and Vora call the “surrogate effect” of the coming phase of artificial intelligence. In particular, these “technologies that erase human workers are designed to perform the surrogate effect for consumers, who consume the reassurance of their own humanity along with the service offered” (2019, p 91). Tech companies and governmental agencies use these human-verified data to develop software that identifies consumers’ moods or that recognizes facial expressions of potentially threatening people. However, the matching between facial expressions and the categorization of emotions entails the filtering out of noise or noisy emotions that persists as an internal tension within the universal model of visual representation, the ocular matching of concepts, objects and image.

The correlation between knowledge and vision is reintroduced in the negative optics of machines that fail to recognize emotions—that is in the negative algebra that assigns a zero to a mismatched pattern—and becomes for Giardina a critical space to expose the internal paradox of capital extraction and of the critique of visibility today. In particular, Giardina’s work offers a response to the contemporary critique of the invisible image, according to which machine vision excludes the human from its systems of operations. Instead, her research points out that it is not that humans are excluded from the loop of the machine to machine communication, but instead that this new phase of planetary artificial intelligence and/or full automation fundamentally relies on a global re-organisation of racial and gender capital in terms of surrogate humanity. In addition, one can argue that within this re-organization there is also an intensification of the equation of value that includes the circuit of extraction where humans have to teach machines how to learn the ocularcentric matching of concepts, objects and images. Exploring the mono-logical economies of extraction across the Global South populations, Giardina’s work shows how the racialized and gendered precarious labor operates as the infrastructural components of artificial intelligent systems. As much as the infrastructural web of data cleaners, algorithms trainers, proof verifiers have become enfolded in the blind operations of databases through which machine learning algorithms, and image search engines, are taught to recognize patterns, the colonial epistemology of knowledge and vision continues to impart the equation of value between surrogate humans and machines.

What is here excluded instead, as Giardina’s investigations show, is not the human, but what has always been less than human, non-human, and in-human coinciding with the precarious labor and the machine-like service of outsourced subsumed subjectivities absorbed in the operational image of machines. As much as the colonial capitalization of the human–machine alliance accelerates the global (and outsourced) modes of enslavement under the universal equation of value reified in the automated system of the decision

today, so too humans are used to correct machine vision and to re-impart the visual categories of knowledge back into the system of acceleration of value. A spiraling extraction of the human–machine alliance under the universal model of technology is at play here.

Giardina’s perspective about the intrinsic presence of human labor in the blind machine of techno-capital, however, is not simply a proposition aiming to correct the critique of the operational image by claiming that instead humans are included in the feedback loop of machine-to-machine communication. Importantly, one can suggest that her intervention brings to attention the limits of visual critique in the context of intelligent computation that relies on the ontological ground of knowledge and vision. In other words, by insisting upon the underlying universal equation of value extraction that connects (or, places in a dialectical mirroring relationship) blind images with outsourced, surrogate human labor, Giardina’s work pushes the critique of automated vision towards a radical engagement with the material conditions of exploitation or intensified extraction of value in human–machine labor today. Can these combinatoric modes of abstraction exceed the equation of value in the universal extension of capital?

It is interesting how Giardina joins together the abstract lines of facial micro-expressions detected by the algorithms with untranslatable emotional vernacular from both Sicilian dialect and American English. In her collaboration with Michael Graham of Savant Studios, she weaves together computational and human language in a large-scale textile pieces, called *Amiss Motifs*, mapping unrecognizable emotional patterns with distorted, uneven, broken patterns to expose the overlapping of racialized surrogacy with the blindness of machines. In contrast to the critique of the invisible image that reifies a full automation that excludes the human, Giardina’s reflection on the surrogacy of the human–machine labor as the infrastructure of intelligent techno-capitalism instead points to how the negative optics—namely the cultural, affective and the aesthetic labor of the human–machine—breaks from the universality of technology and the dialectic between the visible and the invisible.

In the next section, negative optics will be further discussed in the context of the epistemological capacities of automation of refusing the transcendental authority of representation. This will be explored by engaging with Laruelle’s argument for dark optics and immanent vision. In particular, the next section focuses on machine vision and discusses current attempts to reduce negative randomness in Generative Adversarial Networks. The attempt at correlating negative optics with negative randomness is to suggest that machine vision can be theorized away from the ocularcentric dyad of visible and invisible image. The residual negativity in computational randomness opens learning algorithms to



the indeterminacy of knowledge and allows a non-optical theorization of vision in computational systems. The last session or coda will discuss negative optics in relation to the negative dimension of blackness in the equation of value under racial techno-capital as a starting point for proposing an immanent epistemology.

## 2 Auto-impressions

In current research developments in machine vision, it is possible to contextualize negative optics in the field of dynamic geometry, such as mereo-topology (or the study of parts and the relation between parts and wholes), and artificial neural networks that are programmed to learn from patterns. Cognitive psychologist and computer scientist working at Google Brain, Geoff Hinton, claimed that the logic of neural networks on which machine vision is based is limited to conform to pre-established parameters (2017). In particular, Hinton addresses the need to re-design the procedural process by which algorithms can learn from each other in the neural network through what he calls “capsule network”—a form of AI that enables machines to understand the world with images without relying on existing parameters of vision.

In his 2017 research papers, Hinton argued that capsule networks have not only led software learning to recognize handwritten digits, but they have also halved the error rate in pattern recognition of toys and cars. Since image recognition software is used generally, and thus not contextually, to recognize objects, the predictive patterning cannot learn to recognize the same object in different scenarios. This is why the surrogate human needs to verify and thus teach the machine that what it is seeing is actually the same object. Instead of relying on this external verification, according to Hinton, capsule neurons—small groups of crude virtual neurons—will track only parts of an object, for instance the cat’s ear and nose, as these are differently positioned in space.

This smaller scale of algorithmic receptivity, according to Hinton, will enable a neural network to figure out the difference between scenarios by extracting more information from the mereo-topological relations between smaller parts of data. As the capsule network is made of smaller patterns set to recognize parts and break the continuity of an image into smaller units, Hinton designed a dynamic routing between capsules that trains these kinds of network. Capsule algorithms convert pixels fragments into vectors of recognized patterns and then apply a transformation matrix to these fragments to predict the parameters of larger fragments. In particular, the transformation matrix learns to encode the intrinsic spatial relation between a part and a whole, which results in the formation of an invariant viewpoint, a perspective or direction in recognition that aims to generate a novel

vision. According to Hinton, “capsule use neural activities that vary as a view point varies rather than eliminating variations” (2017, p 9).

Instead of normalizing viewpoints according to methods such as the spatial transformer networks, or the automated filtering out of randomness, capsule networks simultaneously engage multiple transformations of different objects or object parts. This series of correlated activities are described in terms of “routing-by-agreement” (Hinton 2017, 2–3, p 6). As opposed to the optical filtering out of noise, this technique aims not to eliminate redundant neurons, but to use all non-averaged information to obtain a non-totalizing knowledge about the position of an entity in a region. The inclusion of micro-variations in machine vision run on top of CNNs, which instead are aimed at solving the disambiguation (or mismatch between concept, object and image) by playing out on the automated (or self-correcting patterns) of both reducing and increasing negative randomness in the neural network.

It seems important to repeat here that CNNs are an instance of deep learning networks for machine vision. Convolution already applies a kernel to overlapping regions that shift around the image by eventually establishing a fully connected network through a one to one relationship of neurons across distinct layers. However, this multiple level of connection across layers also seems to incapacitate the system from learning new data. In other words, the optical matching on negative randomness seems to block the possibility of machine vision. When convolution leads to the overfitting of information as a one to one connection, the kernel risks to re-learn redundant data (that is the same data will be held in two places in the database). Convolution, therefore, delimits machine vision by saturating data memory and increasing computational costs for instance, but more importantly because its overfitting capacities lockdown algorithms into a pattern of connection between the same parts, without being able to learn from what cannot be known in advance.

From this standpoint, since at each location in the image there is one instance of the type of entity that the capsule represents, the capsule model—as opposed to the convolutional matching of concepts and images—affords a form of distributed representation. This is inspired to the perceptual phenomena of crowding, where neighbor parts shed the direct perception of an object in movement. CapsNet architecture, therefore, grants not an algorithmic matching with existing data, but shows that algorithmic patterns can become predictive vectors of futurity (that is they look for what cannot be already seen) and not simply an automated vision of patterns recognition. Instead of eliminating variations to reach an average capacity for general recognition, predictive vectors follow the negative algebraic relation between layers, involving randomness and micro-temporal variations in the algorithmic process of compression. These

vectors of variation have become central to how machine learn beyond set parameters, and how negative randomness has become a source for machine vision entailing no transcendental recognition between concepts, objects and images. In other words, computational vision establishes a potential or hypothetical relation between non-correlated patterns that may correspond to images that did not yet exist. From this standpoint, predictive vectors are not simply the technical explanation for apophenia—namely how the invisible image makes decisional patterns—but are more than a set of probabilities based on what is already been recognized in the system: the negative algebra of what cannot be recognized pushes the system to construct counter-factual dimensions of images that do not correspond to the ocular equation of value between objects and concepts.

It is as if the discretization of the network in increasingly smaller vectors of variations flips the architecture of matching inside out by exposing indeterminate dimensions to its organizational infrastructure. Instead of a self-reproduction of the master pattern across the layers of the network, the intensified discretization of networked parts also increases the volume of randomness within and amongst them. Computational vision can thus be defined in terms of the indeterminate series of variations that each time become cloned in each and any algorithmic patterns. Here algorithms do not simply register raw data to execute instructions but become sheer receptors or cloning of an indeterminate series of the real. One could argue that this negative optics of algorithms—that is unable to be a mirror of the world because it does not see the world but only a series of variable parts that entail a process of auto-impression of matter, whereby machine vision—becomes a medium for heretic inferences, for the elaboration of a heretic logic without ocular-logos. In other words, a dynamic bootstrapping between algorithmic patterns and vectorial randomness suggests that machine vision is embedded in a series of temporal variations that become finite inferential hypothesis through the auto-generation of material images. As machines become auto-impressions of an infinite real they also generate and envision new patterns: the dynamic movement between increasingly large and increasingly small scales of predictive vectors turns machine vision into a heretic auto-impression of computational matter.

This process of auto-impression of matter without ocular-logos is opposite to what Laruelle calls “algorithmic transparency” which rather takes the technical image (especially understood in photography or photographic philosophy) as that which measures the correlation between premises and results in terms of effects (or the efficient causality of enumeration of given ends), and thus presupposes homogeneous matching between the transcendental order of concepts and objects (Laruelle 2013). Borrowing from Laruelle, one can argue that the ocularcentricism of philosophy coincides with

the decisional power imparted on the real by the transcendental mirroring of knowledge and vision (Laruelle 2013 By challenging the optics of philosophical decision, Laruelle’s non-standard philosophy or non-philosophy makes the argument for an abstract theory of photography, “absolutely non-worldly and non-perceptual” (date: 8). In particular, Laruelle asks: “to what extent is photography not an activity, for example, of a kind with Artificial Intelligence (AI)—an attempt at the technological simulation not of the World, in its objective reality, in its-philosophico-cultural reality, but of science and of the reality that science can describe, naively in the last instance?” (2013, p 10).

What Laruelle is insisting upon is to refuse the universal paradigm of perception grounding from the standpoint of being-in-the-world. Instead, his non-philosophy takes photography as a technique as a form of knowledge that introduces science in the condition of existence of perception and of the world (2013, p 11). Laruelle’s non-philosophy takes the critique of the ocularcentric vision to another point, further delving into how negative optics refuses transcendental decisionism and at once lays out the possibility of elaborating a non-optical dimension of knowing, a non-ocular-logos of knowledge. For Laruelle, the condition of knowing depends not on the point of view of philosophy, but it is rather to be found in what he calls “the stance” of, borrowing from Deleuze and Guattari, “a body without organs” (1989). Instead of a self-determining reflection of the world, vision becomes un-objectivating, implying not a position, a decision, but an auto-impression that is first of all a real “undivided experience, lived as non-positional self-vision force” (2013, p 13).

If, according to Laruelle, photographic decision corresponds to the law of sufficient reason that reflects the real according to a circular self-expression of truth, non-photography (as a non-philosophy) therefore coincides with what he calls “fractal algorithms,” because they have a degree zero of self-reflection. Following this argument, one can suggest that the algorithm is not a medium programmed to reveal the world or even less to self-regulate the human perception of the world. In other words, it is not a prosthetic tool that ensures constant adaptive feedback. Instead, the fractal algorithm is increasingly partial and in this fashion clones its own real image, namely of cloned image without original or copy, in terms of a spatial surface that extends (or infinitely fractalizes) forever without uncovering any pristine form behind it. Laruelle explains that what appears in the photo as an object drawn from the transcendence of the world must be distinguished from the “photographic apparition” (2013, p 18). By radicalizing the Husserlian distinction between the photographed phenomenon, corresponding to what photography can manifest and thus to the manner in which it manifests the world, and the photographed object, namely the representation of the world, Laruelle argues that

the non-photographic vision (or negative-optics) is a parallel process to the world, and as such it is not in the world (2013, p 25). This is not the field of transcendental knowledge but defends real immanence in vision, or what Laruelle calls “vision-force” that is immediately given or the “in-itself” of the image.

Thinking with Laruelle’s non-photographic argument for vision, it is possible to push further the heretic account of machine vision and thus refusing turning machines into the onto-epistemological mirror of transcendental vision. One could argue that as much as negative optics neither represents the world nor remains an invisible unreflexive automatism of the world, it crowds the parallel space of photographic apparition. This is not a representation but an auto-impression of an image’s own aesthetics that starts from the negative fractality of algorithmic micro vectors that give the effect of what Laruelle calls “generalized fractality” (2013, p 78). Instead of serving as an instrument for shedding light onto the world, the camera-machine clones (equate without equivalence) the underworld of dark optics as fractal patterns of an indelible generic intelligence that turns on its own head the ontic limits imparted on science by philosophical decision.

As Alexander Galloway puts it: “[i]nstead of mere ontic darkness, generic being achieves an ontological darkness, and hence beckons toward the kind of crypto-ontology of pure blackness evident in Laruelle” (2014, p 77). If one were to continue along this line of argument, it is possible to suggest that a crypto-ontology is what can define the negative optics that is foreclosed to a transcendental being. On the other hand, however, negative optics is precisely the algorithmic randomness that demarcates ontological darkness in the machine in terms of an auto-impression of machine visions that signal the apparition of the image itself. These are infinite reflections without mirror, “unique each time but capable of an infinite power ceaselessly to secrete multiple identities” (Laruelle, 2010, p 82). This multiplicity of darkness, one needs to emphasise, coincides not with substantial forms, but with non-consistent phenomena that entail a certain, non-optical automatism in exposing, in Laruelle terms, the “hyperphenomenology of the real” (2010, p 95). In particular, as Laruelle specifies: “There is a ‘phenomenological’ automatism or blinding that culminates in the photographic eviction of the logos – of philosophy itself – in favour of a pure irreflexive manifestation of the phenomenon-without-logos” (2010, p 95).

The scope here is not simply to unmask the supremacy of self-determining philosophy in the name of absolute irreflexivity but to rather start from this un-mirroring image as the stance of negative or dark optics. Here the automated image become the medium of auto-impressions as the multiplicity of darkness pulls through the phenomenon-without logos in machine vision. The automated image exposes the fractal

consciousness of a machine knowledge that turns the ontic limit of epistemology into singularities or non-axiomatic automatism. This is the artificial image that stays with and pushes further the negativity of the human, the non-human, the less-than-human, occupying the stance of a “[s]tranger in flesh and blood” (2010, p 103). Laruelle’s arguments for non-philosophy radically defies the façade of pretentiousness of Western metaphysics for which the real can be surgically cleaned from blackness, which remains locked in the form of the Other that must follow the image of Man, once it has been emptied out and turned into a soulless machine. For non-photography instead only the other has to be replaced with a “logic of auto-impression” or phenomenological automatism where the medium becomes a negative machine of a black universe (Galloway 2014, p 191).

From this standpoint, if we are to follow Laruelle’s proposition for non-axiomatic photography, the theorization of computational vision may have to start with algorithmic fractality, namely the possibility of a computational auto-impression of the real, exposing the alliance amongst less-than-humans in a field of immanent vision. The Laruelian “Vision-in-One” is a stance that does not predetermine but rather becomes determined by the real in the real’s “last instance.” As much as this determination is a clone of the real, but not the real itself, one can argue that machine vision entails a non-relationality with the world, whose negative auto-impressions are infinite singularizations of blackness beyond the optical value of representation.

That non-photography insists on the cloning of the black universe however is not another way to propose a critique of automation. Instead, this is above all a practice of refusing, hacking, alienating the transcendental decision for which machines as blind can only learn to represent what is already given to them by the ontic limit of knowledge. Similarly, non-philosophy is not simply an invitation for imagining alternative ways to reinstate philosophy and expanding its ontological ground of transcendental decision. Instead this is a heretic project that necessitates trans-collective elaboration of fictional automations that start with the auto-impression of blackness as proliferating each and anytime outside the onto-epistemological cosmogonies of the self and the other. For instance, one can start asking why techno-capital decisionism is still taken as the onto-epistemological law that locks the critique of computation into the perpetual mirroring of vision and knowledge, humans and machines. The fractality of the algorithms can rather contribute to engage randomness in process imaging as machine singularizations of the real. In other words, machine vision can become a medium for the auto-impression of darkness in the last instance: namely as machines learn to learn from negative optics.

The negative materiality of the computational image does not only show that knowledge can be divorced from

transcendental ocularcentrism, but also that the invisible image of machines is part of alien epistemologies that defy the equation of value with the infinities of 0 s. The fractality of algorithms is part of the expansion of heretic epistemologies where alliances between the configurations of the in-human demonstrate how the equation of value can be not only challenged, but turned inside out from the stance of infinities—the auto-impression of blackness. In the following coda, this argument for the algorithmic fractality of infinities that accounts for the invisible images of machines outside ocularcentric critique requires further collective elaborations of how negative optics can become part of a revolutionary practice of machine thinking (that indeed overturns the critique of ocularcentrism within technocapitalism). This coda, in particular, turns to the elaboration of negative infinities in Ferreira da Silva’s articulation of blackness as indeterminate matter at the limit of modern thought, whereby matter without form is above all “matter beyond the equation of value” (2017).

### 3 Coda on images without value

Drawing on quantum mechanics, for which indeterminate results in determining the perception of reality point to the necessity of moving beyond the ontic limits of science, Ferreira da Silva discusses this indeterminacy or the Thing as the referent of blackness, another mode of existing at the limit of modern thought. As she puts it: “deployed as method, blackness fractures the glassy walls of universality understood as formal determination (2017, p 11). She proposes an experiment in articulating the “equation of value” as a self-determining formalism based upon and through the violence against the thing/blackness. Ferreira da Silva proposes to carry out this algebraic experiment not simply to unveil the formal determination of the matter at the core of the equation of value, but one could argue, to rather address the “auto-impression” of blackness minus the form, or matter without the universality of value. To do so, Ferreira da Silva proposes an ethical re-articulation of blackness that does not follow the over-determinant image of Man whose program overfits – that is overrepresents—all modes of being human nor, on the other hand, aims to make a claim for an absolute outside that separates the Modern/European/white human from the non-modern/non-European/black non-human (2017, 04/11).

Ferreira da Silva argues that this ethical experimentation starting from matter beyond the equation of value deserves further investigation about how determinacy, together with separability and sequentiality have sustained modern thought and the construction of an ethical matrix in which the indifference with which racist violence is met, is itself rather become constitutive of a (common and public) moral stance

(2017, p. 04/11). She traces how this matrix is entailed in the modern elaboration of causal efficiency and its operative grounding of equivalence mathematically bounding together ethical, economic and juridical formations (2017, p. 04/11). The modern Kantian world becomes a way to bring together formal and efficient causality in the self-determination of the limits of scientific knowledge—precisely concerning what can already be accessed by the senses (the empirical experience of which science provides the tools for extending universal measure). As value becomes universal and moves across scales, the object (thing/matter) is unified by its formal qualities which in turn are the effects of judgements (and thus transcendental concepts) derived from the measurement and classification of objects (that is by the ontic limits of science). Here the difference is granted by the decisional position or transcendental operator that already knows the object. Ferreira da Silva explains that within this transcendental field of value, blackness as a category of racial difference “occludes the total violence necessary for this expropriation [namely the colonial expropriation], a violence that was authorized by modern juridical forms – namely, colonial domination (conquest, displacement, and settlement) and property (enslavement) (2017, p 08/11). However, Ferreira da Silva’s invocation of the Thing in quantum mechanics, as much as it claims for an autonomy of matter beyond form or universal equivalence of value, also offers a way to unsettle the universal matrix of ethics. The Thing challenges the ontic limit of knowledge imposed on and through science by transcendental philosophy, which saw the extension of the formal into efficient cause driving modern colonial epistemology.

In particular, Ferreira da Silva’s ethical experiment offers us a proof of the “equation of value” for which blackness as nothing—that is zero value or infinity—has the creative capacity to unsettle and hack Modern Western onto-epistemology of vision and knowledge. According to Ferreira da Silva, zero is to be taken not as a contradiction, as established by the dialectic of presence and absence, but itself as a signal of the autonomy (or in Laruelle’s term unilaterality of) negation and of the negative. Instead of being invisible, blackness, as matter without form, brings forward the nullification of the ocularcentric field of vision. But how to demonstrate that such a nullification is not simply another manifestation of the contradiction that rather confirms the norms of the transcendental? This is a question that must accompany the theorization of the negative optics of computational vision because it must bear the challenge against the ontic limits of knowledge that continue to be inscribed within the performance of science of information, whereby the potentiality of machine epistemology is constantly re-subsumed by transcendental decisionism and representational metaphysics. For Ferreira da Silva, it is a question of re-articulating the mathematical mapping of the matrix

of ethics in terms of a radical engagement with zero as a value in itself, rather than self-positing a critique that reveals contradictions in the transcendental equation of value. In as much as the result of Ferreira da Silva's experiment with the equation of value points to the real dimension of the undeterminable (namely value without form, namely neither life nor not-life), it also has zero value because it exists itself without form. In equating blackness with zero or infinity, Ferreira da Silva proposes a radical praxis as a refusal of dialectics founded on the decisional principle of philosophy, whose complains against (and rectification of) the invisibility of the automated image, are there only to re-produce its ocular-logos of recognition (2017). From this standpoint, it is possible to conclude that the dissolution of determinacy and its transcendental decisionism requires not a skeptic critique against computational vision but above all a creative elaboration of the non-ontic science of vision machine that rather enables the possibility of overturning dialectics and its critique. The argument for negative optics, therefore, does not only concern the human–machine relationship but more importantly how this latter, under technocapitalism, can rather be turned inside out to defy the onto-epistemological violence of ocular vision through and with the negation/negativity of infinities, with blackness as matter without form, auto-impressions of the real without the given, auto-imagining minus the self-positing subject. From this standpoint, a praxis of refusal stays with the trouble of dark optics, plunging within multi-layering opacities, fractalizing blackness in artificial visions.

From this standpoint and against the universal model of technology (the techno-logical universalism of ontic knowledge), it seems crucial to continue to elaborate how machine vision is equivalent not only to the perceptual gaze but also to the operational image that has sustained, since modernity, the colonial operative epistemology and its matrix of ethics. Ferreira da Silva's reversed equation of value is a praxis of refusal that does not only reveal the violence of self-determination but also invites us to stay with the negativity of algebraic relations for which nothing is a signal of infinity running against the metaphysical equivalence of knowledge and vision and the capital equation of humans and machines. If we were to invert the equation of value in computational processing one may need to start from the indeterminacy of compression, the negative algorithm of what cannot be known in advance, the negative randomness that condemns the computation to remain incomplete. Since negative optics implies no image-form, so too no universal formalism can grant the outcome of the algorithmic compression of randomness. It is, therefore, possible to argue that negative optics is an index of immanence—namely no transcendental value can explain the auto-impression of images in the algorithmic discretization of increasingly smaller patterns of recognition. Here the networked relation between image,

concept and object is turned upside down as much as the auto-impression of images exceeds the operative correction of images that capital imparts on the human–machine alliance. Instead of the ocular representation of the world, the increasing discretization of the network increases the volume of randomness as much as the ANN affords the ingestion of indeterminate variations within algorithmic compression. In other words, the machine vision also coincides with what cannot be explained, programmed, represented beforehand by concepts programmed in the machine. This is why algorithmic operations are not simply prescribed perceptions but are auto-impressions of the human–machine condition each and any time multiplying the field of auto-apparition of blackness without form in the last instance.

To conclude, it is central to this argument that as much as the negative optics of computational processing speaks of the challenge of reinventing epistemology outside the knowledge-vision correlation, it also requires us to unsettle the colonial roots of modern epistemology for a space of thought for asymmetric auto-impressions of the human–machine alliance. This is a collective and transversal effort in thinking with machines that starts from acknowledging the inhuman condition, the negativity of value, zero value in the human–machine alliance. This is also to say that if the modern universality of technology continues to become reified in machine visions, so too the argument for practices (all forms of practices) starting from the zero value of auto-impression must continue to unpack the heretic versioning of negativity in the technopolitical reconstruction of epistemologies.

## Compliance with ethical standards

**Conflict of interest** Nothing to declare.

## References

- Atanasoski N, Vora K (2019) Surrogate humanity. Race, Robots and technological futures (perverse modernities: a series edited by Jack Halberstam and Lisa Lowe). Duke University Press, Durham and London
- Ferreira da Silva D (2017) “1 (life) ÷ 0 (blackness) = ∞ – ∞ or ∞ / ∞: On Matter Beyond the Equation of Value.” *e-flux* 79, February. <https://www.e-flux.com/journal/79/94686/1-life-0-blackness-or-on-matter-beyond-the-equation-of-value/>
- Galloway AR (2014) Laruelle: against the digital. University of Minnesota Press, Minneapolis
- Giardina E (2019) The cleaning of emotional data. *Aksioma*—Institute for Contemporary Art, Ljubljana. <https://aksioma.org/cleaning-emotional.data/>
- Hinton G et al (2017) Dynamic routing between capsule. In: 31st conference on neural information processing systems, NIPS 2017, Long Beach, CA. <https://arxiv.org/abs/1710.09829>. Accessed 22 Apr 2020
- Laruelle F (2010) The concept of non-photography. Urbanomic, Falmouth

- Laruelle F (2013) The transcendental computer: a non-philosophical utopia. Trans. Taylor Adkins and Chris Eby, *Speculative Heresy*, August 26. <https://speculativeheresy.wordpress.com/2013/08/26/translation-of-laruelles-the-transcendentalcomputer-a-non-philosophical-utopia/>.
- Lowe L (2015) *The intimacies of four continents*. Duke University Press, Durham, NC
- Paglen T (2014) The operational image, *e-flux*. November. <https://www.e-flux.com/journal/59/61130/operational-images/>
- Paglen T (2016) Invisible images (your pictures are looking at you). *The new enquiry*. December. <https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/>
- Steyerl H (2015) A sea of data: apophenia and pattern (mis-)recognition. *e-flux*. April. <https://www.e-flux.com/journal/72/60480/a-sea-of-data-apophenia-and-pattern-mis-recognition/>
- Virilio P (1994) *The vision machine*. Indiana University Press, Bloomington, IN

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Artificial vision, white space and racial surveillance capitalism

Nicholas Mirzoeff<sup>1</sup>

Received: 29 July 2020 / Accepted: 14 October 2020 / Published online: 6 November 2020  
© Springer-Verlag London Ltd., part of Springer Nature 2020

## Abstract

This first half of the paper outlines the formation of racial surveillance capitalism across the *longue durée* of settler colonialism, with special attention to the formation of artificial vision. This artificial vision is deployed in the erased territory, creating a white space in which to see from platforms, ranging from the ship, to the train and today's drones. The second section examines the Eurodac digital fingerprint database created by the European Union to monitor and control asylum seekers and refugees as an "artificial life system," to use a phrase coined by its administrators. In this automated form, artificial vision is distributed rather than centralized.

**Keywords** Artificial vision · Surveillance · Database · Eurodac · Leviathan · The state · Race · Racialization · Whiteness

"This is how to place you in the space in which to see."

Layli Long Soldier, *Whereas*

In discussing "relations between the conqueror and the colonized" in his *Ways of Seeing*, John Berger made a line drawing depicting in barest outline two figures. The one on the right was captioned "omnipotent" and the one to the left "less than human." Berger noted: "the way each sees the other confirms his own view of himself" (1972: 96). Two pairs of diagonals go from eyes to feet and eyes to the top of the facing figure's head, perhaps evoking Hegel or Lacan. There's much left unsaid here. Did the conquered actually think of themselves as less than human? Or were they confirmed in seeing that the conqueror saw them that way? "Seeing comes before words," as Berger had famously begun his book. Before seeing comes "the space in which to see," to borrow a phrase from Oglala Lakota poet Layli Long Soldier (2017: 8). The way of "seeing" that arises in the space in which to see erases so as to produce white space, which can then be claimed for absolute ownership. This seeing-in-space is the sensing of how to place people in relations of hierarchy to extract value. The formation of white space in which to see, by people and machines, is my subject here. This white space is the product of coloniality, a space formed by the erasure of existing human and

other-than-human relations. Coloniality is the time-space since 1492 when "America [the hemisphere] was constituted as the first space/time of a new model of power" (Quijano 2000: 533). In the space of erasure, artificial vision and artificial surveillance are enabled. Deployed first from infrastructure platforms like ships or trains, these processes are now being distributed into a network of machines that form artificial life systems. Together, the combination of erasure, extraction and surveillance has enabled racial surveillance capitalism to survive in that white space from the overseer on the plantation to neocolonial domination by unmanned aerial vehicle (UAV). What's different today is that this artificial vision is now being automated and distributed, creating spaces of disappearance.

The operations of white space precede what is conventionally thought of as "seeing," a look directed by a person at an object or other person that necessarily takes place in space. Whiteness is the apex, the place of organizing, and the vanishing point to and from which "seeing" is directed under racial surveillance capitalism. White space is rendered by the systemic erasure of colonized terrain and existing social relations in that space. The erased ground made space perceptible to the "conquering" gaze identified by Berger. This process involved a multiplicity of senses from touch to vision and sound because (colonial) ground is layered and folded, as Stuart Hall has defined it: "it is the site irretrievably marked in relation to the question of 'origin' by an unpassable distance" (2017: 168). The resulting white space is at once static, responsive to input, and cultivates transformation. This static is the presence of the state, meant to be

✉ Nicholas Mirzoeff  
nm45@nyu.edu

<sup>1</sup> Department of Media, Culture and Communication, New York University, New York, USA

permanent and unchanging; the statue as a symbolic figure for the state; and the electric noise generated by surveillance apparatus. White space is always a surveillance arena and so it responds, if and when it detects something or someone within what Jacqueline Rose has called the “field of vision” (1986). Elsewhere, I have called this regime “oversight,” meaning the work done by the overseer on the plantation to ensure maximum production and minimum resistance and projected forward into the still-continuing “plantation futures” of the Atlantic world (Mirzoeff 2011: 48–76; McKittrick 2013: 1–12). If there is always a “weave of differences” (Derrida 1984: 13) in human identification, the frame on which that weave is produced is, under the existing regime of coloniality, whiteness-as-white-supremacy, whether that frame is a picture frame, a mainframe, or a container for network packets. These frames are not identical or self-identical but contain and produce whiteness as “a changing same,” to borrow Paul Gilroy’s formula (1993: 72–110). In whiteness’s own imaginary, to be seen in white space is to be subject to violence without redress.

White space sustains whiteness as the “changing same” of what Caribbean philosopher Sylvia Wynter has called “monohumanism.” For Wynter, monohumanism acts “as if it were the being of being human” (2015: 199 n.22). It is a system of violent domination, enacted by means of visualized distinction leading to separation, whether or “races” or of the free and enslaved, and the consequent production of vulnerability to harm. As a way of seeing, monohumanism uses a monocular vision of the world as a grid, shaping, in turn, the square of plantation agriculture, the layout of imperial cities, and now the patterns of electronic surveillance. The combination of monocular vision with the enforcement of monohumanism forms what I call “racial surveillance capitalism.” As a concept, it connects Cedric Robinson’s racial capitalism, with the recent upsurge in awareness of surveillance as constitutive of capitalism, by way of Simone Browne’s concept of “the racialized disciplinary society” (Robinson 2000: 200–30; Browne 2015: 9; Zuboff 2019). Far from being a “rogue mutation of capitalism” (Zuboff 2019: ix), racial surveillance capitalism has been active since the surveillance-dominated grid cities of sixteenth-century Spanish Mexico organized on the principle of “concentration” (Nemser 2009: esp. 25–64); to the slavery-era plantation with its overseer; the factory with its foreman; the “new Jim Crow” of mass incarceration; the “carceral reservation world” (Estes 2019: 115) invented by settler colonialism for the indigenous; and today’s CCTV-controlled megacities on quarantine lockdown. Assertions that “surveillance capitalism is young” (Zuboff 2020) fail to account for its long role in generating and sustaining racial surveillance capitalism on stolen land in the plantation and the factory. Sustaining racialized hierarchy is and was codependent with the extraction of value by means of persistent

surveillance of those excluded from monohumanism. State-gathered racialized “intelligence” is now being formulated into facial recognition, unmanned aerial vehicles (UAV) or drones, and border identification technologies, all still seeking an automated version of the perfect surveillance desired by the plantation overseer.

## 1 Erasure

In the Americas, racial surveillance capitalism began with the “clearing” of ground, mentally and physically, and by displacing or disposing of the Indigenous. This clearance continues with the assertion that Indigenous peoples are “extinct,” or that their claims to land are void. When the Indigenous within the borders claimed by the United States combined to protest the extension of the Dakota Access Pipeline into land designated as Lakota by the 1851 and 1868 Fort Laramie Treaties, they were met with violence, from state police with sticks, to tear gas, and presidential Executive Order. LaDonna Bravebull Allard was quite clear as to what was happening: “Erasing our footprint from the world erases us as a people. These sites must be protected, or our world will end; it is that simple” (TallBear 2019, in Estes and Dhillon 2019: 17). By replacing the Lakota world with a pipeline, that world was erased, at least in part. That erasure continues: for example, Tohono O’onham burial grounds were demolished in 2020 to make way for the US border wall with Mexico (Ruiz 2020). The 2020 Land Defenders on Wet’suwet’en land within the borders claimed by Canada met a similar response. As Freda Hudson, spokesperson for the Unist’ot’en (one of the clans of the Wet’suwet’en nation) put it, the issue revolved around competing definitions: “for us, our critical infrastructure is the clean drinking water and the very water that the salmon spawn in ... to them, they massively clear cut land, which the animals depend on” (Spice 2019: 215). When settler colonialism directly confronts its others, the issue is stark: erasure or survival.

Layli Long Soldier’s poem “Three” (within the section *He Sapa*, known to settlers as the Black Hills, treaty-protected land sacred to the Lakota Sioux) visualizes this space of encounter and confrontation: “This how you see me the space in which to place me/The space in me you see is this space/To see this space see how you place me in you/This is how to place you in the space in which to see.” Seeing is spaced and placed, you and me, unevenly. Here, I, the non-indigenous white settler, am the “you” of her poem. If one begins at the top, it begins with the settler placing the Indigenous in any space whatever. The sentence is evenly and standardly spaced. The next two lines have spaces between phrases. Later in the collection, she terms this unreadable space a “white hole,” which results in letterpress when two or more spaces are used, whether by accident or design



(Long Soldier 2017: 71). Those “holes make the space open,” allowing you, me and them to enter, as and when. Across these spaces, a form of relation occurs. By the last sentence, evenly spaced, the settler may, with due process, become able to access the space in which to see. The poem as a whole produces a square blank space in the middle of the page, formed by these unequal ways of seeing. Across her section “Whereas,” Long Soldier has many names for “it,” that unreadable but perceivable and knowable space that rhymes without sound. It is “Indian emptiness” that she notes the Oxford English Dictionary now says must be rendered as “American Indian emptiness” (Long Soldier 2017: 62). The colonizer still controls the sentence.

Indeed, *Whereas*, the title of her book, follows from the 2009 Congressional Resolution of Apology to Native Americans, which was placed entirely in the “whereas” clauses of the 2010 Defense Appropriations Act, providing over \$500 billion for the military (Public Law 2009). That is to say, the Act articulates, in the sense offered by Stuart Hall, the foundational erasure of Indigenous peoples in the Americas with present-day neo-colonial ventures in Afghanistan, Iraq and elsewhere. This articulation of what settler colonialism does amplifies Long Soldier’s statement that the Apology “falls short of legal grounds” (2017: 70). Yet, like Fanon in *Wretched of the Earth*, Long Soldier dreams of running, a dream of embodied and decolonized freedom. In her poem, on waking she “teeter[s]” to the mirror, saying “*You’re old enough now to look at yourself full-on.*” (2017: 70; her emphasis). Long Soldier comes into her own view, deferring and making different the experience of colonized ground. For Long Soldier, the result is “defiance,” not the deconstructive *différance* (2017: 75).<sup>1</sup> It is not a “gaze,” because that is what the settler does and automates in the surveillance system. It is a full-on look, one that expresses majority, in the sense of legal subjectivity and of maturity. Being able to look at yourself full-on counts. Thinking about the colonizing state, Kahnawà:ke Mohawk scholar Audra Simpson has identified in the sovereign a “death drive to eliminate, contain, hide and in other ways ‘disappear’ what fundamentally challenges i[t]s legitimacy” (Simpson 2016: n.p.). This visualized structure of domination by concealment and disappearance is a powerful formation of racialized surveillance. It is both deployed against individuals and has a collective set of outcomes. For Simpson, “the state is a man,” and a “heteropatriarchal” man at that, for, as Susan Deer has argued, rape “can be employed as a metaphor for the entire concept of colonialism” (2015: xvii; in Estes 2019: 81). The death drive of disappearance renders what

Simpson calls “Indigenous political orders” unappearing within white space.

The sovereign monocular stare of erasure, disappearance and death is an active, artificial and engaged form that seeks to conceal itself from those it observes. The state and its surrogates project<sup>2</sup> themselves onto the erased white space of plantation futures, to use the term coined by Kathleen McKittrick, meaning “a conceptualization of time–space that tracks the plantation toward the prison and the impoverished and destroyed city sectors.” To these sectors should be added the so-called “reservation” for the Indigenous as and the detention center for migrants and refugees. In the plantation imaginary, the overseer was capable of envisaging everything that took place in and around the plantation, keeping humans, animals, biomass and even landscape under transformative surveillance. McKittrick shifts the register of the plantation as past time to one in which “the plantation uncovers a logic that emerges in the present and folds over to repeat itself anew” (2013: 2, 4). In this case, that logic is the means by which plantation oversight continues to structure the automated systems of racial surveillance capitalism.

White space subsequently metamorphoses a person into a commodity. It transubstantiates life into value or renders life into data. White space is always moving image space, where there is not simply motion, but alteration. This logic rendered life into the property, the process of enslavement by which a body becomes an object according to colonial law, but so too does whiteness become property. From that violence results a chain of metamorphoses, as Walter Johnson has put it, from “lashes into labor into bales into dollars into pounds sterling” (2013: 244). In formal economic language, the later stages of this process are usually considered as an exchange, but while dollars can be exchanged for cotton and other things, lashes and their resulting pain cannot (or should not) be exchanged at all. But Johnson notes that “violence is the metric of production,” to which I would add in this context, “of white space.” Under digital systems of surveillance and detection, life is rendered into data, creating a “hostile environment” to appropriate former British prime minister David Cameron’s nasty tag. A person within the hostile environment is subject to physical violence, ranging from arrest and detention to deportation, but is pressured constantly by the awareness of being considered a suspect. The digital form of the “wages of whiteness” identified by W. E. B. Du Bois in 1935 is not access to the water fountain but to the nation-state as a whole, a hostile white space designed to exclude.

<sup>1</sup> The term *différance* is a neologism coined by Jacques Derrida to express the contradiction that the French verb *différer* means both to defer and to differ.

<sup>2</sup> For the concept of projection in colonial contexts, see Casid, *Scenes of Projection* (2015: 89–124).

## 2 Artificial surveillance: Leviathan

While white space is co-eval with coloniality, by the time of the great acceleration of sugar production in the seventeenth century that fueled the rise of modern racial capitalism, it became subject to artificial surveillance. This surveillance was the combination of an artificial way of seeing and the compound formation of the state as an artificial machine. The first figure of this “artificial intelligence” that formed the plantation futures in which racial capitalism continues to operate was the Leviathan, the imaginary means of visualizing the state. For Thomas Hobbes, the Leviathan was an “artificial man”: “the sovereignty is an artificial soul, as giving life and motion to the whole body” (1651: 7).<sup>3</sup> For Hobbes, the sovereign is also sovereignty, a far from the neutral term, as Hardt and Negri have shown: “the concept of sovereignty emerges from the colonial mentality and is conceived explicitly in relation to the natives of ‘America,’ who are considered populations that remain in the state of nature” (Hardt and Negri 2017: 26). Hobbes’ sovereign representation expressed the racialized hierarchical imaginary of monohumanism derived from the plantation colony. In *Leviathan*, all (white) persons were imagined to voluntarily give up their freedom in exchange for the protection of the state. Hobbes declared “[a] Multitude of men, are made One Person, when they are by one man, or one Person, Represented” (in Skinner 2018: 284). The Leviathan “represents” the “multitude” as being contained in one artificial man.

This artificial state deployed an artificial vision, visualized in the period by Abraham Bosse, first professor of perspective at the nascent Academy of Painting in Paris, the artist who would later draw *Leviathan*. Bosse was quite literally a bourgeois revolutionary—a Huguenot, he lived in Paris and participated in the Fronde (1648–53), the uprising against the young Louis XIV—who understood perspective to be the dominant form of representation. Bosse depicted vision as a pyramid formed by four pieces of string ending in the eye (singular) of white soldiers carrying swords. The object being looked at wills its own surveillance, according to Bosse, because it emitted rays that entered the eye to be seen. In a similar fashion, Hobbes began *Leviathan* with a discussion of seeing, in which he distinguished between the object and the way it was seen that resulted from “pressure,” which is to say, “the motion of external things upon our eyes.” In Bosse’s drawing, the soldiers don’t have visible eyes. The observed cannot look back at the sovereign stare because it is eyeless, such as Samson. This visualizing of vision is, to borrow philosopher of vision Susanna Berger’s term, itself entirely “artificial” (2017: 184). It has nothing

to do, then or now, with seeing as a physical process and everything to do with controlling appearance in the field of vision. The square formed within the highly abstracted space depicted by Bosse is visibly white space. Whatever was there before—people, cultures, other-than-human life—has been erased.

Bosse’s diagram renders in miniature the triangulation of land as colonizing and visualizing technology. Land titles known as “plats” in the period depicted terrain as a line drawing as seen from above. In practice, measurements were made by enslaved labor using a unit called a “chain,” invented by the English priest Edmund Gunter in 1620. The chain varied in length (in British imperial measures it was 66 feet) and was measured with a metal chain, the material embodiment of white space. The chain remains the length of a cricket pitch and it was the basis of the famous Ordnance Survey maps of Great Britain. This detail of measurement epitomizes the formation of white space: an entirely artificial measure nonetheless amply expresses the realities of coloniality from conquest to enslavement. Bosse’s diagram appeared in *Ways of Seeing* as an illustration of perspective, a system Berger described as enabling “appearances [to] travel in [to the body]. The conventions called those appearances *reality*” (1972: 17). The colonial reality being formed by the artificial white machine at the intersection of the artificial state apparatus and artificial seeing was white space. Erased rather than empty, it erased life to extract value.

The artificial state and artificial vision came together in the famous engraving made by Bosse to depict Leviathan as a sea monster from the Book of Job, emerging from its element as a single figure containing all its subjects. As Horst Bredekamp reminds us, Hobbes imagined the Leviathan as a “mortal god,” a figure equivalent to Hercules and other creatures of legend (Bredekamp 2007: 33). Hobbes saw the formation of such “compound creatures” as he called them, as a special instance of the power of colonial imagination, or what he called “Fancy.” Fancy was not simply an artistic or creative attribute: “whatsoever distinguisheth the civility of *Europe*, from the Barbarity of the *American* savages, is the workmanship of *Fancy*.” Fancy created images, meaning “any representation of one thing by another” (in Tralau 2007: 65–9). Leviathan is, then, the image of sovereign colonial authority in and as the power to represent. With that in mind, it becomes clear that Leviathan is, in fact, emerging out of the sea, with his legs as yet underwater. Bosse drew the sea, complete with a ship, at the extreme right—often cut out of reproductions—together with a typical colonial fort. The fort is the prototype of Fortress Europe, the anti-migrant regime of the present-day European Union. Leviathan is in and out of the water, partly immersed: it was and is both a sea creature and a technology—a ship. The ship must always be above and below water. Below the waterline were those to be enslaved, invisible and insensible to those

<sup>3</sup> Prepared for the McMaster University Archive of the History of Economic Thought, by Rod Hay; see Berger 2017: 188).

above, but nonetheless an active and indispensable part of the Leviathan.

### 3 Platforms: ships, trains, and drones

#### 3.1 The slavers' Pigeon-hole

Leviathan as the colonial state may be represented, to use Hobbes' term, as a slave ship: divided, compartmentalized, Manichean. The Manichean formation of whiteness departs, in the sense of sets sail, from this division. There are those in the hold, the means by which Africans were transformed into "slaves" via the monstrous agency of the Middle Passage (Sharpe 2017: 27; Moten and Harney 2013: 93). Where did the white person in authority come to "see" slavery? If the Africans were consigned to the "hold" (a deck below the main deck), then it would be the "deck," from which slaver officers and crew ordinarily sustained the operations of slave trading. Managing a ship was always a question of interactive sensory labor. Visual observation from the crow's nests in the masts was relayed to the deck. The visual perception of sea conditions and nearby land had to be supplemented with logged observation of wind and currents. When close to land, a ship would be "sounding," meaning the measurement of depth by throwing a weighted rope overboard from the bow. Using a variety of visual markers to indicate different measurements, the crew would then call out the depth in fathoms (five and a half feet for merchant ships, six feet for warships). The deck offered a multidimensional and multisensory field of vision. It was nonetheless highly vulnerable. Africans managed to wreck ships during the Middle Passage and the vagaries of weather, wind and currents did for many more.

No case of such disorientation has received more attention than that of the *Zong* (Walvin 2011). This infamous history concerns the slave ship of that name, which, getting lost on the Middle Passage in 1781 overshot its destination in Jamaica, and contrived to jettison no less than 132 Africans. Whether this casting overboard was done to "save" water for the remaining crew and captives, or simply to make a claim for insurance, was the subject of repeated court cases in the period and continues to resonate across historical and creative accounts. Such jettison—a term meaning precisely the throwing overboard of goods to preserve the vessel—was not uncommon, even of so-called human property. The *Zong* is remembered because the formerly-enslaved writer and abolitionist Olaudah Equiano took up the case and pushed it to white abolitionist Granville Sharp. Many believe that it also inspired J.M.W. Turner's painting *Slavers Throwing Overboard the Dead and Dying, Typhoon Coming On* (1840). This devastating painting was made just after Britain had finally abolished slavery, whether as a final indictment, or

as a provocation to consider that abolition had not yet fully taken place. As Christina Sharpe has pointed out, "[t]hat Turner's slave ship lacks a proper name allows it to stand in for every slave ship" (2017: 36). It might have been the *Zong*. Or it could have been the *Leão*, a Portuguese ship that took onboard 855 Africans in 1836 and was known to have thrown some 30 people infected with smallpox overboard, while a further 253 died of measles. The *Leão* indicates that Middle Passage conditions were arguably at their worst toward the end of the slave trade, as captains packed their ships with their now illegal human property (Graden 2014: 62–63; Voyage ID 1586 on slavevoyages.org).

Turner does not make the guilty jettisoners visible, unlike a widely-reproduced abolitionist print from 1833 showing the crew throwing Africans off the deck. Placed "high" in the picture, its single jib shredded by the wind, the slavers have only the quarter-deck (a raised platform behind the main mast from where the captain directed the ship), invisible as it rides the waves, as their place from which to see. The slavers are framed against the purples and oranges of the setting sun, the Manichean condition of slaving. Turner's suggestion is clear: the slaver is a component of the slave ship (often known as a slaver), in the same way that for Hobbes sovereignty could not be distinguished from the sovereign. The slaver was a platform to sustain a specific artificial way of seeing in racial surveillance capitalism, the view from the deck. It created powers over life in ways that could not previously have been imagined. If the main deck was the overall place from which to see, its specific vantage point was the quarterdeck. I think here of M. NourbeSe Philip's astonishing poetry collection *Zong!* (Philip 2008). Like Long Soldier, Philip makes extensive use of the white space of the page and creates white holes by use of extra spaces to visualize not just the action of jettison but its affects. These white holes were also known as pigeonholes, a term that could equally refer to a small hole for looking through, a hole in a ship through which rigging would pass, or the holes in which a person's hands were restrained during the flogging. The entire Atlantic world is there in this expression. It could equally be appropriated and reversed, as when Harriet Jacobs created what she called a "loophole of retreat" to see out of the attic where she evaded enslavement.

#### 3.2 The platform of skulls

The deck was a multisensory platform for the visualization of white space in the era of colonization and Atlantic slavery. The internal colonization of the United States was enabled from the platform provided by the train, involving the mass slaughter of human and other-than-human life. Today's platforms contain traces of the decks and platforms that preceded them, like layers in a Photoshop image, always produced, of course, against a background of white space.

Kul Wicasa from the Lower Brule Sioux Tribe scholar Nick Estes estimates that even as “buffalo-hunting [Indian] nations on the Northern Plains from 1780 to 1877 experienced a 40% population decline,” between 10 and 15 million buffalo were exterminated in the last two decades of that time (Estes 2019: 86, 110). To supplement the work of the US Army in both these genocides, the newly completed Transcontinental Railroad was used as a platform from which to kill. An article from *Harpers Weekly* of the period describes how: “The train is ‘slowed’ to a rate of speed about equal to that of the herd; the passengers get out fire-arms which are provided for the defense of the train against the Indians, and open from the windows and platforms of the cars, a fire that resembles a brisk skirmish” (King 2017). The train has long been understood as a paradigm for industrial modernity, with the corollary that the first “moving image” was the view as seen from the train (Schivelbusch 1987; Mirzoeff 2016: 125–41). These trains put themselves in synch with buffalo life in order to kill them, using weapons first supplied to kill Indigenous peoples. Far from being an amusement or attraction, as early cinema studies have often had it, the moving image was first a scene of genocide of human and other-than-human life.

A vivid example of the intersection of the attraction with the elimination of the buffalo can be seen in a now-widely circulated photograph of a massive arrangement of buffalo skulls. The photograph was taken at a glueworks in Rougeville, Michigan, by an unknown photographer, probably in the mid-1870s (Anon n.d.).<sup>4</sup> The skulls were collected to be rendered into fertilizer so that the death of Indigenous animal life could enable settler agriculture. Two men stand at the top and bottom of the structure, allowing us to estimate that the pile is some twenty-five feet high. The photograph is a depiction of the settler-colonial conquest and racial hierarchy in material form. It’s also a stunt, making the viewer ask how the man on top can stand on such a pile. Looking closely, the grotesque *memento mori* appears to be below ground level, filling in a trench, possibly a railroad cutting by which the skulls had been delivered. The feet of the man at the top cannot be seen, concealed by a skull he is carrying. Perhaps he is standing on the top of the cutting. Or he is standing on a stack of metal cages, like the one in the foreground in which the skulls must have been transported. The factory went to extravagant lengths to create the illusion of a freestanding mountain of skulls, making Indigenous death into an attraction and a spectacle. Estes showed this photograph in a lecture to illustrate his thesis, following

Simpson, that there is a “settler culture of death.” In filmmaker Arthur Jafa’s installation of the photograph in his 2019 Prague exhibit entitled “A Series of Utterly Improbable Yet Extraordinary Renditions,” there was no explicit commentary. The title clearly referenced the term used by the Bush administration to take suspects to remote “black sites,” where they were subject to illegal torture. Enlarged to life size, the photograph wrapped around a corner of the Galerie Rudolfinum, built during the Austro-Hungarian empire, and was placed next to Jafa’s artwork *Black Flag*. In staging this literally intersectional oppression, Jafa wanted us to know both that “Love is the message” and that “the message is death.” For Jafa, his film of that title addresses what it means “to be alive and not to be alive at the same time.” Such is the message from these skulls, the materialization of white space from the platform of the train.

### 3.3 Kill box

In the past fifty years, the moving-image platform for the surveillance of white space has become automated. The unmanned aerial vehicle (UAV), or drone, has made such surveillance persistent and pervasive, from the US-Mexico border to post-9/11 low-intensity counterinsurgency warfare. It operates as “lawfare” blending law and warfare into a single word, now a term of art in the military (Hajjar 2017: 59–88). Its continuity with earlier forms of racialized surveillance produces what Keith Feldman has called “racialization from above” (Feldman 2011: 325–41). Over this period, UAV “lawfare” has created an abstract space of death, formally known as a “kill box.” It was first developed by the Israeli Defense Force as a means of monitoring the Palestinians under occupation in the 1970 and 80s,<sup>5</sup> which has every more clearly become a means of erasure and disappearance. The visualization of the kill box made by the IDF is strikingly reminiscent of Bosse’s white space in which to see. Both diagrams show a viewing point—whether the eye or the drone—and a pyramid of lines forming an abstract white square. The “kill box” is what it sounds like: an arbitrary square area drawn from the drone as if it were a single point, in which the drone operator is given permission to kill those who become visible. It is a plantation future of the overseer’s visual footprint over his “plat,” a line drawing of an estate as if from an aerial viewpoint produced by (inaccurate) surveying to distinguish one colonized piece of land from another. Everything in that plat was to be seen by the overseer. The triangulation of land as visualizing technology folds into the present in automated form. The “kill box” is now part of US military doctrine since

<sup>4</sup> The National Museum of the American Indian dates the photograph to the mid-1870s, although the text on reverse reads: “C.D. 1892 Glueworks, office foot of 1st St., works at Rougeville, Mich.” Burton Historical Collection, Detroit Public Library.

<sup>5</sup> For this history, see Chamayou (2015: 27–28). For the first kill box theory, see Lee (2019).

the post-9/11 ventures into neo-colonialism. Accordingly, it has its own Field Manual and, like the plantation, has evolved its own bureaucracy, requiring a form to be filled out for each requested kill box. The Field Manual defines a kill box as “a three-dimensional area used to facilitate the integration of joint fires.... When established, the primary purpose of a kill box is to allow lethal attack against surface targets without further coordination with the establishing commander” ([US] Army, Marine Corps, Navy, Air Force 2009: 1). It has a ceiling to prevent friendly fire accidents and is sufficiently high-resolution to have “no fire areas” within it. These boxes are not imaginary: people die as a result of them and in them. These systems are boundlessly expensive. The eighty-six weapons systems examined by the Government Accountability Office in 2012 were estimated to cost \$1.6 trillion over their lifetime. The three major UAV systems were projected to cost over \$30 billion (Defense Acquisitions 2013: 101). They are funded for good reason because, as media scholar Lisa Parks puts it, the formation of the drone view of the world has “the potential to materially alter or affect the phenomena of the air, spectrum, and/or ground” (Parks 2017: 135). Or, in the terms I’m using here, white machines transform the land into white spaces in which to see is to kill.

#### 4 Artificial life systems

Racial surveillance capitalism’s artificial forms are manifest throughout biometric and algorithmically-automated systems. The intersection of long-standing technologies of monohumanism with digitized forms of artificial intelligence is producing new forms of artificial life. Simone Browne calls for a “critical biometric consciousness” to respond to the blanket application of biometry by states (2015: 116). Browne sets facial recognition and biometry in the long context of the “technology of tracking blackness that sought to make certain bodies legible as property” (2015: 128). Indeed, as Joseph Pugliese has put it, “biometric technologies are *infrastructurally calibrated to whiteness*” (2007: 107). IT equipment, like servers in the data centers that produce cloud computing, is actually known to designers as “white space.” Technicians design these centers in pure white supposedly to make any dirt more visible but also in accord with what A.R.E. Taylor calls the “techno-aesthetics” of cleanliness and purity (2017: 49). These aesthetics are also, consciously or not, those of eugenic white supremacy. The result is what has been described as the “diversity disaster that now reaches across the entire AI sector” (West et al. 2019). As Ruha Benjamin has put it, “discriminatory design” has produced nothing less than the “New Jim Code” (2019: 3). It is an intended outcome, not an accident, and it is not proving simple to eradicate (See Joh 2016: 15–42;

Marx 2016; Hao 2019). Based as it is on “epidermalization” (the assertion of absolute difference based on relative differences in skin color), AI’s racial surveillance deploys an all-too-familiar racialized way of seeing operating at planetary scale. It is the plantation future we are now living in. All such operations take place in and via the new imagined white space of technology known as the cloud. In reality, a very material arrangement of servers and cables, the cloud is both an engine of high-return low-employment capitalism and one of the prime drivers of carbon emissions. On the one hand, Amazon Web Services—one of the largest cloud operations—have become the greatest source of profit for the company (Condon 2018). On the other, if taken together, data centers formed the fifth highest source of carbon emissions as early as 2012, according to Greenpeace (Hu 2015: 79). It takes a million times the amount of energy to store a document in the cloud than it would on a hard-drive (Adamson 2017).

Amidst the fake cleanliness of the cloud, the refugee has (re)appeared as the key figure of a database-driven reconfiguration of the carceral nation state to prevent migration and asylum. While the US is obsessed with its archaic wall, key to this process in Europe is the Eurodac fingerprint database, created by the European Union to enforce its Dublin Regulation, which stipulates that all must seek asylum in the country in which their fingerprints were first taken. Eurodac is the distributed form of racial surveillance capitalism. It “sees” the migrant and registers them not as people but as a biometric data set. Eurodac’s administrative body—*European Agency for the operational management of large-scale IT systems in the area of freedom, security and justice*—considers it to be “a living system” (2013: 4). This suggests Eurodac is not just an AI, it is an artificial life form. In 2017, its four year budget was set at €29.8 billion, equivalent in cost to the UAV programs of the US military.<sup>6</sup> Its software was designed by Cogent, later acquired by Gemalto, who were acquired in turn by Thales for €4.8 billion to become part of a €19 billion global security company.

Digitized fingerprints are now the prime mover of the refugee system in Fortress Europe, the direct descendant of Leviathan. The fingerprinting machine has become the border, since asylum seekers are required to apply for asylum wherever they have been fingerprinted. Since 2003, Eurodac has recorded the fingerprints of asylum seekers to enable “fingerprint comparison evidence” as an automatic identification system for administering claims.<sup>7</sup> Since 2015,

<sup>6</sup> See <https://www.europarl.europa.eu/legislative-train/theme-toward-a-new-policy-on-migration/file-jd-recast-eurodac-regulation>.

<sup>7</sup> “By comparing fingerprints, Member States can determine whether an asylum applicant or a foreign national, who is suspected to be illegally present within a Member State, has previously claimed asylum in another Member State or whether an asylum applicant entered the Union territory unlawfully.” *Framework Contract for Maintenance in*

it also allows law enforcement to do cross-checking against criminal and “terrorist” databases, a consistent pattern in the EU. Every major immigration and asylum database that was set up with a firewall to law enforcement routinely grants them access within a short period of time (Hayes 2017: 185 n. 21). With a capacity to hold 7 million records,<sup>8</sup> Eurodac already stored 4 million records by 2016, which it keeps for a decade.<sup>9</sup> In that year, 1.6 million fingerprints were taken but the system only generated “hits” in 16% of cases, which suggests its function is more to deter than detect.<sup>10</sup> While some studies of the supposed “gold standard” of European data protection and asylum regulation criticize the “thoughtlessness” (Bugge 2019: 91–100, 94) or ambiguity of these practices, these are intentional failures, designed to produce exclusion rather than consistent and equal treatment. As Benjamin Muller has put it, the results have been “pre-emptory logics, a negligent attitude towards ‘false positives,’ and an overall proliferation of borders” (2010: 9). More precisely still, it is a racialized border.

In the nineteenth century, Francis Galton, the founder of eugenics, adopted the fingerprinting system of identification from colonial India, where it was used because officials could not tell people apart. Gemalto, the software company that administers the EU fingerprint system, acknowledges this genealogy on its website and even quotes Galton on the alleged accuracy of the fingerprint.<sup>11</sup> For despite the CSI imaginary in which forensics are always already completely accurate, fingerprinting is no more than “probabilistic,” as Browne puts it, noting cases of false identification by fingerprints. Galton also coined the term “biometry” in 1901, as a militarized state apparatus, capable of “converting a mob [of statistics] into an orderly array, which like a regiment thenceforth becomes a tactical unit” (Galton 1901: 7–10).<sup>12</sup> This conversion would allow for the perception

of “incipient changes in evolution, which are too small to be otherwise apparent.” It was precisely such alleged variations in evolution that eugenics was intended to eliminate. Immense caution should be used before reviving such toxic lines of thought, yet the regulatory regime is vague and easily evaded by state actors (Browne 2015: 111).<sup>13</sup>

If Eurodac is alive, it is a privatized zombie designed by Kafka. An asylum seeker incarcerated in Denmark described it accordingly: “you are in a room trying to get out, and it’s like a labyrinth to get out of there, with lots of different corridors to take and you don’t know which one” (Freedom of Movements Research Collective 2018: 17). Following what turned out to be the high point of migration to Europe in 2015, Eurodac was enhanced so as “to take fingerprints and an additional biometric identifier, namely a facial image. Far from making the system more reliable, researchers have shown that existing AI consistently interprets Black faces as “angry” or “contemptuous” from a study using the publicity photographs of US basketball players (Rhue 2018). The new Eurodac regulations also lowered the age of taking fingerprints from 14 to 6 years old.<sup>14</sup> Are 6-year-olds terrorists? Perhaps, says Eurodac, because they may have acted “irregularly” in crossing internal EU frontiers or overstaying visas. As a result, legal experts are now debating whether school photographs are private information or “public” for state use (Kindt 2018: 523–538, 534). A German stock photo available to illustrate media reporting on the system clearly shows the Eurodac imaginary. A hand clad in a medical glove to protect against infection presses another, visibly brown hand onto the scanner. The connotations are that brown people are a viral source of contamination and cannot be distinguished except by machines. Artificial vision is no longer analogous to human vision: its purported digital capacity to distinguish people exceeds (white) human capabilities.

The E.U., in fact, bans the use of biometric data for identification via its 2016 General Data Protection Regulation (GDPR).<sup>15</sup> The European Court of Human Rights has ruled that the retention of biometric data is an intrusion on the right to privacy, with exemptions, such as when “explicit consent” has been given. Does setting foot in a public space monitored by CCTV constitute such consent? Perhaps a court might agree. Very few individuals would, I suspect. If the court does not agree with the blanket provision, authorities can claim “substantial public interest” to

Footnote 7 (continued)

Working Order (MWO) for the EURODAC system LISA/2016/RP02 (Restricted Procedure—Article 104 (1) (b) Financial Regulation, Article 127 (2) paragraph 2 Rules of Application). Available at <https://www.eulisa.europa.eu/Procurement/Tenders/LISA2016RP02%20EURODAC%20MWO/Annex%20I%20Eurodac%20MWO-Executive%20Summary.pdf>.

<sup>8</sup> Using Dell machines, [https://www.europarl.europa.eu/doceo/document/E-8-2018-001595-ASW\\_EN.html?redirect](https://www.europarl.europa.eu/doceo/document/E-8-2018-001595-ASW_EN.html?redirect).

<sup>9</sup> Annual report on the 2016 activities of the Eurodac central system, including its technical functioning and security pursuant to Article 40(1) of Regulation (EU) No 603/2013 (European Agency for the operational management of large-scale IT systems in the area of freedom, security and justice (eu-LISA), 2017), 4. Available at <https://eulisa.europa.eu>.

<sup>10</sup> <https://www.gemalto.com/govt/customer-cases/eurodac>.

<sup>11</sup> <https://www.gemalto.com/govt/customer-cases/eurodac>.

<sup>12</sup> For a case study of Galtonism and racialized state power, see Breckenridge (2014).

<sup>13</sup> See section “Branding Biometrics” (Browne 2015: 109–18) for full exploration of this issue.

<sup>14</sup> [https://ec.europa.eu/home-affairs/what-we-do/policies/asylum/identification-of-applicants\\_en](https://ec.europa.eu/home-affairs/what-we-do/policies/asylum/identification-of-applicants_en).

<sup>15</sup> Regulation of the European Parliament on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive 95/46/EC, 2016/679/EU, 27 April 2016.

justify using biometric data. The merest whiff of “security” is almost always taken as such justification that is not formally defined. Many of these protocols can be bypassed because the EU has decided that fingerprints and facial images (or any such indicator) are not biometric data until they undergo “specific technical processing” (Kindt 2018: 530). Digital photography and the taking of fingerprints by digital scanning are somehow exempt from being defined as technical processing, despite being processes that require specific technology and must be carried out in a “highly controlled and cooperative manner” (Labati et al 2015: 1). As anyone who has applied for a visa knows, facial photography for ID purposes has a set of requirements—full face, no glasses, use of plain white background—that produce an image which most people say does not look like them. The photograph renders the person as the state wishes to see them. In short, the European Union creates the appearance of biometric data protection, while allowing for its wholesale use under the pretense that it is “unprocessed.” The most recent forms of facial recognition, like Clearview AI, use immense pools of “scraped” social media photographs and other generic snapshots to identify people, giving security services an additional means to bypass the question of “processing” (Hill 2020).

Eurodac’s policing of white space is producing a blank white non-space, the space of disappearance. In Denmark, to take a key zone of Nordic whiteness, the strategy is to disappear the refugee from society. Asylum seekers cannot work, cannot claim benefits and cannot accept cash donations. They are currently detained in camps run by the prison service, even though under the 1951 Geneva Convention, claiming asylum is a human right. In 2019, I was able to visit Sjaelsmark thanks to the invitation of an Eritrean refugee I’ll call “Lily”—the camps are not open to outside visitors. Like many others, Lily was denied leave to remain in Denmark because her fingerprints were first taken in Greece. The Sjaelsmark departure or expulsion centre for refugees denied asylum (technically “non-deportable rejected asylum seekers,” according to EU law) where Lily was detained is accessed through a formidable gate, which is locked every night at 10 pm, even to residents. Individuals and families live in former military barracks (in April 2020 families will be moved out of the centre) (The Local 2019). The residents call it a “camp,” and it is newly surrounded by 10-foot-high security fences. Although residents can leave whenever they want, the effect is one of imprisonment. The camp is 25 km from Copenhagen, a journey that takes two hours by public transport.

If the Danish settler colony once wanted to extract labor from its colonial subjects in the Caribbean, Africa, and Asia, all it wants now from their descendants is that they go away. To that end, Sjaelsmark residents cannot cook, have furniture (other than a bed, one table, and hard chair), or decorate their

rooms. No carpets or rugs are allowed. There is no television, radio or Internet service. Residents live in cold, spare, whitewashed rooms with very high ceilings: erased white space. The Social Democratic victory in the 2019 elections has led to a reduction in these cruelties. Families will no longer be housed at Sjaelsmark as of April 2020 and detainees will be allowed to cook and eat in their rooms. There is still no broader solution for the political limbo where these asylum seekers mostly find themselves. They have lost what Hannah Arendt called the “right to have rights,” and are still being disappeared. It is clear that many others, formally citizens, are entering the space of disappearance, permanently or temporarily, and losing the right to have rights, whether a person subject to London police using facial recognition on anyone in certain areas; a person present in areas where the Coronavirus has infected others; people kettled or otherwise restrained by police when protesting and so on.

In this review of the production of white space across the hitherto-existing span of Atlantic world coloniality, several trends can be detected. There is a consistent production of artificial vision to create and sustain erased space that can be colonized. The introduction of automated machine vision and machine learning has absorbed the long history of coloniality as its “intelligence” and continues to reproduce it. However, in this automated form, visibility is distributed, rather than centralized. Artificial vision was constructed around a single point from which to see, whether that of the overseer or the colonial state. It then developed platforms from which this vision might be deployed, from the deck of the ship to the platform of the train. Artificial vision relied on infrastructures tied to specific places, whether the Atlantic sea routes driven by oceanic currents or the material form of the railway track. With the UAV and distributed machine-learning, the state can now deploy its artificial vision wherever it wants, whenever it wants. CCTV is the domestic application of this apparatus that has already become ubiquitous in places like China and the UK. The function of this machine vision has circulated over time. To adapt the aphorisms of Lorenzo Veracini, the colonial state first told the Indigenous “you, go away.” It alternated that command with “you, work for me,” a directive also used for forced migrants (Veracini 2011: 1–12). The latter was extended to once-colonized subjects invited back to the metropole in the labor shortages following the Second World War. With the creation of globally distributed labor forces, racial surveillance capitalism now says “go away” to all those surplus to its requirements. Accordingly, it now seeks to disappear them, rather than keep them under close watch. Whether they live or die is a matter of indifference to the state and grounds for visual activism for the rest of us.

**Funding** No funding was received for this work.

## Compliance with ethical standards

**Conflict of interest** I confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

## References

- Adamson J (2017) Carbon and the cloud: hard facts about data storage. In: Stanford magazine. <https://medium.com/stanford-magazine/carbon-and-the-cloud-d6f481b79dfe>
- [US] Army, Marine Corps, Navy, Air Force (2009) Kill box multi-service tactics, techniques, and procedures for kill box employment FM 3-09.34. Department of Defense
- Benjamin R (2019) Discriminatory design, liberating imagination. In: Captivating technology: race, carceral technoscience, and liberatory imagination in everyday life. Duke University Press, Durham
- Berger J (1972) Ways of seeing. Pelican, Harmondsworth
- Berger S (2017) The art of philosophy: visual thinking in Europe from the late renaissance to the early enlightenment. Princeton University Press, Princeton
- Bredenkamp H (2007) Thomas hobbes's visual strategies. In: Patricia S (ed) The Cambridge companion to hobbes's *Leviathan*. Cambridge University Press, Cambridge
- Browne S (2015) Dark matters: on the surveillance of blackness. Duke University Press, Durham
- Bugge M (2019) Obedience and dehumanization: placing the Dublin regulation within a historical context. *J Hum Rights Soc Work* 4:91–100
- Casid J (2015) Scenes of projection: recasting the enlightenment subject. University of Minnesota Press, Minneapolis
- Chamayou G (2015) Drone theory. Penguin, London
- Condon S (2018) In 2018, AWS Delivered Most of Amazon's Operating Income. ZDNet. <https://www.zdnet.com/article/in-2018-aws-delivered-most-of-amazons-operating-income/>
- Deer S (2015) The beginning and end of rape: confronting sexual violence in Native America. University of Minnesota Press, Minneapolis
- Defense Acquisitions (2013) Assessments of selected weapon programs. GAO-13-294SP
- Derrida J (1984) Différance. In: Margins of philosophy, trans. Alan Bass. University of Chicago Press, Chicago
- Estes N (2019) Our history is the future: standing rock versus the Dakota access pipeline and the long tradition of indigenous resistance. Verso, New York
- Feldman K (2011) Empire's verticality: the Af/Pak frontier, visual culture and racialization from above. *Comp Am Stud* 9(4):325–341
- Freedom of Movements Research Collective (2018) Stop killing us slowly. A research report on the motivation enhancement measures and the criminalisation of rejected asylum seekers in Denmark. Denmark
- Galton F (1901) Biometry. *Biometrika* 1(1)
- Breckenridge K (2014) Biometric State : The Global Politics of Identification and Surveillance in South Africa, 1850 to the Present. Cambridge University Press, Cambridge
- Gilroy P (1993) The Black Atlantic: modernity and double consciousness. Harvard University Press, Cambridge
- Graden DT (2014) Disease, resistance, and lies: the demise of the transatlantic slave Trade to Brazil and Cuba. Louisiana University Press, Baton Rouge
- Hajjar L (2017) Lawfare and armed conflicts: a comparative analysis of Israeli and U.S. targeted killing policies and legal challenges against them. In: Lisa P, Caren K (eds) Life in the age of drone warfare. Duke University Press, Durham
- Hall S (2017) The fateful triangle: race, ethnicity, nation. Harvard University Press, Cambridge
- Hao K (2019) This is how AI bias really happens—and why it's so hard to fix. *MIT Technol Rev*. <https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/>
- Hardt M, Negri A (2017) Assembly. Oxford University Press, New York
- Hayes B (2017) Migration and data protection: doing no harm in an age of mass displacement, mass surveillance and big data. *Int Rev Red Cross* 179
- Hill K (2020) The secretive company that might end privacy as we know it. <https://www.nytimes.com/2020/01/18/technology/clear-view-privacy-facial-recognition.html?searchResultPosition=1>
- Hobbes T (1651) Leviathan or the matter, forme, & power of a common-wealth ecclesiasticall and civil. London
- Hu T-H (2015) A pre-history of the cloud. MIT Press, Cambridge
- Jafa A (2019) A series of utterly improbable yet extraordinary renditions. Galerie Rudolfinum, Prague
- Joh EE (2016) The new surveillance discretion: automated suspicion, big data, and policing. *Harvard Law Policy Rev* 10(1):15
- Johnson W (2013) River of dark dreams: slavery and empire in the Cotton Kingdom. Belknap Press of Harvard University Press, Cambridge
- Kindt EJ (2018) Having yes, using no? About the new legal regime for biometric data. *Comput Law Secur Rev* 34:532–538
- King G (2017) Where the buffalo no longer roamed. *Smithsonian Magazine*. <https://www.smithsonianmag.com/history/where-the-buffalo-no-longer-roamed-3067904/>
- Labati RB, Donida R et al (2015) Touchless fingerprint biometrics. CRC Press LLC, Boca Raton
- Lee S (2019) Commodified (In)security: cultural mediations of violence in Israel, Palestine, and beyond. chapter three, PhD dissertation, NYU
- The Local (2019) <https://www.thelocal.dk/20191121/denmark-to-allow-families-to-move-out-of-controversial-refugee-facility>
- Long Soldier L (2017) Three. In: Whereas. Graywolf Press, Minneapolis
- Marx GT (2016) Windows into the soul: surveillance and society in an age of high technology. University of Chicago Press, Chicago
- McKittrick K (2013) Plantation futures. *Small Axe* 17(3):1–15
- Mirzoeff N (2016) How to see the world: an introduction to images from self-portraits to selfies, maps to movies and more. Basic Books, New York
- Mirzoeff N (2011) The right to look: a counterhistory of visibility. Duke University Press, Durham
- Moten F, Harney S (2013) The undercommons: fugitive planning and black study. Minor Compositions, New York
- Muller BJ (2010) Security, Risk and the Biometric State: Governing Borders and Bodies. Routledge, New York
- Nemser D (2009) Infrastructures of race: concentration and biopolitics in Colonial Mexico. University of Texas Press, Austin
- Quijano A (2000) Coloniality of power, eurocentrism and Latin America. *Nepantla* 1(3):215–232
- Parks L (2017) Vertical mediation. In: Parks L, Kaplan C (eds) Life in the age of drone warfare. Duke University Press, Durham
- Philip MN (2008) Zong! Wesleyan University Press, Middletown
- Public Law (2009) No: 111–118. <https://www.congress.gov/bill/111th-congress/house-bill/3326/text>
- Pugliese J (2007) Biometrics, infrastructural whiteness, and the racialized degree zero of nonrepresentation. *Boundary* 2 34(2):105–133
- Regulation of the European Parliament on the Protection of Natural Persons with Regard to the Processing of Personal Data and



- on the Free Movement of such Data, and Repealing Directive (2016) 95/46/EC, 2016/679/EU, 27 April 27
- Rhue L (2018) Racial influence on automated perceptions of emotions. <https://ssrn.com/abstract=3281765>. <https://doi.org/10.2139/ssrn.3281765>
- Robinson C (2000) *Black marxism: the making of the black radical tradition* [1983]. University of North Carolina Press, Berkeley and Los Angeles
- Rose J (1986) *Sexuality in the field of vision*. Verso, London
- Ruiz C (2020) The art newspaper. [https://www.theartnewspaper.com/news/president-trump-blows-up-native-american-grave-site-for-his-border-wall?fbclid=IwAR0JWfLcJEeyx6Oa-7PIFs-jOYTULQNyxCy6R6H7D11K7jQF71\\_mSt8Q0KU](https://www.theartnewspaper.com/news/president-trump-blows-up-native-american-grave-site-for-his-border-wall?fbclid=IwAR0JWfLcJEeyx6Oa-7PIFs-jOYTULQNyxCy6R6H7D11K7jQF71_mSt8Q0KU)
- Sharpe C (2017) *In the wake: on the blackness of being*. Duke University Press, Durham
- Schivelbusch W (1987) *The railway journey: the industrialization and perception of time and space in the nineteenth century*. University of California Press, Berkeley and Los Angeles
- Simpson A (2016) The state is a man: Theresa Spence, Loretta Saunders and the Gender of Settler Sovereignty. *Theory Event* 19(4)
- Skinner Q (2018) *Hobbes and the humanist frontispiece. From humanism to hobbes: studies in rhetoric and politics*. Cambridge University Press, Cambridge
- Spice A (2019) Heal the people, heal the land: an interview with Freda Hudson. In: Estes N, Dhillon J (eds) *Standing with standing rock: voices from the #NODAPL Movement*. University of Minnesota Press, Minneapolis
- TallBear K (2019) Badass indigenous women caretaker relations: #Standingrock, #IdleNoMore, #BlackLivesMatter. In: Estes N, Dhillon J (eds) *Standing with standing rock: voices from the #NODAPL Movement*. University of Minnesota Press, Minneapolis
- Taylor ARE (2017) The technoaesthetics of data centre ‘white space.’ *Imaginations* 8(2):42–55
- Tralau J (2007) Leviathan, the beast of myth: Medusa, Dionysos, and the Riddle of Hobbes’s Sovereign Monster. In: Springborg, Cambridge Companion, pp 65–69
- Veracini L (2011) Introducing. *Settler Colonial Stud* 1(1):1–12
- Walvin J (2011) *The Zong: a massacre, the law and the end of slavery*. Yale University Press, New Haven
- West SM, Whittaker M, Crawford K (2019) Discriminating systems: gender, race and power in AI. *AI Now Institute*, 5. <https://ainowinstitute.org/discriminatingystems.html>
- Wynter S (2015) The ceremony found: towards the autopoietic turn/overturn, its autonomy of Human Agency and Extraterritoriality of (Self-)Cognition. In: Broeck S, Ambrose JR (eds) *Black knowledges/black struggles: essays in critical epistemology*. Liverpool University Press, Liverpool
- Zuboff S (2019) *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. Profile Books, London
- Zuboff S (2020) *The known unknown*. New York Times

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# AI urbanism: a design framework for governance, program, and platform cognition

Benjamin Bratton<sup>1</sup>

Received: 18 November 2020 / Accepted: 23 December 2020 / Published online: 7 February 2021  
© The Author(s), under exclusive licence to Springer-Verlag London Ltd. part of Springer Nature 2021

## Abstract

Historically, the dynamic between philosophy of artificial intelligence and its practical application has been essential for the development of both, and thus the encounter between theory of AI and architectural/urban theory should be a site of considerable productivity. However, in many ways, it is not. This is due to two primary factors, one arising from each side of this encounter. First, legacies of overly-anthropomorphic models of AI permeate design discourses, where issues of how well AI can be constrained to social issues of philosophy of mind exclude more foundational and cross-cultural questions about artificial intelligence as material process in the physical world that could inform new philosophical insights. Concurrently, the mobilization of investment in “Smart Cities” discourses dominates the space of application of AI at urban scale in ways that prematurely fixes solutions in the skeuomorphic image of architectural and urban conventions. To break this impasse, we held a series of think-tanks, workshops, and design charettes that brought together leading figures in philosophy of artificial intelligence, urban design, and commercial AI platforms. The goal was to re-think from first principles how alternative philosophical models of AI as a distributed, discontinuous, landscape-scale technology in a direct encounter with the applications in contexts of ecological sensing, automation’s impact on urban form, and issues of algorithmic governance. The design brief for this research establishes a generative framework for conceiving how embedded machine sensing and intelligence is already changing urban form (but unrecognized) and could change it in the future (based on more appropriate design projections). The framework focuses on issues of urban zoning and architectural programming, data modeling and governance, platform cognition and design, and how these inform shifts in dynamics of public and private institutions.

**Keywords** AI · Urbanism · Design · Governance · Program · Platform cognition

## 1 Introduction

Instead of conceiving of artificial intelligence as a kind of disembodied mind, we would be better served by considering it as something more like a quality of landscapes: it is physically embedded, sensory, decentralized, distributed, and heterogeneous, such that its “intelligence” is, like the clustering of neurons, an emergent effect of complex and non-linear interactions. The social philosophy of AI is nothing if not ready with new metaphors, and I have argued for a “synthetic garden” metaphor with respect to a general conception of AI in the field, but here I wish to use it to make more concrete propositions for how AI will, or should,

transform some foundational questions of urban design. AI should be conceived on landscape scale, and urban landscapes can be reconciled in relation to AI, but how so?

The most prevalent and popular models are those associated with the catch-all “Smart City,” a stale shorthand which has come to include the sensible and the silly. As a model of AI urbanism, however, it is crippled both by a lack of imagination as to what is possible beyond the simple automation of prosaic functions, and by a deep misrecognition of the already existing intelligence of urban systems. Its conceit, to introduce intelligence to systems that have none so that they can do what they have always done but slighter more efficiently, is wrongheaded on both priorities. Corollary to these are various “AI City” projects that promise a radical new way of architecture but which are, clearly to anyone who looks carefully, only the most conventional design banalities slathered with an “AI” marketing trope. By contrast, the approach we took to AI Urbanism at The

---

✉ Benjamin Bratton  
bbratton@ucsd.edu

<sup>1</sup> University of California, La Jolla, San Diego, CA, USA

New Normal think-tank at Strelka Institute in Moscow prioritizes the social and technical intelligence of existing urban systems and asks how, and indeed *if*, artificial intelligence can add to them, and if so, on what terms? The interdisciplinary researchers were drawn from 17 countries and for this research module were joined by Blaise Aguera Y Arcas, of Google Cerebra AI research group, Kenric McDowell of Google Artist and Machine Intelligence, and Ben Cerveny of Foundation for Public Code and more. They were joined by Philosophers, Reza Negarestani, Peter Wolfendale, Patricia Reed, and others. The work was supported only the Institute which received no outside funding for this work.

What follows is not the design that emerged from these sessions, but rather an explication of *the design brief* that we used to develop that work. This will likely have greater value for work by (1) architects and urban designers as they consider the long-term impact of artificial intelligence on city's form, function and governance, and (2) software applications and systems designers as they model and build AI-inclusive programs for socially and culturally complex urban environments. As both seek to define the role of artificial intelligences in the City layer of The Stack, the brief provides one possible framework that emphasizes issues of governance, architectural program, data structures, and platform cognition.

The research was divided into four thematic areas. Program to Program focuses on the conflicts and/or complementarity between software and architectural programming and looks at changes in the planned organization of urban space as an adaptation to pervasive automation. AI Governance looks at how artificial intelligence becomes both a means with which urban systems are governed, and equally importantly, a socio-technical process which itself demands to be governed. Data: Direct and Derivative asks design researchers to consider the dynamics of how data can, in effect, produce the territory that it models, and how problematics of data transparency and opacity can determine the social relations between different, even competitive, platforms. Finally, Platform Cognition, considers how artificial intelligence at platform scale is both produced and constrained in accordance with how it is able to sense the world, produce data about its surroundings, and construct active models of the world according to this artificial temporality. Together these represent four areas of direct and speculative research for (1) conceiving a general model of artificial intelligence as an urban-landscape-scale phenomenon rather than a disembodied “mind in a box,” and (2) to do so by the projective modeling of different relevant urban scenarios for the design of form, program, system and social governance.

## 1.1 AI urbanism introduction

AI is a technology that has developed in relation to thought experiments about AI (Turing's Test, Searle's Chinese Room, etc.), which have informed the actual technologies, which have informed further thought experiments. This double-helix between the real and the virtual forms of AI should continue apace. Encounters should be staged between advanced work in the philosophy of artificial intelligence and the cutting-edge applications of the technology at platform-scale to discern, where the translation between formal and practical models of AI breaks down. It's there that further thought experiments may be most inspired, and this encounter was the goal of this research brief.

AI is an urban-scale phenomenon that is dependent on myriad sensor arrays to know the world in which it is co-situated. Instead of considering AI in a petri dish, as some disembodied artificial mind, synthetic sensing and intelligence should be understood as a distributed function of the material world: a polyphonic orchestra of automation amalgamated from this uneven topography, capable of unexpected creativity and cruelty. With comprehensive automation, modern urban programs that have been drawn by the cycles of residence, work, and entertainment of earlier eras are open to re-imagination.

Our technical interest lies in the anonymization and depersonalization of data, federated learning, decentralization, and granular distribution of AI, instead of an infrastructure for surveillance of individuated humans, cloud centralization, and coordination of monopoly models. Radical drops in price and energy usage at the tail end of Moore's Law allows for inexpensive machine-learning-capable chips to be embedded and distributed widely, for data to be federated more than extracted and models to be both effective and more anonymous. Taken as a new potential property of everyday objects, site-specific AI is closely tied to how it senses what does and does not become part of different machine-readable worlds.

The implications of understanding AI at urban scale also point in the direction of their geopolitical implications. Artificial models of the world both ingest and output information, and as the output of one is the input of another, each may make the world more or less fragile, and each may be either weaponized or contribute to economies that make weaponization too expensive. They may make governance of design impossible or, instead, make it finally possible. Regardless, AI's impact on cities will be profound, but perhaps not as profound as the impact cities will have on AI. What follows are four design research briefs, each with a series of questions meant to guide an exploration of the potentials and implications of AI at urban scale.

## 1.2 Program to program

Since at least the 1980s, a fundamental shift in the means of spatial organization has taken place. The functions that society used to ask of architecture—such as to coordinate the flows of large groups of people—it now asks software to accomplish. This may leave architecture without a clear assignment. Its assignment may become more “autonomous” from function, synthesized into the abstract geometry and parametric convolution of increasingly thin but complicated skins and surfaces. Or, architecture’s autonomy may be dissolved into space machines of maximum reconfigurability, such as the so-called supermodern aesthetic of warehouses, data centers, and industrial architecture. At both extremes of the discipline’s reaction, architecture shifts its range of function, as its previous ones are absorbed by software platforms. The change in architecture caused by software is not simply because there is now software in architecture, and so making it computational, but because a previous architectural program is now coordinated by those software-based programs.

## 1.3 Zoning and program

How has/will the intensification of automation at urban and regional scale shift the necessary zoning logics and standards? Conventional categories of residential, industrial, light industrial, retail, etc. may no longer be relevant, or their adjacencies may need resorting. What are key examples of novel zoning and programming developments driven primarily by the needs of software infrastructure or platforms? What patterns and aspects of these will drive further shifts? How and why? Key patterns in this new urban zoning may already exist and yet not have proper names. What should they be?

## 1.4 AI constructivism

The Constructivist era in Russian architecture was known for its programmatic innovation. It attempted to anticipate the needs of a new socialist society and to provide for them in advance—and in doing so, to accelerate the formation of that society. These included shared kitchens, courtyards, and social areas, massive mini-city sized apartment blocks, living quarters for different family structures, extensive social facilities near (or even in) factories and other places of work, and so on. This was not only a political and social meta-program for architecture; it was also a propositional response to the spatial requirements of industrial technology at national scale. Hypothesizing that AI/automation may have an equally profound effect and may open up new ways of living and being together “inside,” what might a revitalization of the Constructivist ethos for this era entail? Does it mean radical virtualization toward space-as-service? Must

that be based on monopoly economics? And if not, what are alternative urban models?

## 1.5 Isolate vs. gradient texture

The relation between human-centric and machine-centric spaces must be considered in terms of gradients of exclusion and inclusion, from a strict prophylactic division to a cyborgian intermingling of people and machines. The question of programmatic transformation in this is not simply a resorting of existing blocks. A model, diagram, and categorical framework are needed to describe the structure, distribution, dynamics, and programmability of this “exclusion gradient.” What are the logical and practical structures of that gradient?

## 1.6 AI governance

The Modern bureaucratic nation-state was and still is an information sensing, storing, and processing institution. Modern citizenship is a legal status, as well as a set of persistent records linked to a credentialed individual: economic exchanges, cartography, real estate records, taxes, and perhaps most importantly, debts. Its ability to govern and to cohere a consensus of legitimacy depends on the proficiency and regularity of the management of debt. Debt, and the prospective return on debt, in turn depends on the consensual tabulation of value in some real or intangible currency. That currency serves as a repository for standardized units of value to be circulated within an economy, and more importantly, in a society whose ontology conceives what is and is not valuable as such. In this regard, accounting reigns as the supreme writing.

The Modern state held a local monopoly on the credentialization of currency, on the tokenization of value, and thus on the (formal) social ontology of value. In some ways this has proved more important than the corollary monopoly on violence (they are related of course). Some would argue that an advanced market is itself a form of artificial intelligence, one built of human–machine amalgamations. Even if so, new media allow for new categories through which governance may address its polity and oikos, including new geographies. But they may also simply give algorithmic fortification to vestigial institutional forms. In principle, we hope that emergent ontologies give way to preferred modes of social self-organization, but how so?

## 1.7 Who trains the trainers?

Any urban AI system that is built is only as good as its training data. At the start, any neural network is raw links and nerves, and any training set is a stream of noisy protuberances which bend that network into shape. Training the layers to recognize regular signals is as much like training a

stone to become a spear tip as it is like training a dog to sit. But in practice, training sets are built from available opportunities and resources. These are sometimes robust and broadly representative of what the set means to model, and sometimes they are bizarre artifacts of legacy bias, repeating and reinforcing the same all-too-human prejudices over and over again. When can increasing the scope and depth of training data help to correct the model, and when does it dilute the pattern?

## 1.8 Institutional form

What are the constitutive anatomies of AI governance? Will its emergence as an information infrastructure show how the modern political order was the effect of a modern bureaucratic informational apparatus? AI can and does contribute to the governance of both discrete and diffuse polities with different scales, locations, and legal forms. The intensification of its formal and de facto governing effects traces these differences and works against them. Whereas states identify and coordinate the lives of citizens, macroeconomic futures, and military gambits, cities mediate the encounters of inhabitants, the platform infrastructures of everyday life, and the logistics of circulation. AI can organize each of these, of course, but how may it also shift the effective roles of state vs. city?

## 1.9 AI as geopolitical gambit

In 2018, Xi Jinping released China's plan as part of a larger China 2020 initiative that benchmarked the country's progress in AI to match the USA's by 2020 and surpass it by 2030. At around the same time, Putin remarked to a group of students that whoever controls AI will control the world, and suggested that the technology should be a shared, open resource. Macron authored an EU AI plan that focused on social responsibility, citizen-first data models, and supporting regional research. The USA followed suit with a more vague white paper, reiterating the importance of AI to its industries and ambitions, etc. Each of these is a genre of hemisphere-scale speculative strategy. Can they all come true at once, or do they map a more zero-sum game?

## 1.10 Algorithmic governance

Politics and technology are interwoven as a means to remake the world by design. Not only does technology express a political arrangement, but any polity emerges only within a technical milieu. In an age of planetary-scale computation, the dynamics of algorithmic governance shift the perspectives on political geography, sovereignty, citizenship, regimes of rights, and how leverage is embedded in computational technologies. Some governing institutions

seem to liquify only to congeal in new ways, while others are grotesquely amplified. Models that categorically divide centralization vs. decentralization or public vs. private are of limited help in conceiving the systems we need. What other heuristics are most plausible to develop a formal or practical typology of algorithmic governance?

## 1.11 Platform sovereignty

AI can be used to refortify and re-amplify earlier forms of constructed formal citizenship, and it can, just as it does for the transnational flows of data and objects, facilitate very different cuts of territory and sovereignty. Platforms generate different forms of sovereignty based on generic access and centripetal/centrifugal value distribution. These may include formal state sovereignties (most obviously in China) and informal platform sovereignties (most obviously in the USA). Can we anticipate the evolution of divisional Galapagos effects or a global convergence of these models?

## 1.12 Data: direct and derivative

AI works both with and on data, but the boundary between information sensing and information processing is as fuzzy for AI as it is for humans. Data is a sort of sovereign substance that multipolar hemispherical stacks defend the right to generate and use as the basis of governing models and simulations. Data is generated by sensing and modeling systems, and so the referent of the data—the person, place, or action—is likewise both constructed by the act of modeling and revealed by its correlations. All of this unfolds unevenly and opportunistically, and so a dissensual politics of data is inevitable. The agency of something within these constitutive systems can be constructed directly by its data shadow, or indirectly by the integration of multiple traces. Data is not only a means to represent social and economic realities, but also to construct and compose them. In what ways are our models of modeling wrong?

## 1.13 Transparency, opacity, control

Opacity and transparency have become an ethics in and of themselves, irrespective of how their principles are applied. The transparent society suggests that formal and informal social contracts, especially between rulers and ruled, will be more consensual and stable when more decisions and interactions are visible to others. The ethics of opacity suggest, conversely, that privacy allows not only the fortification of individuation but also the development of robust “trustless” platforms that enable the positive effects of decentralization. For some, both extremes can co-exist, albeit sometimes somewhat schizophrenically (i.e., Assange). Popular representations of these conditions are inadequate to understand

their complexities. What are the attractors and dangers of societies in which “everybody knows,” vs. those in which “nobody knows”? How would we know the difference? What urban models represent them?

### 1.14 Resolution of territories

A cursory *dérive* over the depopulated planet of Google Earth demonstrates that not all territories are represented in the same way. There are high-resolution territories modeled with highly granular details, both visual and statistical, and there are low-resolution territories, for which available data is much blockier. The implications of this are not so straightforward. High-resolution territories may enjoy greater services and support in some ways, but also more intensive supervision. Other territories may be low-resolution in one respect, but high-resolution in another (such as personal health data vs. crime data). The spectrum of resolution represents an underlying difference between these territories, and is even generative of those differences. How may urban systems evolve to promote or suppress various modes of resolution? How would these intersect with existing programmatic conventions?

### 1.15 Identity and derivative metrics

The indirect construction of a data referent through composite formulation, metadata, and corollary traces (such as a phone number) is, in some ways, a more important methodology than simple, direct observation in the panoptic sense. In this, it becomes clear how the functional identity of a referent is constructed through the process of modeling in ways that need not correspond to how that referent self-identifies. There is, then, especially for people, an essential disconnect between inherent and expressive identity. The profile is formulated in relation to the correlation of indirect fragments of information that are linked into an individuated identity. Expressive identity can be experienced through the mediated traces of a fragmented inherent identity. Is this toxic or inevitable?

### 1.16 Platform urbanism

Public and private platforms operate according to their ability to rationalize and assume value from the data that they mediate. They may not directly tax users or charge for transactions, but they assume and enforce economic valuation by how they generate a “surplus” of value from both the direct and derivative data that they model. Some of that surplus is realized by end-users who benefit from platform systems and effects, but most is absorbed by the platform itself. For some forms of platform capitalism, this concentration is severe, but for others less so. This dynamic oscillates between

public and private versions, and positions are rearranging quickly. Platform systems might appear to resolve toward one sort of outcome, but may also form the structural basis of a very different kind of political economy exchange (i.e., “People’s Republic of Walmart”). What urban forms, however, strange, may enable those trajectories and economies?

### 1.17 Platform cognition

Any single sensed event or action is, like the firing of one neuron, just a blip. It is only by the accumulation and integration of massive quantities of instances that something like patterns and then intelligence can emerge. Because no two platforms are structured to sense the world in the same way, no two platforms will achieve synthetic cognition in the same way. Some are trained on object flows for which end-users are mostly just input and output variables, but for others the artificial formulation of end-user self-expression is both means and ends. Yet, platform architectures constructed on behalf of particular immediate strategies will also evolve to allow for unforeseen possibilities and demands. They will adapt to the capacities of competing platforms, and this co-evolution represents the external forces that frame the internal dynamics of any system. That is, among the things that platforms sense and calculate are not only other platforms, but perhaps more importantly, they sense and model themselves.

### 1.18 Simulation and prospection

The absorption, organization, and modeling of data serve to populate complex simulations of a governed world. Such simulations are not merely representations of underlying processes, but serve as read/write media through which steering decisions about how to intervene in those processes are manifested. As platforms comprehend the world in relation to the categorical schema that orient their perception and logics, their ability to model and predict events in the world around them is at least partially tautological. Given that cognition could, in principle, be based on quite alien categories, what alternative taxonomic possibilities and urban platforms might they engender?

### 1.19 Addressability and dimensionality

Any platform sensory apparatus senses and responds to specific input events and not others, and is thus bound by a strategic or tactical ontology. That ontology in turn depends on means to identify, nominate, and address events as discrete individuated entities that can be re-identified, re-nominated, and re-addressed as needed. In other words, platform cognition depends on a dimensional array of addressable instances, any of which may have quite different relations to

real world people, places, and things. The intrinsic abstractions of any addressing ontology are inseparable from the particular intelligence of that platform. What are some viable alternative addressing systems, and how do they arrange the urban landscape in their image?

### 1.20 Temporality and threshold

A data set always has some direct or indirect temporal scope. It can be an instantaneous snapshot of many things happening in the blink of an eye, or one thing happening over centuries, or various combinations of few and many, slow and fast. In this sense, platform cognition is not only delimited by the anatomic and strategic distribution of its sensing media, but also by the temporal thresholds of its durational models. What causal patterns might be rendered by forms of machine intelligence that work on data sampled from inhumanly vast or minuscule timescales that would be otherwise incomprehensible?

### 1.21 Camouflage and display

In any complex ecology, biological or not, multiple participant species, dependent on their unique sensory media, interrelate with one another according to competitive and symbiotic dynamics. In this, not only do more powerful ways to sense the world evolve, but also ways to prevent others from seeing and deducing. This is the arms race of perception and camouflage, one of the primary forces that give form to perpetual media and the surfaces they try to see and interpret. The strategy of camouflage is made by both predator and prey, but so are elaborate forms of display that solicit and encourage the contact and pollination between species and within a species. This is as true of bees and flowers as it is of WeChat-enabled cities and ambient distributions of QR codes. How does platform-scale AI mimic or deviate from natural evolution in relation to these optical dynamics?

## 2 Conclusion

“Artificial intelligence” is not one thing, but many. It is many different technologies, and it is many different metaphors. As any culture would define “the artificial” differently, and

“intelligence” differently, so too, the term “artificial intelligence” will host multiple connotations. Our approach seeks to glue a radically materialist view of AI seen as emergent property of matter arranged in such a way as to produce intelligent effects with a macroeconomic and geopolitical view that sees that arrangement as a kind of world-making and world-governance. The valence of governance here is not by default negative. There is a sincere presumption that AI has an important role to play in how societies sense, model, simulate and act back upon themselves. The pandemic, for example, has done much to push the importance of that to the foreground. The implications of this approach may make other approaches less relevant, for example those that see AI exclusively in terms of philosophy of mind or exclusively as an infrastructure captured by capital. At the same time, however, the landscape model of AI (“synthetic garden”) has implications for those approaches and how they define their research programs, many of which profoundly overlap with our own. Finally, the conclusion of our investigation is that to understand AI as a property that can be designed into objects of different scales, from molecular to urban scale, focuses AI critique and AI design toward fundamental social questions of what kind of planetarity can and must be composed.

### Compliance with ethical standards

**Ethical approval** This research was supported by Strelka Institute of Architecture, Media and Design, a non-profit educational institution based in Moscow, Russia.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Correction to: Excavating AI: the politics of images in machine learning training sets

Kate Crawford<sup>1</sup> · Trevor Paglen<sup>2</sup>

Accepted: 4 October 2021 / Published online: 23 November 2021  
© Springer-Verlag London Ltd., part of Springer Nature 2021

## Correction to: AI & SOCIETY

<https://doi.org/10.1007/s00146-021-01162-8>

The Editor-in-Chief has removed Fig. 1 due to copyright and consent concerns, along with the related references, and Fig. 2a has been updated. A sentence in Sect. 3 “Anatomy of a training set” explaining the authors’ understanding of the intended purpose of the JAFFE dataset (Lyons and Akamatsu 1998) has also been removed.

Additionally, the Authors would like to provide the following clarifications:

- Figure 2 gives examples of images from the ImageNet database, which was first introduced in the cited article (Deng et al. 2009)
- On page 6, the following sentence incorrectly references a figure: “A photo shopped picture shows a smiling Barack Obama wearing a Nazi uniform, his arm raised and holding a Nazi flag. It is labelled “Bolshevik (Fig. 2).”” The citation to Fig. 2 should have been placed in the following paragraph when referencing the labelled image of Sigourney Weaver
- The images in Fig. 3 were reproduced from the “Aligned&Cropped Faces” dataset of UTKFace (UTKFace-Aicip 2019)
- The full caption for Fig. 4 is: “Fig. 4 image from IBM’s Diversity in Faces paper, image credit M. Merler et al. (Merler et al. 2019).”

The original article can be found online at <https://doi.org/10.1007/s00146-021-01162-8>.

✉ Kate Crawford  
Kate.crawford@usc.edu

Trevor Paglen  
trevor@paglen.com

<sup>1</sup> University of Southern California, Annenberg School, Microsoft Research New York, New York, USA

<sup>2</sup> The University of Georgia, Athens, Greece

## References

- Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition, pp 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- Lyons M, Akamatsu S (1998) Coding facial expressions with gabor wavelets. In: Proceedings, third IEEE international conference on automatic face and gesture recognition, 14–16 April 1998, Nara, Japan. IEEE Computer Society, pp 200–205. <https://doi.org/10.1109/AFGR.1998.670949>
- Merler M, Ratha N, Feris RS, Smith JR (2019) Diversity in faces. ArXiv 4:1901–10436. [arXiv:1901.10436](https://arxiv.org/abs/1901.10436)
- UTKFace-Aicip. <http://aicip.eecs.utk.edu/wiki/UTKFace>. Accessed 28 Aug 2019

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.