The Social Lives of Generative Adversarial Networks

Michael Castelle
University of Warwick
Centre for Interdisciplinary Methodologies
January 29th, 2020
M.Castelle.1@warwick.ac.uk

ABSTRACT

Generative adversarial networks (GANs) are a genre of deep learning model of significant practical and theoretical interest for their facility in producing photorealistic ‘fake’ images which are plausibly similar, but not identical, to a corpus of training data. But from the perspective of a sociologist, the distinctive architecture of GANs is highly suggestive. First, a convolutional neural network for classification, on its own, is (at present) popularly considered to be an ‘AI’; and a generative neural network is a kind of inversion of such a classification network (i.e. a layered transformation from a vector of numbers to an image, as opposed to a transformation from an image to a vector of numbers). If, then, in the training of GANs, these two ‘AIs’ interact with each other in a dyadic fashion, shouldn’t we consider that form of learning... *social*? This observation can lead to some surprising associations as we compare and contrast GANs with the theories of the sociologist Pierre Bourdieu, whose concept of the so-called habitus is one which is simultaneously cognitive and social: a productive perception in which classification practices and practical action cannot be fully disentangled. Significantly, Bourdieu used this habitus concept to help explain the reproduction of social stratification in both education and the arts. In the case of learning, Bourdieu showed how educational institutions promote inequality in the name of fairness and meritocracy through the valorization of elite forms of ‘symbolic capital’; and in the arts, he often focused on the disruptive transitions in 19th-century French painting from realism to impressionism. These latter avant-garde movements were often characterized by a stylistic detachment from economic capital, as “art for art’s sake”, and this cultural rejection of objective-maximization—a kind of denial of an aesthetic ‘loss function’—can in turn help highlight a profound paradox at the core of contemporary machine learning research.

KEYWORDS

generative adversarial networks, sociological theory, habitus, bias, game theory

1 INTRODUCTION: FROM PHOTOGRAPHY TO NEUROGRAPHY

> “Each day art further diminishes its self-respect by bowing down before external reality; each day the painter becomes more and more given to painting not what he dreams but what he sees. Nevertheless it is a happiness to dream, and it used to be a glory to express what one dreamt. But I ask you: does the painter still know this happiness?” — Baudelaire, 1859 [6]

At the present moment it is hard to be sure, but there is a possibility that we stand today towards the technological subject of this book—the *generative adversarial network* or *GAN*—much as Baudelaire and other contemporaneous critics stood towards the introduction of photography: in a critical contemplation of a novel technology which seems to force us to revise the conceptual boundaries of our relations to the world. While the tools for creating this earlier generation of durable images were developed at the hands of scattered European inventors like Daguerre and Talbot, the tools of generative adversarial networks derive from a triple relationship between large institutions of academic research, government sponsorship, and high-tech corporations. This may inspire us to ask: will the future techniques, styles, and even standardization of GANs be determined by the intrinsic scientific and mathematical properties of these technologies themselves? Or, like photography, will they be determined by a variety of both amateur and professional practitioners, in interaction with audiences and one another?

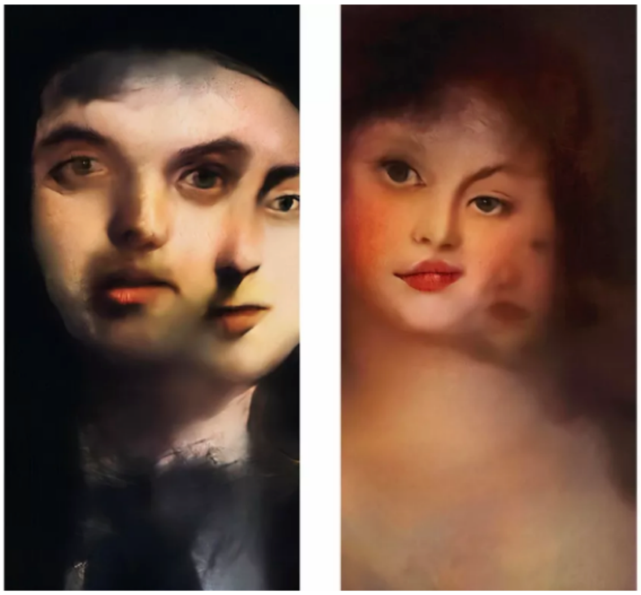

Figure 1: Mario Klingemann, Memories of Passersby I (2018).

This intuition of the potential cultural and intellectual impact of generative adversarial networks has been captured by one of the most prominent practitioners of GAN-based art, Mario Klingemann, in his use of the term *neurography*, and in his analogy we can see an outline of a path to understanding (see Fig. 1 for an example of Klingemann’s work, trained on a corpus of pre-19th century portrait paintings). First, if neurography is to become a respected art form—perhaps eventually with its own museums, as those of photography, holding retrospective revivals of prominent neurographic careers—we can, for the moment, decline the question of the supposed non-human *agency* of neural networks, and instead focus simultaneously on an increased *technical* and *social* understanding of these artists' machines. This in turn can enlighten not just our understanding of neurography with respect to photography (as critics like Baudelaire attempted to understand photography with respect to painting), but also provide new ways to interpret, and *think with*, this technological and/or artistic genre.

However, there are many potential 'levels' of such techno-social understandings. In the case of this essay we will not descend to the level of the programming languages and libraries (Python, Tensorflow, PyTorch, etc.) used to implement neurographic machines (which in their inert form can and will be called *models*), nor will we too closely examine the more abstract mathematical formalizations, published in research papers, which those libraries aim to implement. Instead, we will remain at the slightly more conceptual level of the so-called *architecture* of GANs, and include diagrams much as a guide to photography might provide a diagram demonstrating the physical flow of optical photography (in which, e.g., photons from a light source haphazardly reflect off surfaces, refract through a camera lens, alight on silver halide crystals, which are immersed in a sequence of chemical baths, magnified, and printed to paper). As we will see, the basic GAN architectures have a distinctive *interactional*, *dyadic* form—a relationship between a 'generator' network and 'discriminator' network—which distinguishes them from previous computer-generated art systems.

This dyadic and arguably 'pedagogical' form will then provide an occasion to reflect on something noticed early on by Alan Turing, but somehow forgotten in much of AI research the late 20th century; that in the drive to create artificial intelligence, we should recognize that "[i]t would be quite unfair to expect a machine straight from the factory to compete on equal terms with a university graduate. The graduate has had contact with human beings for twenty years or more. This contact has been modifying his behaviour pattern throughout that period. His teachers have been intentionally trying to modify it" [89]. This is to say that human learning is intrinsically *social* and also takes place over a long developmental period. In their (limited) emulation of these social and developmental qualities, then, GANs can also help us reflect on the claims of some contemporary AI practitioners to ultimately be able to reach some kind of "general" or human-level intelligence.

In this chapter, I will specifically be arguing that the form of the GAN's distinctive architecture—a duality between a system which classifies and a system which generates—has a close analogy with a heretofore wholly disconnected theory of social classification and cultural reproduction: namely, the French sociologist Pierre Bourdieu's notion of what he calls the *habitus.* Simultaneously cognitive and social, Bourdieu's *habitus* concept can, I will argue, provide a significantly deeper understanding of the novelty and conceptual appeal of GANs, especially for those who are concerned with the potential in machine learning for blindly reproducing cultural biases in society. In addition, because Bourdieu himself used his concept of the habitus to explore both literary and artistic fields, we can more easily explore the potential connections between social theory and the practical technoaesthetics of GANs.

In what follows, I will consider the architecture of the GAN, which is classically segmented into by a 'generator' and 'discriminator' network. I will then show how both connectionism (i.e. the use of neural networks) and Bourdieu's theory were equally, but separately, inspired by a rejection of cognitive and social theories based on rules. Once these technical and intellectual qualities have been established, I will consider the metaphoric structure of the GAN, one of a relationship between a productive, developing apprentice and a critical teacher. Then, I will raise the issue of the reproduction of *bias* in machine learning and how trained GAN models can be seen as a kind of partially embodied, material form of 'cultural capital'—the largest of which can be seen as kind of infinitely productive 'epistemic consumption object'. Finally, I will explore the formalism of (and/or metaphor of) game theory and games which are present in the technical formulation of GANs and the descriptions of Bourdieu's theory of the habitus and its fields of operation, respectively.

2 THE GAN ARCHITECTURE AS A SOCIAL AI

The re-emergence of the topic of artificial intelligence in the public sphere in the 21st century is not so much a revival but a recurrence—specifically, a recurrence of the distinctive computational architectures known as multilayer (and often *convolutional* and/or *recurrent*) neural networks which came of age in the late 1980s and early 1990s, in the wake of the publication of the so-called "PDP volumes" on what was then dubbed *Parallel Distributed Processing* [76]. The multilayer convolutional neural network (or *CNN*), in particular, was inspired indirectly by the so-called 'feed-forward' flow of information posited for the human visual system [48] and more proximately by a previous neural net architecture known as the Neocognitron [38], and was famously honed by Yann LeCun and others at Bell Labs for the purposes of recognizing numbers from small black-and-white bitmap images of handwritten digits, such as on bank cheques and other forms [55, 56]. This type of learning system—in which an artificial neural network is trained on a large amount of hand-labeled data and tested on a smaller quantity of data where the labels are hidden from the system—is known as *supervised learning*; as we will soon see, generative adversarial networks provide an interesting twist to this approach.[1]

The neural networks performing these classifications are conventionally referred to as *models*; but while they might be *inspired* by, e.g., neurological aspects of animal or human perception, they do not correspond to the classical conception of a mechanical or physical model which is designed to *represent* the structure and/or behavior of some real-world system.[2] Instead, neural networks for computer vision are so-called '*data models*' [4, 36, 53], which are

[1] For the purposes of this essay we will primarily constrain ourselves to discussing GANs which use convolutional neural networks and not recurrent neural networks (which are more common in dealing with sentences and documents in the field of natural language processing (NLP)); and we will limit discussion to supervised learning (in contrast to *unsupervised learning*, in which no explicit labels are provided for input data, and/or *reinforcement learning*, where networks dynamically adapt to a more complex system of rewards/punishment). It should be noted that AI researchers' notion of the possibility of learning without a teacher or in some otherwise "socially uncontaminated" way is itself ideological [7, 84].

[2] On representationalist ideologies of modeling vs. alternatives, see Morgan and Morrison [64] and Knuuttila [54].

less formalisms designed to have close correspondence with real-world cognition and are instead more like artifactual tools oriented to a casual, data-centric pragmatism: e.g., in the case of LeCun's CNNs, how well do they classify previously unseen images? As such, supervised multilayer neural network models can be thought of more like semiotic *engines*, which consume computational energy to transduce one type of sign (e.g. bitmaps of handwritten digits) to another (a symbolic integer from 0 to 9).

Notably, the 'parameters' or weight values of these models—the set of real-valued numbers which determine each stage of matrix multiplications which constitute a convolution—are not fixed beforehand. Instead, the model is *trained* through presentations of successive batches of labeled images, and depending on how bad its guesses (or *predictions*) are, proceeds to update those weight parameters in such a way that it will do better next time—performed by a standard technique specifically known as *backpropagation.* So unlike the 'difference engine' of Babbage (an early, hand-powered, prototypical computational device which calculated tables of polynomials) and the logical and procedural programming techniques which succeeded it, neural networks can be thought of, with reference to Derrida, as a kind of '*différance engine*' whose artifactual reproduction of future (or *deferred*) acts of classification is produced through many small observations of *difference*—namely, the mathematical distance between a guess and the true answer [41].[3]

Unlike other forms of machine classification prevalent in earlier data-mining regimes, the CNN—composed as it is primarily of a sequence of filtering and downscaling operations (see Fig. 2)—is careful to preserve and account for the spatial organization of its input at each layer. Each set of small 'filters' learned by each convolutional layer operates across its entire input—in two-dimensional sweeps of multiplication operations known as convolutions—and thus permits what mathematicians like to call a spatial 'invariance' with respect to different inputs. In this, they are similar to generic filters in Photoshop which successfully blur, or sharpen, or detect edges in an image without attending to the content of the image itself; the difference is that while a default blur/sharpen/edge-detection filter has fixed values, the filters of a CNN are gradually learned over time, using backpropagation [55, 75]. While it was in the 1990s hoped that this prototypical CNN architecture, known as 'LeNet' after its lead author, could eventually be applied to help classify larger (and full-color) images, researchers for many years found it difficult to successfully scale the successes of these systems up to larger tasks.

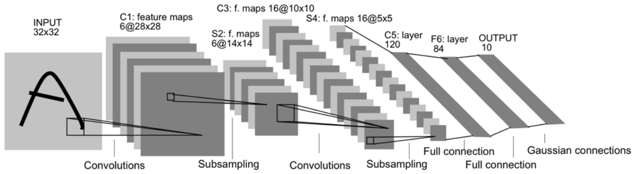

Figure 2: The 'LeNet' Convolutional Neural Network (CNN). From LeCun et al [56].

It was indeed only after the development of certain technical (and sociotechnical) innovations—including the use of graphics processing units or GPUs [86]; the large-scale distributed labeling of massive datasets by Mechanical Turk workers laboring below minimum wage [32]; and clever "tricks of the trade" such as Geoff Hinton's *dropout* technique —that these models took newer and deeper forms (i.e. with higher numbers of layers), such as VGG16 [83] and ResNet [46] which succeeded on more impressive labeling tasks (detecting 1000 different objects in 224x224 full-color images instead of detecting 10 digits within 28x28 grayscale images). So clever did these models seem that, for some, entire subfields of computer vision appeared to have become outdated overnight [28]. The news of these successes thus inspired the beginnings of modern-day AI hype, and these models began to be popularly referred to not as 'models' but as artificial intelligences or 'AIs' in their own right. Moreover, fueled by the successes of related reinforcement learning models on highly constrained yet complex tasks like old Atari games and the game of Go—in combination with an imaginative millenarian ideology—it was somehow thought in the mid-2010s by otherwise educated individuals that these multilayered classification models would evolve at an exponential rate to achieve superhuman intelligence. While the intensity of these beliefs has receded somewhat, the distinctive form of these architectures continue to capture the imaginations of new generations of students while continuing to perform well on various tasks across computer vision, natural language processing, and other fields.

The reason I have spent so much time in the retelling and re-describing of an example deep neural network architecture is to simply point out that the generative adversarial network, as devised by Ian Goodfellow in 2014, is composed of *two* such structures—or, if you will, two 'AI's. One network, the *discriminator*, is a classifier in the deep neural network lineage, which takes input images and successively performs a number of linear and nonlinear transformations to produce some numerical output—specifically, it is trained to output some number between 0 and 1 indicating the extent to which it thinks an input image is "real" (i.e. plausibly drawn from the training data) or not.[4] The other network, the *generator*, is a kind of horizontally flipped or inverted discriminator which, instead of transducing an input image to some vector of output values, transduces an input vector of values to some output image—a *generated* image, which is ideally similar to, but not copied from, the types of images observed in the training process (see Fig. 3). The training of what we call a GAN actually involves a back-and-forth training between these two models, whose weight values are not directly visible to each other; and as the discriminator gradually improves its ability to distinguish real images of digits from those produced by the generator, the generator in turn gradually improves its ability to concoct images of digits which can fool the discriminator. The simple task of this essay is to ask: if a discriminator neural network is considered 'an AI'—which, as mentioned, has inspired outlandish imaginaries of robotic sentience—and a generator neural network is also 'an AI', then shouldn't we think of generative adversarial

[3] The rather fascinating connections between postmodernism and connectionism could likely be profitably revived today for a new kind of 'digital humanities' which restores the primacy of the latter category over the former.

[4] The earliest GANs, such as those originally described by Goodfellow [44], used what are called *fully-connected* layers (imagine a single large "filter" which does not have to be "swept" across its input, as in a convolutional layer). Later developments such as the Deep Convolutional GAN or *DCGAN* [73] showed how the dyadic GAN architecture could be made to work with convolutional generators and discriminators.

networks as… *social*? The extent to which this claim holds or does not hold water is the subject of this essay.

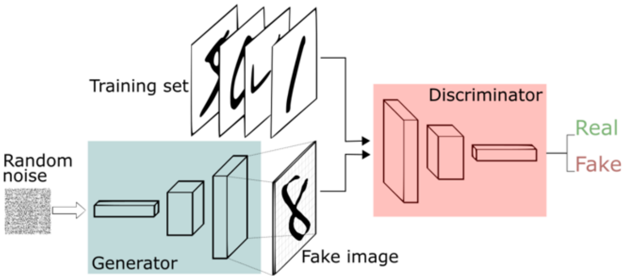

Figure 3: The basic GAN architecture.

As we will see, in some ways, the idea that GANs are 'social' is a powerful analogy with very intriguing connections to existing theories of social classification and cultural reproduction, and I will suggest that such associations may be sometimes implicitly, if not always consciously, at the core of the specific interest in GANs over other neural architectures. But in other ways, as I will show, GANs can be considered just as limited and methodologically hermetic as earlier models in cognitive science. In particular, even though connectionism (i.e. the use of neural networks) is often seen as a revolt against cognitivism (i.e. the ideology underpinning "old-fashioned" symbolic AI, which analogized the human mind to a rule-following computer program), deep learning models still operate largely in isolation from their surrounding material and social environment. In addition, the primary mathematical formalization of GANs in Goodfellow's original paper uses the framework of game theory, a particular conception of social interaction deriving from wartime research in statistics and economics [57], and at the conclusion of this chapter I will discuss the potential limitations of viewing GANs through this framework.

3 BOURDIEU'S HABITUS: GENERATIVE AND DISCRIMINATIVE

Graduate students in sociology for decades have at some point been confronted with the following notoriously challenging passage from the French sociologist Pierre Bourdieu's *Outline of a Theory of Practice* [11], in which the author, in the course of proposing a social theory which would move beyond the phenomenology of Merleau-Ponty and the structuralism of Lévi-Strauss, defines something known as the *habitus*:

> "The structures constitutive of a particular type of environment (e.g. the material conditions of existence characteristic of a class condition) produce *habitus*, systems of durable, transposable *dispositions*, structured structures predisposed to function as structuring structures, that is, as principles of the generation and structuring of practices and representations which can be objectively "regulated" and "regular" without in any way being the product of obedience to rules, objectively adapted to their goals without presupposing a conscious aiming at ends or an express mastery of the operations necessary to attain them and, being all this, collectively orchestrated without being the product of the orchestrating action of a conductor." [11, p. 72]

Bourdieu had long been concerned with the reproduction of social stratification, as illustrated in his earlier collaborations with Jean-Claude Passeron, *The Inheritors* [23] and *Reproduction* [22]. Both authors had shared the experiences of growing up in provincial regions of France and managing to reach the *grandes écoles*; they suspected that formal public schooling was far from egalitarian, and instead might be in fact responsible for reproducing the very stratification it might be expected to mitigate. The concept of the habitus was, in part, developed in order to help address this puzzle. It was, they argued, through the cultural inculcation of an embodied and partially unconscious *habitus*—this "durably installed generative principle of regulated improvisations" [11, p.78]—that, they argued, students from the upper classes are given an advantage which is only further reinforced throughout their educational trajectories.

How does the habitus operate? First, it works as an interiorization of the past, effected through socialization, which, secondly, allows one to carry out practical activities in everyday life, without necessarily being conscious of how those activities are generated by the mind and/or body. As such it concerns both 'the internalization of externality' and the 'externalization of internality', showing "the way society becomes deposited in persons in the form of lasting *dispositions*" but also how these dispositions can "guide them in their creative responses to the constraints and solicitations of their extant milieu" [91]. One can further distinguish between the *primary* habitus—those embodied dispositions acquired from one's close family during one's childhood—and a *secondary* habitus typically acquired through the explicit pedagogy of schooling [92].

Bourdieu's conception of the habitus can be difficult to grasp because, unlike a conventional cognitivist view of the mind as a largely inert information processor, the habitus is simultaneously both cognitive *and* social; in this view, one cannot disentangle isolated and 'structured' mental feats from generated and 'structuring' actions taken in the social world. These two mutually interlinked components have been described by the contemporary sociologist Omar Lizardo as corresponding to (1) "the habitus as a perceptual and classifying structure" and (2) "the habitus as a generative structure of practical action" [59]; put simply, the habitus is composed of both a classifier and a generator process, and thus, at first blush, appears to have the same basic architecture as a GAN.

What we want to explore in the following sections are the implications of this interesting (and clearly unintentional) isomorphism between Bourdieu's habitus concept and the architecture of GANs. By analogizing the generator and discriminator in a GAN model to the two unique functions of the habitus—and, later, by showing how both the habitus and the GAN are conceived by their authors, Bourdieu and Goodfellow, as part of a kind of strategic game—we can ask whether deep learning research itself is, or is not, independently rediscovering and reinventing a kind of social theory.

4 AGAINST RULES: STRUCTURALISM, CONNECTIONISM, AND CRITICAL TECHNICAL PRACTICE

Bourdieu developed the concept of habitus in part as a response to what for him was a long-standing question: "how can behaviour be regulated without being the product of obedience to rules?" [14, p. 65]. This question implicitly demonstrates the similarity of Bourdieu's ideas with a distinctive way of thinking that developed in response to the hegemony of a set of largely *symbolic* approaches to artificial cognition—focused around institutions like MIT and CMU in the United States—now known as "good old-fashioned AI" (GOFAI).[5] According to these latter symbolic AI proponents, intelligence can be modeled by conceptually disembodied "physical symbol systems" [65] which, through a combination of explicit rule-following, broader heuristics, and hierarchical planning, can find solutions to tasks in a broad "search space" [9].

By contrast, the critical responses to this worldview—along with its so-called "strong AI" version in which such systems can be considered intelligent or even conscious, and thus serve as a model for human cognition—came from philosophers and social scientists across the 1970s, 1980s, and 1990s with distinctive but overlapping perspectives. Hubert Dreyfus, inspired by Heidegger's *Being and Time*, suggested that orderly behavior might emerge without symbolic rule-following in an *embodied* and dynamic human environment [33]; Lucy Suchman emphasized the role of action as *situated* and context-dependent [87]; and Phil Agre attempted to implement *deictic* representations in software [1]. Each of these authors, then, understood that intelligence was intrinsically *indexical* [68, §2.227–2.308] and not something that could exist in a hermetic closed world; with Phil Agre in particular hoping that technical work in AI could be integrated with these views to create the possibility for a reflexive "critical technical practice" [2]—a concept which resonates in Bourdieu's own later search for a reflexive science of society [24].

Bourdieu's attempt, on his part, to understand situated and embodied social regularities without rule-following emerged as a response to the popularity of *structuralist* approaches in the social sciences, especially as articulated by the anthropologist Claude Lévi-Strauss in his analyses of the mental relations determining the diverse classification practices and mythical systems in various human societies [58]. Bourdieu saw structuralism as a canonical example of what he called 'objectivism': a worldview detached from the everyday practice of the agents under discussion, such as in Lévi-Strauss' previous use of genealogical data to explain the formal structure of kinship relations in different cultures. In addition, Lévi-Strauss had been inspired by the Bourbaki school of abstract mathematics and strove to view mythology as something which could be composed and manipulated as a kind of algebra, an arguably atemporal view which, for Bourdieu, seemed unable to account for contextual indeterminacies and change [16].

But interestingly (and in a way which is relevant to our argument), the mid-1980s resurgence in neural networks—which took place largely independently from the aforementioned critical AI literature of Dreyfus and Suchman—was *also* intellectually inspired by a revolt against rules-oriented logics; in a short essay preceding the publication of the main PDP volumes, David Rumelhart wrote:

> "Discussion of cognition, especially of language and of though often revolves around a discussion of *the rules of language* and *the rules of thought*… As neat as these accounts have seemed, there are serious problems with them. There are characteristic flaws in our reasoning — sometimes we don't follow the rules. Similarly, language is *full* of exceptions to the rules… It has seemed to me for some years now that the "explicit rule" account of language and thought is *wrong*" [74].

Rumelhart's pragmatic demonstration of his argument was to build a connectionist system which would attempt to learn the phonological representation of the past tense of English verbs [76]; the claims of relative success for this system inspired much debate and controversy [8, pp. 955–57]. It was soon argued by a different member of the PDP group, Paul Smolensky, that such a system proved that it was possible to work without recourse to rules and symbols and instead only depended on so-called 'subsymbolic' representations, in which input and output are represented as numerical vectors: as he put it, '[u]nlike symbolic tokens, these vectors lie in a topological space in which some are close together and others far apart" [85].[6] This description succinctly expresses the paradigmatic view of the *vectorization* of data [60] which, as it turns out, has since become standard practice in deep learning.

For some of the aforementioned humanistic critics of symbolic AI, connectionism seemed newly promising in a way which was aligned with their intellectual influences. As Hubert Dreyfus (with his brother Stuart) put it,

> "If multilayered networks succeed in fulfilling their promise, researchers will have to give up the conviction of Descartes, Husserl, and early Wittgenstein that the only way to produce intelligent behavior is to mirror the world with a formal theory in mind.… [n]eural networks may show that Heidegger, later Wittgenstein, and Rosenblatt were right in thinking that we behave intelligently in the world without having a theory of that world." [34, p. 35]

While we may today be indeed on the cusp of such an intellectual moment, the broader success of multilayer neural networks clearly did not happen for some time; in fact, the primary representatives of deep learning today, such as the aforementioned LeCun, were severely marginalized during much of the 1990s and 2000s [28]. By contrast, Bourdieu's theories around this time period were increasingly appreciated within U.S. sociology, which (along with the rest of the social sciences globally) rather studiously ignored (or were ignorant of) contemporaneous developments in machine learning methodology, such as the popular 'shallow" machine learning models for tabular data such as Support Vector Machines (SVMs) [30, 67, 78].

[5] Unless otherwise specified, I will use the term 'symbolic' in the sense of C.S. Peirce's *symbol* sign type [68, §2.227–2.308]—namely, signs which refer to their object in the *arbitrary* or *conventional* mode of de Saussure's signifier-signified relationship (e.g. the string of characters *'arból'* referring to a tree) [79], and in opposition to iconic or indexical reference.

[6] The ensuing debate, about whether or not 'subsymbols' were actually symbols, reflected the semiotic limitations of cognitive scientists, who did not have terms like *indexical* or *deictic* as used by Suchman and Agre.

5 TEMPORALITY, DEVELOPMENTAL THEORY, AND MIMESIS

Another group which found inspiration in the late-80s connectionist wave were developmental psychologists who saw in the San Diego-based cognitive scientist Jeff Elman's *recurrent neural networks* (or *RNNs*)—a simple architecture whose output iteratively loops back onto its input, with one or more 'hidden states" updating internally—a useful metaphor for the human process of learning over time [37, 71]. For these psychologists, cognitivism and symbolic AI had obscured both the temporal and social process of knowledge acquisition in practice, but the progressive training process of neural networks—beginning with randomly initialized weight values and achieving better performance with more 'epochs" of presentation data—recalled the more 'dynamic' *epigenetic* psychological theories of human cognitive development of Jean Piaget or Lev Vygotsky. Some even noted analogies between the so-called "U-shaped" learning curves in RNNs and those measured in human learning processes [70].

But while this exploration of RNNs reintroduced the relevance of temporality to cognitive theories of learning, such artifactual learning was not significantly 'social', composed as it is of the repeated presentation of 'supervised' training data (with 'right' answers and strict penalization for guessing the wrong answer); by contrast, the dyadic architecture of GANs explicitly suggests a more nuanced relationship between a 'teacher' (who judges the 'output' of a student) and a 'student' (who is only concerned with learning how to please the teacher). Specifically, like most multilayer neural networks, the weight values for both the generator and discriminator in a GAN are randomly initialized, so that at the beginning of the training process, the generator is "dumb" and only knows how to create 'random' noisy images, and the discriminator is equally "dumb" and cannot, e.g., reliably distinguish between real images of digits and the garbage produced by the generator. But with every back-and-forth step of training, the discriminator learns a little more about real images of digits; and the generator learns a little more about what the discriminator thinks is a valid digit. As such—at least in the original GAN formulation—the discriminator is, ideally, just a little bit "ahead" of the generator at any given training step.[7] One can analogize this situation to so-called peer learning [88], in which learning from someone near or just above one's knowledge level can be productive by virtue of taking place within what Vygotsky [90] called the *Zone of Proximal Development* (ZPD).

While Bourdieu only infrequently mentions Piaget or Vygotsky directly, Lizardo [59] has explained the context of Bourdieu's early writings on the habitus and shows that Bourdieu was influenced by Piaget's "unique blend of structuralism and developmental cognitive psychology". One can look at the following suggestive passage from Piaget's 1971 book *Genetic Epistemology*:

> "..I think that human knowledge is essentially *active.* To know is to assimilate reality into *systems of transformations.* To know is to transform reality in order to understand how a certain state is brought about. By virtue of this point of view, I find myself opposed to the view of knowledge as a copy, a passive copy, of reality. . . . Knowing reality means constructing systems of transformations that correspond, more or less adequately, to reality. . . Knowledge, then, *is a system of transformations that become progressively adequate*" [69, p. 15]. (emphasis added)

While Piaget does not give a name to this type of learning process, it should again remind us of GAN generators, which do not learn a simple copying procedure but instead learn a "system of transformations" which—as repeatedly confronted by the discriminator's judgments—"become[s] progressively adequate". Similarly, Bourdieu describes "the process of acquisition" of the habitus as a "practical *mimesis*", as explicitly opposed to "an *imitation* that would presuppose a conscious effort to reproduce a gesture, an utterance or an object explicitly constituted as a model. . . " [14, p. 73]. We can thus see a correspondence between the training of the habitus and the way a GAN learns to generate and/or classify without recourse to a 'conscious' or explainable representation. We can also see that, in the case of generative networks, although there is indeed an 'explicit constitution of a model', the images generated are definitively not an imitative, pure replication of examples in the training data, because the generator network *never sees* the training data—it only receives judgments on plausibility from the discriminator. In this way, GANs, more so than other types of generative models, seem closer to Bourdieu's *mimesis* than his *imitation.*[8]

At the same time, the use of the term *mimesis* highlights one of the deeply un-social aspects of GANs, which is that the generator typically only interacts with a discriminator, and does not learn from *other* networks. In this sense, the mimesis of a GAN is nowhere near to the implied complexities of the theories of mimesis of, e.g., René Girard, who sees mimetic behaviors (and desires) as proximally dependent on observations of others' behaviors and desires in a potentially unbounded (and politically perilous) recursiveness [40]. While more recent alternative GAN architectures have attempted to introduce designs with multiple discriminators [35] and multiple generators [39], researchers have yet, as of this writing, to conceive of generative networks as agents in a *community* of artists and critics or as part of some broader sociological "art world".[9]

6 BIAS, CULTURAL CAPITAL, AND THE EPISTEMIC CONSUMPTION OBJECT

The appeal of the specific type of mimetic reproduction in GANs, then, is that they can produce new images which *appear* to be drawn from the training data, but are not in fact imitative or overt copies of images in the training data. In a potentially related manner, Bourdieu's habitus is what he calls both *durable* and *transposable*: relatively stable, but capable of being deployed dynamically in novel

[7] In other formulations, such as the *Wasserstein GAN* [3], the discriminator is trained more extensively than the generator at each step.

[8] Bourdieu's distinction between imitation and *mimesis* and the comparable behavior of GANs have an interesting parallel with the literature on data models vs. "structurally isomorphic" models in the philosophy of science [54]. In both cases, a simple perspective of isomorphic 'copying' is confronted with a more 'ungrounded' process of reproduction which can occur without direct reference to the 'reality' supposedly being modeled.

[9] Although we do not explore the wide cornucopia of GAN variants here, it is worth mentioning the intriguing and impressive *CycleGAN* model which consists of two discriminator-generator pairs simultaneously learning to, e.g., translate horses into zebras and zebras into horses in still images as well as video [94].

and varying social situations—of "being capable of becoming active within a wide variety of theatres of social action" [61].[10] The increased discussions in the past few years around the topic of *bias* in machine learning [5], I would argue, can be understood as concerns about the increased materialization of precisely this kind of durability and transposability, often at the hands of machine learning techniques. Earlier in the decade, for example, (shallow) neural networks were (and still are) used for training what are called *word embeddings* (or *word vectors*); these embeddings, trained on large corpuses of text, represent individual words (detached from their greater context) as high-dimensional vectors of real numbers for use in other machine learning models.[11] Two papers independently discovered that these vectors represent—and, when deployed on widely used platforms, can help reproduce—cultural biases, such as female names being associated more with 'family' words than 'career' words [10, 27]. Such biased embeddings could be reflected in search results and other interactional situations, and these works raised concerns precisely because of the increased standardization of word embedding data in various software products and the increased facility on the part of Google, Facebook, etc. to blindly deploy such biased interpretations at scale in everyday sociotechnical life.

In recent years, studies of bias in machine learning have examined not just models built on structured, tabular data (as in the case of the U.S. recidivism classification algorithm known as COMPAS) or text (as in word embeddings), but have addressed the potential discriminatory aspects of convolutional neural networks in facial recognition [49], gender classification [26], and in object recognition in images in general [66]. It is here that we can theorize *generative* models (including GANs) as a distinctive genre of bias in machine learning which emphasizes a limited form of *practical action* which takes advantage of, but is in part distinct from, a "trained" perceptual and classifying structure. Put simply, GANs reproduce bias not just through their facility for stereotyped classification, but through their potential for generating *new* biased data.[12] They differ from the "algorithms of oppression" of Google's search and recommendation engines, whose biases also exist, but which must be taken up and reproduced by humans in the loop taking practical action of their own. In Bourdieu's terms, traditional machine learning biases can be considered *durable* but not *transposable*. GANs bring us closer to this kind of transposability—but certainly not all the way, for while a standard GAN model can produce novel objects similar to its training data, it does so identically in each contextual environment in which it is deployed.

One way to understand this durable-but-not-precisely-transposable quality of GANs is to see them through the lens of a different concept in Bourdieu's oeuvre, that of *cultural capital*. This category is distinguished by Bourdieu from the more conventional notions of *economic capital* (i.e., something which can be converted directly to monetary value, like cash or private property) or *social capital* (which involves the relationship of agents to each other and/or the resources available to such agents by virtue of their position in a social network) [13]. *Cultural* capital, by contrast, can be loosely defined as locally advantageous forms of embodied and/or externalized knowledge—or, more accurately, 'know-how'. It is this genre of capital which, Bourdieu initially argued, is created through the relationship between the family and the educational system, and which can come in three general forms: *incorporated* (embodied as part of the habitus), *objectivized* (manifested in material form), or *institutionalized* (sanctioned by an institution of some kind) [12]. In the case of a schooling institution, one could say that students arrive with incorporated cultural capital acquired from their parents and upbringing (embodied in the aforementioned *primary* habitus); they would subsequently develop relationships with forms of objectivized cultural capital like textbooks, educational software, and the apparatuses of testing; and in turn they are rewarded with institutionalized cultural capital such as qualifications and degrees (themselves symbolically objectivized in the form of a diploma).

Such a conceptual framework now suggests a question: what type of capital do trained neural networks, and GANs specifically, represent? The many well-known deep learning models whose architectural source code is available online—and for which pretrained models are often similarly made available—should probably not be considered as a form of economic capital or social capital. By contrast, the ability of neural networks to efficiently classify objects in images *sounds* like a kind of cultural capital; and Bourdieu indeed has described—in his essay entitled "Outline of a Sociological Theory of Art Perception"—how 'art competence' can be understood as "the preliminary knowledge of the possible divisions into complementary classes of a universe of representations" [18]. From today's perspective, this sounds a lot like a (discriminator-type) CNN, which can (in the case of models trained on the large corpus known as ImageNet) distinguish among 1000 different classes of objects in 224x224-pixel color images; while for Bourdieu, the operating metaphor was the ability of a spectator to identify the "author" of a given painting, a skill which varies among those with different amounts of cultural capital (especially that inculcated by the family and different types of schooling). He continues:

> "A mastery of this kind of system of classification enables each element of the universe to be placed in a class necessarily determined in relation to another class, itself constituted by all the art representations consciously or unconsciously taken into consideration which do not belong to the class in question. The style proper to a period and to a social group is none other than such a class defined in relation to all the works of the same universe which it excludes and which are complementary to it" [18, p. 221].

In the case of either a) a CNN-like neural network classifier or b) a so-called *class-conditional GAN*—which learns to generate novel images of objects from a variety of categories, as in Mirza and Osindero [63]—this kind of cultural capital is arguably *incorporated*, especially if we consider the flowing form of neural network architectures as a kind of embodiment; from Bourdieu's perspective,

[10] Machine learning practitioners might say transposability is a kind of *generalizability*.
[11] While the use of neural networks for the construction of word embeddings is not strictly necessary [42], deep neural networks are consistently used for more recent iterations of embeddings which take greater amounts of so-called "context" (word co-occurrence at the scale of paragraphs and documents instead of just short sequences) into account.
[12] This issue was raised in an interview with Mario Klingemann, regarding his use of so-called "Old Masters" portraits which literally represent the historical dominant classes. Klingemann acknowledged this implicit reproduction of stratification and replied "[b]y accident I do political art" [72].

however, the relatively limited corporeal dynamism of said architectures would probably restrict such an identification. But GANs are certainly *objectivized*, i.e. converted into an inert and stabilized form; and the *institutionalization* of the cultural capital of GANs is a process currently under way, with recent attempts by Christie's, Sotheby's, and various smaller galleries in London and Paris to deploy their own instutionalized value towards the performative conversion of GANs into economic capital. As such, we can see that while machine learning models can be implicated in a process of objectivized and therefore biased cultural reproduction, they still require the assistance of human agents in society with respect to embodiment and institutionalization.

In addition, Bourdieu's above description of how one can learn a consistent 'style' through a process of relational exclusion could be seen as similar to the training process of these conditional GANs, which progressively learn to imitate different image classes by virtue of the discriminator network's progressive ability to distinguish between those classes. It is specifically interesting, however, to consider conditional GANs which learn from large datasets like ImageNet—a massive 1.4 million-image dataset of pictures downloaded from the web—such as the recent BigGAN [25]. The BigGAN model architecture, which can generate novel images from any class in the ImageNet dataset (e.g. leopards, container ships, mushrooms, dalmatians, etc.) starting with a single vector of 128 real-numbered values (i.e., the "latent vector") as input, thus has the potential to produce a practically uncountable number of possible images. It is an object which has thereby absorbed a massive amount of cultural capital, in the form of information about the vast space of amateur and professional photography of all the different classes in the ImageNet classification system (a dataset itself with known geographical and cultural biases [81]); but it is also an object which in its generative capacity appears to *exceed* (or demonstrate the unquantifiable quality of) the cultural capital absorbed.[13]

Brock et. al.'s BigGAN model, which easily generates images of a certain structural coherence but often with have a profound alien (and alienating) quality can thus be understood as a very particular type of entity known by a certain school of "postsocial" sociological theory as an *epistemic consumption object*: something which is "materially elusive" and which has a "lack of ontological stability" that renders it "a continuous knowledge project for consumers" [95]. One such object is the world of real-time financial markets—often referred to holistically as "*the* market"—an entity which can never be fully observed and yet provides an endless fount of information as it is probed by communities of investors and traders. The understanding of BigGAN as an epistemic consumption object is best illustrated by Joel Simon's 2018 site *ganbreeder.app*, in which users can interactively "explore" the latent vector space of the BigGAN model and use a form of genetic algorithms to "hybridize" said vectors [82]. Such exploration, performed manually, allowed Mario Klingemann to improvise a virtual and aleatory "tour" of this latent space when the model was first released (see Fig. 4); the apparent pragmatic infinity of such a generative object can help illustrate some of the challenges in the potential for interrogation of (and/or making 'interpretable') neural models' underlying and enculturated biases.

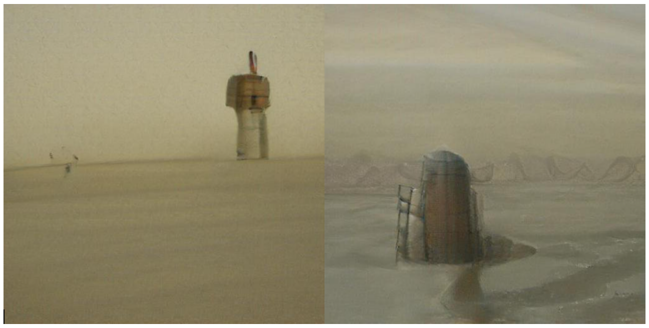

Figure 4: "[A] hideaway in the wastelands of #BigGAN" [52].

7 THE LIMITS OF GAME THEORY

One of the more intriguing and suggestive qualities of Goodfellow's original GAN formulation is its use of the formalisms of *game theory*, a subfield of economics founded with von Neumann and Morgenstern's 1944 book *Theory of Games and Economic Behavior*. It is intriguing in part because it helps motivate the otherwise strange, dyadic GAN architecture with reference to one of the only conceptions of the social with which computer scientists are typically familiar; but also because, as we have seen, the language of Bourdieu often uses economic metaphors itself (as in the concept of cultural capital). And in fact, Bourdieu often spoke of the way the habitus learns to both classify and generate activity—which takes place in a "relatively autonomous" social arena he calls the *field*—as a process of incorporating "a 'feel for the game"' [14, p. 66]. He describes this *field* as "a space defined by a game offering certain prizes or stakes", which is in a dual relationship with the habitus, defined as the "system of dispositions attuned to that game" [19, p. 18]. However, Bourdieu stressed that unlike a board game or a football game in which the rules can be consciously understood as "an arbitrary social construct" by its players, these fields are "the products of a long, slow process of autonomization" in which "one does not embark on the game by a conscious act, one is born into the game, with the game" [14, p. 67].

In an effort to finally resolve our primary question—namely, to what extent GANs are or are not like the habitus (in a given field)—the final questions I wish to address, thus, are somewhat 'bidirectional':

- To what extent can the GAN training process be understood in terms of games (and/or game theory)?
- To what extent can the habitus/field relationship be understood in terms of games (and/or game theory)?

In both cases, the primary authors in question (Goodfellow for GANs and Bourdieu for the habitus) use game and/or game-theoretic metaphors extensively, and the purposes of this final section is to judge their relative appropriateness and/or equivalence.

In his initial paper on GANs, Ian Goodfellow describes the GAN architecture as a "two-player minimax game with value function

[13] The 'absorption' of the conventions of amateur photography in an artifact like BigGAN is a topic of some potential interest, and is related to an earlier work by Bourdieu from 1965 entitled (in English) *Photography: A Middle-Brow Art* [15].

V(G,D)", meaning that there is a single abstract function whose output value the discriminator is trying to maximize and which the generator is trying to minimize: namely, the ability of the discriminator to distinguish between images drawn from the training data and images created by the generator [44].[14] In the value function specified by Goodfellow, the discriminator's optimal situation is when it can consistently output 1.0 for real data and 0.0 for fake data; and the generator's optimal situation is that the discriminator outputs "0.5" for all data, meaning that it is "maximally confused" [31]. This function follows the logic of a "minimax game", developed in the 1920s by John von Neumann, a Hungarian mathematician and polymath known for his contributions to physics and early computing; he argued that in a so-called *zero-sum* game (in which the gain of one player is equal to the losses of the other), the optimal strategy of both players is to attempt to *minimize* their *maximum possible* losses at every turn—hence, "minimax".[15]

However, the type of game discussed by von Neumann in the first half of the 20th century—typically represented as a small table called the *payoff matrix* which enumerates the wins and losses depending on two players' strategic actions—and the type of 'game' involved in the training of a GAN have some significant differences. Specifically, because GANs are trained in an alternating turn-taking mode, they correspond to what is called the game-theoretic "extensive form" or "dynamic form" as opposed to the traditional "matrix form" [45, p. 45]. In addition, the 'action' taken in each turn—the generation of new data points by the generator, or the assertion of fake-to-real judgments by the discriminator—obscures the rather radical transformation of the entire generator or discriminator agent at the hands of the backpropagation algorithm, potentially updating millions of different parameters in every round of training. As such, the value function *V(G,D)* is not stable but in fact dynamically changing at every timestep. This dynamism means that it can be difficult to successfully converge on what is called a *Nash equilibrium*—a state in which the generator can no longer improve based on the discriminator's judgments, and in which the discriminator cannot improve at distinguishing real and fake images [43, 77].

The basics of game theory—bringing together *players*, *strategies*, and *preferences* reflected in value-laden *payoffs*—can thus be thought of as a kind of attempt to formalize a "theory of interdependent decision making" [29, p. 3]. It brings with it a set of assumptions, some implicit, and some explicit, which may or may problematize it with respect to more complex theories of human behavior. As laid out by Hargreaves-Heap and Varoufakis [45, pp. 7–33], these assumptions include:

- *Individual action is instrumentally rational*: i.e., each agent has a *preference ordering* of outcomes, highlighted by the way the value function (also sometimes called a *loss function* or *utility function*) produces a single, unidimensional value which (in the case of GANs) describes the extent to which the generator fooled the discriminator or the extent to which the discriminator successfully classified real vs. fake images. (This assumption may be familiar from the stereotype of *homo economicus* who strives only to maximize their utility function.)
- *Common knowledge of rationality*: each agent in a game-theoretic scenario chooses a strategy based on the assumption that the *other* agents are operating in the manner described above, i.e., as guided by a single-dimensional utility function which determines the preferences for their actions.
- *Knowledge of the rules of the game*: each agent is assumed to be familiar with the full spectrum of possible actions and their equivalent payoffs.
- *Segregation of the rules of the game from actions taken*: the rules of the game are fixed and cannot be affected by actions taken, nor can the rules themselves affect the preference ordering for particular actions.

While each of these assumptions are by definition involved in the game-theoretic formulation of GANs, not all of these assumptions are acceptable to sociologists like Bourdieu whose theories simultaneously rely on both the habitus—that classifying and generative "structuring structure"—and the *field* in which said habitus is deployed (despite Bourdieu's characterization of the field as a kind of game-like arena). While we earlier deferred the question of autonomy and agency of generative networks, we can now contrast the limitations of the GAN architecture with the artistic and literary fields analyzed by Bourdieu to better understand the limits of these "AIs" in comparison to those of human artists. For while (as previously mentioned) connectionism was originally posited as a rejection of a rule-oriented cognitivism, the dependency of neural network training on loss/utility functions means that deep learning and GANs maintains a very close link with instrumental rationalism. The GAN, however, improves on the *fixed* loss functions of traditional supervised learning with a training process that gradually *learns* a loss function (i.e. as the parameters of the discriminator and generator evolve).[16]

Regardless, for Bourdieu, such a utility-based approach to artistic creation could not be more crude when compared to its social reality: utilitarianism is, for him, "the degree zero of sociology", by which he means an isolated, inert, and amodal—and therefore not particularly sociological—starting point [19, p. 76]. To truly model a generating agent in an empirically plausible 'artistic' field would be considerably more complex and could never be reduced to instrumental rationality, as demonstrated in this passage describing how to approach the production of music:

> What makes [the study of the economics of music production] so difficult is that, in the field of cultural goods, production implies the production of consumers, that is to say, more precisely, the production of the taste for music, the need for music, belief in music. To give an adequate account of that… would mean analysing the whole network of relationships of competition and complementarity, complicity in competition, which hold together the whole set of agents concerned, famous and unknown composers

[14] The specific minimax equation used in Goodfellow et al. [44] is $\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$. The discriminator wants to maximize the first term (the accurate judgment of multiple images $\boldsymbol{x}$ drawn from a conceptually infinite set of 'real' images) and minimize the second (the incorrect judgment of multiple images drawn from the very large set of images which can be created by the generator, initialized with some random vector $\boldsymbol{z}$).

[15] For a history of game theory which includes the cultural context of early 20th-century Hungary and World War II, see Leonard [57].

[16] The idea of GANs as representing a "learned loss function" can be attributed to Philip Isola [50] and the Berkeley lab of Alexei Efros more generally.

> and performers, record producers, critics, radio producers, teachers, etc., in short, all those who have an interest in music and interests that depend on music, musical investments – in both the economic and psychological senses – who are caught up in the game and taken in by the game" [19, p. 106].

As such, we suggest that while GANs can be seen as representing an artistic habitus in a kind of limited 'supervised' framework, this emulated habitus does a poor job at incorporating the true social complexities of an artistic field. For Bourdieu, actors in artistic fields past and present are not subject to a single all-powerful 'discriminator'—not only are there multiple critics (including one's peers), but multiple critical spheres, each of which compete with each other for authority.[17] Bourdieu has analyzed precisely the emergence of this "plurality of competing cults with multiple uncertain gods" in his close readings of the 'symbolic revolution' wrought by Édouard Manet in the 1860s with now-famous paintings such as *Le déjeuner sur l'herbe* (1863) and *Olympia* (1863) [17, 21]. The case of Manet, which dynamically transformed the way art is interpreted and valued, allows Bourdieu to reflexively interrogate his own theories of cultural reproduction [62]. It suggests that a truly *autonomous* artist would be one who can not only work *within* the learned 'rules of the game' but to overthrow them altogether.

Finally, because of Bourdieu's interests in both artistic and economic fields, he is particularly concerned with the emergence in the 19th century of a so-called *bohemian* culture, which is characterized primarily by its inversion of financial incentives, in which failure is a kind of success, and "selling out" (i.e. maximizing profit) worst of all:

> "The game of art is, from the point of view of business, a game of 'loser takes all'. In this economic world turned upside down, one cannot conquer money, honours (it is Flaubert who said that 'honours dishonour')… in short, all the symbols of worldly success, success in high society and success in the world, without compromising one's salvation in the hereafter. The fundamental law of this *paradoxical* game is that one has an interest in disinterestedness: the love of art is a crazed love [*l'amour fou*], at least when one considers it from the viewpoint of the norms of ordinary, 'normal' world put on to the stage by the bourgeois theatre." [20, p. 21]

This is to point out that if GANs—in their formulation as relentless optimizers of a loss function—intentionally or unintentionally replicate an economistic rational actor model, then they may be fundamentally at odds with the "value function" of many human artists-in-development across history. To this extent, the GAN habitus differs radically from the cultural norms of an originary bohemia but not necessarily from that of the contemporary art world, which—as we have indeed seen with the auction market for generative neural network art, and with 'residencies' provided by, e.g., Google for AI-related artists—has come to incorporate commercial incentives while at the same time preserving normative aspects of an *avant-garde* field.

Because of these fundamental incompatibilities of the GAN training methodology and the sociological intricacies of the artistic field, however—whether it be the relationships of multiple agents in the space, the assumptions of the game-theoretical framework, or the specific paradoxes of the loss or utility function—we can argue, along with Hertzmann (2018), that such models/algorithms cannot be considered as 'autonomous creators' at present, and to do so would involve taking into account far more of the fundamentally social qualities of artistic production and reproduction into the training process.[18] While some developments in reinforcement learning appear to be moving in an intriguingly 'field-like' direction—of a multi-agent framework where each agent itself models other agents as in Jaques et al. [51] — for now, the future of neurography will no more involve the autonomous GAN than the history of photography featured the autonomous camera.

8 CONCLUSION

In this chapter I have attempted to show that the novel dyadic architecture of generative adversarial networks (GANs) can be fruitfully understood as potentially corresponding–and sometimes, less potentially corresponding—to the simultaneously cognitive and social *habitus* of Pierre Bourdieu; and that these (and related) architectural developments in the "revolution" of deep learning, primarily understood as a scientific and technical achievement [80], might be understood as constituting a new stage in computing which intentionally or unintentionally constitutes a nascent, independent reinvention of social theory. Through this genre of 'interactional' machine learning models, we can potentially see classic and contemporary sociological theory through a new lens; and, inversely, for those inclined, through explorations in sociological theory one can—perhaps surprisingly, and perhaps to a greater degree than machine learning engineers themselves—more clearly understand both the novelty and potential limitations of artificial intelligence.

ACKNOWLEDGMENTS

Thanks to Noortje Marres, Etienne Ollion, Matthew Berland, Cristián Bravo, and Peli Grietzer for comments and discussion.

REFERENCES

[1] Philip E. Agre. 1997. *Computation and Human Experience.* Cambridge University Press.

[2] Philip E. Agre. 1997. Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*, Geoff Bowker, Les Gasser, Susan Leigh Star, and Bill Turner (Eds.). Lawrence Erlbaum Associates.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *arXiv:1701.07875 [cs, stat]* (Jan. 2017). http://arxiv.org/abs/1701.07875 arXiv: 1701.07875.

[4] Daniela M. Bailer-Jones and Coryn A. L. Bailer-Jones. 2002. Modeling Data: Analogies in Neural Networks, Simulated Annealing and Genetic Algorithms. *Model-Based Reasoning* (2002), 147–165. https://doi.org/10.1007/978-1-4615-0605-8_9

[5] Solon Barocas and Andrew Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (June 2016), 671. https://doi.org/10.15779/Z38BG31

[6] Charles Baudelaire. 1956. The Modern Public and Photography. In *The Mirror of Art: Critical Studies.* Doubleday, Garden City, N.Y., 227–233.

[7] Anja Bechmann and Geoffrey C Bowker. 2019. Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media.

[17] As Loïc Wacquant, a former student and collaborator of Bourdieu, puts it, "every component involved in the forging of habitus is quintessentially collective" [93].

[18] Indeed, to more fully incorporate Bourdieu's logic of practice would imply obliterating the currently strong distinction in machine learning between the training process and deployment, and considering only fully 'on-line' models [47].

Big Data & Society 6, 1 (Jan. 2019), 2053951718819569. https://doi.org/10.1177/2053951718819569

[8] Margaret Boden. 2006. *Mind as Machine: A History of Cognitive Science.* Oxford University Press, Oxford.

[9] Margaret A. Boden. 2014. GOFAI. In *The Cambridge Handbook of Artificial Intelligence.* 89–107. https://doi.org/10.1017/CBO9781139046855.011

[10] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv:1607.06520 [cs, stat]* (July 2016). http://arxiv.org/abs/1607.06520 arXiv: 1607.06520.

[11] Pierre Bourdieu. 1977. *Outline of a Theory of Practice.* Cambridge University Press.

[12] Pierre Bourdieu. 1979. Les trois états du capital culturel. *Actes de la recherche en sciences sociales* 30 (1979), 3–6.

[13] Pierre Bourdieu. 1986. The forms of capital. In *Handbook of Theory and Research for the Sociology of Education*, J. Richardson (Ed.). Greenwood Press, New York.

[14] Pierre Bourdieu. 1990. *The Logic of Practice.* Stanford University Press, Stanford, Calif.

[15] Pierre Bourdieu. 1990. *Photography: A Middle-brow Art.* Polity Press, Cambridge.

[16] Pierre Bourdieu. 1990. The Scholastic Point of View. *Cultural Anthropology* 5, 4 (1990), 380–391. https://www.jstor.org/stable/656183

[17] Pierre Bourdieu. 1993. Manet and the Institutionalization of Anomie. In *The Field of Cultural Production: Essays on Art and Literature.* Polity Press, Cambridge, 238–253.

[18] Pierre Bourdieu. 1993. Outline of a sociological theory of art perception. In *The field of cultural production: essays on art and literature.* Columbia University Press, New York, 215–237. http://web.mit.edu/allanmc/www/bourdieu3.pdf

[19] Pierre Bourdieu. 1993. *Sociology in Question.* SAGE Publications Ltd.

[20] Pierre Bourdieu. 1996. *The Rules of Art: Genesis and Structure of the Literary Field* (new ed edition ed.). Polity Press, Cambridge.

[21] Pierre Bourdieu. 2017. *Manet: A Symbolic Revolution.* Polity Press, Malden, MA.

[22] Pierre Bourdieu and Jean Passeron. 1990. *Reproduction in Education, Society and Culture* (second edition ed.). SAGE Publications Ltd, London ; Newbury Park, Calif.

[23] Pierre Bourdieu and Jean Claude Passeron. 1979. *The Inheritors: French Students and Their Relation to Culture.* University of Chicago Press, Chicago.

[24] Pierre Bourdieu and Loïc Wacquant. 1992. *An Invitation to Reflexive Sociology.* Polity Press, Chicago.

[25] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv:1809.11096 [cs, stat]* (Sept. 2018). http://arxiv.org/abs/1809.11096 arXiv: 1809.11096.

[26] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency.* 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[27] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (April 2017), 183–186. https://doi.org/10.1126/science.aal4230

[28] Dominique Cardon, Jean-Philippe Cointet, and Antoine Mazières. 2018. Neurons Spike Back: The Invention of Inductive Machines and the Artificial Intelligence Controversy. *Réseaux* 5, 211 (2018), 173–220.

[29] Andrew M. Colman. 1995. *Game Theory and its Applications.* Routledge, Oxford England ; Boston, Mass.

[30] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning* 20, 3 (Sept. 1995), 273–297. https://doi.org/10.1023/A:1022627411411

[31] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath. 2018. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine* 35, 1 (Jan. 2018), 53–65. https://doi.org/10.1109/MSP.2017.2765202

[32] Jia Deng, Wei Dong, Richard Socher, Li-jia Li, Kai Li, and Li Fei-fei. 2009. Imagenet: A large-scale hierarchical image database. In *In CVPR.*

[33] Hubert Dreyfus. 1972. *What Computers Can't Do.* MIT Press.

[34] Hubert L. Dreyfus and Stuart E. Dreyfus. 1988. Making a Mind versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint. *Daedalus* 117, 1 (1988), 15–43. http://www.jstor.org/stable/20025137

[35] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. 2016. Generative Multi-Adversarial Networks. *arXiv:1611.01673 [cs]* (Nov. 2016). http://arxiv.org/abs/1611.01673 arXiv: 1611.01673.

[36] Paul N. Edwards. 2013. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming.* MIT Press, Cambridge, Massachusetts.

[37] Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science* 14, 2 (March 1990), 179–211. https://doi.org/10.1207/s15516709cog1402_1

[38] K. Fukushima. 1980. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36, 4 (1980), 193–202.

[39] Arnab Ghosh, Viveka Kulharia, Vinay Namboodiri, Philip H. S. Torr, and Puneet K. Dokania. 2017. Multi-Agent Diverse Generative Adversarial Networks. *arXiv:1704.02906 [cs, stat]* (April 2017). http://arxiv.org/abs/1704.02906 arXiv: 1704.02906.

[40] René Girard. 1979. Mimesis and violence: Perspectives in cultural criticism. *Berkshire Review* 14, 9-19 (1979).

[41] Gordon G. Globus. 1992. Derrida and connectionism: différance in neural nets. *Philosophical Psychology* 5, 2 (Jan. 1992), 183–197. https://doi.org/10.1080/09515089208573055

[42] Yoav Goldberg. 2017. *Neural Network Methods in Natural Language Processing.* Morgan & Claypool Publishers, San Rafael.

[43] Ian Goodfellow. 2016. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv:1701.00160 [cs]* (Dec. 2016). http://arxiv.org/abs/1701.00160 arXiv: 1701.00160.

[44] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]* (June 2014). http://arxiv.org/abs/1406.2661 arXiv: 1406.2661.

[45] Shaun Hargreaves-Heap and Yanis Varoufakis. 2004. *Game Theory: A Critical Introduction* (2nd ed.).

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]* (Dec. 2015). http://arxiv.org/abs/1512.03385 arXiv: 1512.03385.

[47] Steven C. H. Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. 2018. Online Learning: A Comprehensive Survey. *arXiv:1802.02871 [cs]* (Feb. 2018). http://arxiv.org/abs/1802.02871 arXiv: 1802.02871.

[48] D. H. Hubel and T. N. Wiesel. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology* 160, 1 (Jan. 1962), 106–154.2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1359523/

[49] Lucas D. Introna and Helen Nissenbaum. 2009. Facial Recognition Technology: A Survey of Policy and Implementation Issues,.

[50] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2016. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv:1611.07004 [cs]* (Nov. 2016). http://arxiv.org/abs/1611.07004 arXiv: 1611.07004.

[51] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, D. J. Strouse, Joel Z. Leibo, and Nando de Freitas. 2018. Intrinsic Social Motivation via Causal Influence in Multi-Agent RL. (Oct. 2018). https://arxiv.org/abs/1810.08647

[52] Mario Klingemann. 2018. "Found a hideaway in the wastelands of #BigGAN.". https://mobile.twitter.com/quasimondo/status/1064230996793614338

[53] Tarja Knuuttila. 2005. Models as Epistemic Artefacts: Toward a Non-representationalist Account of Scientific Representation.

[54] Tarja Knuuttila. 2005. Models, Representation, and Mediation. *Philosophy of Science* 72, 5 (2005), 1260–1271. https://doi.org/10.1086/508124

[55] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* 1, 4 (Dec. 1989), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

[56] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (Nov. 1998), 2278–2324. https://doi.org/10.1109/5.726791

[57] Robert Leonard. 2010. *Von Neumann, Morgenstern, and the Creation of Game Theory: From Chess to Social Science, 1900âĂŞ1960.* Cambridge University Press.

[58] Claude Lévi-Strauss. 1966. *The Savage Mind.* The University Of Chicago Press, Chicago.

[59] Omar Lizardo. 2004. The Cognitive Origins of Bourdieu's Habitus. *Journal for the Theory of Social Behaviour* 34, 4 (Dec. 2004), 375–401. https://doi.org/10.1111/j.1468-5914.2004.00255.x

[60] Adrian Mackenzie. 2017. *Machine Learners: Archaeology of a Data Practice.* MIT Press, Cambridge, MA.

[61] Karl Maton. 2008. Habitus. In *Pierre Bourdieu: Key Concepts*, Michael Grenfell (Ed.). Routledge, Stocksfield.

[62] Ben Merriman. 2017. Rewriting the Rules of Art: Pierre Bourdieu's "Manet: A Symbolic Revolution". https://lareviewofbooks.org/article/rewriting-the-rules-of-art-pierre-bourdieus-manet-a-symbolic-revolution/

[63] Mehdi Mirza and Simon Osindero. 2014. Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784* (2014).

[64] Mary Morgan and Margaret Morrison. 1999. *Models as mediators: perspectives on natural and social science.* Cambridge University Press.

[65] Allen Newell. 1980. Physical Symbol Systems*. *Cognitive Science* 4, 2 (April 1980), 135–183. https://doi.org/10.1207/s15516709cog0402_2

[66] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press, New York.

[67] Etienne Ollion and Andrew Abbott. 2016. French Connections: The Reception of French Sociologists in the USA (1970-2012). *European Journal of Sociology / Archives Européennes de Sociologie* 57, 2 (Aug. 2016), 331–372. https://doi.org/10.1017/S0003975616000126

[68] Charles S. Peirce. 1931. *Collected Papers of Charles Sanders Peirce.* Vol. 2.

[69] Jean Piaget. 1971. *Genetic Epistemology.* W. W. Norton & Co., New York.

[70] Kim Plunkett and Virginia Marchman. 1991. U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition* 38, 1 (Jan. 1991), 43–102. https://doi.org/10.1016/0010-0277(91)90022-V

[71] Kim Plunkett and Chris Sinha. 1992. Connectionism and developmental theory. *British Journal of Developmental Psychology* 10, 3 (1992), 209–254. https://doi.org/10.1111/j.2044-835X.1992.tb00575.x

[72] Antonio Poscic. 2018. The Pixels Themselves: An Interview With Mario Klingemann. *The Quietus* (Aug. 2018). https://thequietus.com/articles/25188-mario-klingemann-ai-art-interview

[73] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434 [cs]* (Nov. 2015). http://arxiv.org/abs/1511.06434 arXiv: 1511.06434.

[74] David E. Rumelhart. 1984. The emergence of cognitive phenomena from subsymbolic processes. In *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*. 59–62.

[75] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323, 6088 (Oct. 1986), 533–536. https://doi.org/10.1038/323533a0

[76] David E. Rumelhart and James L. McClelland (Eds.). 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1.* MIT Press, Cambridge, MA, USA. http://dl.acm.org/citation.cfm?id=104279.104284

[77] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., USA, 2234–2242. http://dl.acm.org/citation.cfm?id=3157096.3157346 event-place: Barcelona, Spain.

[78] Jeffrey J. Sallaz and Jane Zavisca. 2007. Bourdieu in American Sociology, 1980–2004. *Annual Review of Sociology* 33, 1 (2007), 21–41. https://doi.org/10.1146/annurev.soc.33.040406.131627

[79] Ferdinand de Saussure. 1959. *Course in General Linguistics*. Philosophical Library, New York.

[80] Terrence J. Sejnowski. 2018. *The Deep Learning Revolution*. MIT Press, Cambridge, MA.

[81] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. *arXiv:1711.08536 [stat]* (Nov. 2017). http://arxiv.org/abs/1711.08536 arXiv: 1711.08536.

[82] Joel Simon. 2018. Ganbreeder: Create wild, weird, and beautiful images. https://www.ganbreeder.app/

[83] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (Sept. 2014). http://arxiv.org/abs/1409.1556 arXiv: 1409.1556.

[84] Peter Slezak. 1989. Scientific Discovery by Computer as Empirical Refutation of the Strong Programme. *Social Studies of Science* 19, 4 (Nov. 1989), 563–600. https://doi.org/10.1177/030631289019004001

[85] Paul Smolensky. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11 (1988), 1–74.

[86] Dave Steinkraus, Patrice Y. Simard, and Ian Buck. 2005. Using GPUs for Machine Learning Algorithms. In *Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR '05)*. IEEE Computer Society, Washington, DC, USA, 1115–1119. https://doi.org/10.1109/ICDAR.2005.251

[87] Lucy A. Suchman. 1987. *Plans and situated actions : the problem of human-machine communication*. Cambridge University Press.

[88] Keith Topping and Stewart Ehly (Eds.). 1998. *Peer-assisted Learning*. Routledge, Mahwah, N.J.

[89] Alan M. Turing. 1948. *Intelligent Machines*. National Physical Laboratory Report. National Physical Laboratory report.

[90] Lev S. Vygotsky. 1978. *Mind in Society: Development of Higher Psychological Processes* (new edition edition ed.). Harvard University Press, Cambridge, Mass.

[91] Loïc Wacquant. 2006. Habitus. In *International Encyclopedia of Economic Sociology*, Jens Beckert and Milan Zafirovski (Eds.). Routledge, New York, 315–319.

[92] Loïc Wacquant. 2014. Homines in Extremis: What Fighting Scholars Teach Us about Habitus. *Body & Society* 20, 2 (June 2014), 3–17. https://doi.org/10.1177/1357034X13501348

[93] Loïc Wacquant. 2014. Putting Habitus in its Place: Rejoinder to the Symposium. *Body & Society* 20, 2 (June 2014), 118–139. https://doi.org/10.1177/1357034X14530845

[94] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv:1703.10593 [cs]* (March 2017). http://arxiv.org/abs/1703.10593 arXiv: 1703.10593.

[95] Detlev Zwick and Nikhilesh Dholakia. 2006. The Epistemic Consumption Object and Postsocial Consumption: Expanding Consumer-Object Theory in Consumer Research. 9, 1 (March 2006), 17–43.