Engaging Science, Technology, and Society 6 (2020), 1-19 DOI:10.17351/ests2020.277

Enchanted Determinism: Power without Responsibility in Artificial Intelligence

ALEXANDER CAMPOLO[1]
UNIVERSITY OF CHICAGO

KATE CRAWFORD[2]
NEW YORK UNIVERSITY, MICROSOFT RESEARCH

Abstract

Deep learning techniques are growing in popularity within the field of artificial intelligence (AI). These approaches identify patterns in large scale datasets, and make classifications and predictions, which have been celebrated as more accurate than those of humans. But for a number of reasons, including nonlinear path from inputs to outputs, there is a dearth of theory that can explain why deep learning techniques work so well at pattern detection and prediction. Claims about "superhuman" accuracy and insight, paired with the inability to fully explain how these results are produced, form a discourse about AI that we call *enchanted determinism*.

To analyze enchanted determinism, we situate it within a broader epistemological diagnosis of modernity: Max Weber's theory of disenchantment. Deep learning occupies an ambiguous position in this framework. On one hand, it represents a complex form of technological calculation and prediction, phenomena Weber associated with disenchantment. On the other hand, both deep learning experts and observers deploy enchanted, magical discourses to describe these systems' uninterpretable mechanisms and counter-intuitive behavior. The combination of predictive accuracy and mysterious or unexplainable properties results in myth-making about deep learning's transcendent, superhuman capacities, especially when it is applied in social settings. We analyze how discourses of magical deep learning produce techno-optimism, drawing on case studies from game-playing, adversarial examples, and attempts to infer sexual orientation from facial images. Enchantment shields the creators of these systems from accountability while its deterministic, calculative power intensifies social processes of classification and control.

Keywords

artificial intelligence; deep learning; enchantment; Max Weber; magic; classification

[1] Alexander Campolo, Email: acampolo@uchicago.edu
[2] Kate Crawford, Email: kate@ainowinstitute.org

1. The Return of Alchemy in Artificial Intelligence

In 1961, the School of Industrial Management at M.I.T. celebrated its centennial with a lecture series titled "Management and the Computer of the Future." At its conclusion, John McCarthy, organizer of the 1956 Dartmouth Conference that launched the field of artificial intelligence, made a memorable declaration. Against a consensus that sought to modestly define the different types of tasks that humans and computers were best suited to, McCarthy boldly argued that the differences between human and machine tasks were *illusory*. There were simply some complicated tasks that would take more time to be formalized and solved by machines (Greenberger 1962, 315). The brothers Hubert and Stuart Dreyfus were so struck by this assertion that they submitted additional remarks to the organizers, where they criticized McCarthy's equivalence between mind and machine. Instead, they used a critical metaphor of magic as a call for humility in AI research. Researchers like McCarthy "should run the same risks as the alchemist trying to synthesize gold from base materials: obscurity until success" (Greenberger 1962, 322).

Hubert Dreyfus later expanded these remarks into a report titled "Alchemy and Artificial Intelligence," in which he argued that the excessive techno-optimism in the early years of AI were driven by simplistic and problematic metaphors about intelligence, where the human brain was understood as analogous to a computer. These metaphors were misleading the field and being used to obscure the conceptual limitations and technical pitfalls they were encountering. Dreyfuss expanded his critical comparison between AI and magic, writing "the long range of alchemy has shown that any research which has had an early success can always be justified and continued by those who prefer adventure to patience" (Dreyfus 1965, 85). In other words, the surprising early *efficacy* of both alchemy and AI research served to cover over larger conceptual problems that prevented them from reaching a more respectable scientific status. The polemical comparison of AI with a premodern, protoscientific, magical practice[1] was designed to attack its scientific legitimacy and puncture the overconfidence of its proponents.

Much in AI has changed since the 1960s, including a shift from symbolic systems to the more recent focus on machine learning techniques. Over the last decade, AI has expanded as a field in academia and industry; now a small number of powerful technology corporations deploy AI systems at an international scale. In spite of these changes, contemporary researchers, including leaders in the field, have once again begun to describe the latest deep learning techniques as magical. In a recent interview, the computer scientist Stuart J. Russell reprises this theme:

> We are just beginning now to get some theoretical understanding of when and why the deep learning hypothesis is correct, but to a large extent, it's still a kind of magic, because

[1] Since Dreyfus's early polemical use of alchemy, research in the history of early modern science has problematized overly sharp distinctions between conceptions of alchemy and the modern sciences. This literature instead emphasizes continuities between alchemy, magic, and chemistry as well as the great interest of historical figures like Isaac Newton and Robert Boyle in alchemical practices (Principe and Newman 2001).

> it really didn't have to happen that way. There seems to be a property of images in the real world, and there is some property of sound and speech signals in the real world, such that when you connect that kind of data to a deep network it will—for some reason—be relatively easy to learn a good predictor. But why this happens is still anyone's guess (Ford 2018, 42).

Invoking magic, Russell suggests that deep neural networks can interpret real world phenomena like images, producing effective predictions without theoretical understanding of why this is so. Russell is not the only figure in deep learning to come to this conclusion. Other experts are more critical, adopting Dreyfus's polemical tone. François Chollet of Google recently characterized ad-hoc approaches to modifying learning algorithms as "folklore and magic spells" (Edwards and Edwards 2018). Ali Rahimi, also of Google, invoked "alchemy" directly to describe the lack of understanding of why certain models work (Hutson 2018). A recent event at Princeton University's Institute for Advanced Study convened some of the field's top researchers, broaching the question directly in its title: "Deep Learning: Alchemy or Science?" (Arora 2019).

This article analyzes the significance of this pattern, a discursive event that connects deep learning and magic in historically specific ways (Foucault 1972, 27). What are the features of contemporary deep learning systems and their social applications that have led them to be characterized this way, and what effects do such statements produce?

Many experts, such as Russell, imply that it is only a matter of time until we gain theoretical understanding of deep learning's predictive efficacy. Perhaps. But the discourse of enchantment operates in the present, shaping both social perceptions of these systems and the practices of their designers. It is not reducible to marketing hype or journalistic license, although both of these may reinforce popular perceptions about magic. In fact, this discourse is significant precisely because discussion of magic moves across a wide range of social positions, from experts in the field, to its critics, and to a wider public that is beginning to be exposed to deep learning's social applications.

We term this ensemble *enchanted determinism*: a discourse that presents deep learning techniques as magical, outside the scope of present scientific knowledge, yet also deterministic, in that deep learning systems can nonetheless detect patterns that give unprecedented access to people's identities, emotions and social character. These systems become deterministic when they are deployed unilaterally in critical social areas, from healthcare to the criminal justice system, creating ever more granular distinctions, relations, and hierarchies that are outside of political or civic processes, with consequences that even their designers may not fully understand or control. New problems arise when the lived effects of social prediction and categorization are unknown to their makers and unaccountable to those who are disadvantaged by them when applied in the world. The application of these systems threatens not only legal due process (Citron and Pasquale 2014) but also more expansive forms of political contestation, and social agency, while simultaneously distancing AI designers and the corporations that employ them from ethical responsibility and legal liability.

2. Enchantment and Disenchantment in Deep Learning

It is often the case that new technologies are presented as magical, and contemporary forms of deep learning are no exception. A number of scholars have shown how those with an interest in marketing and profiting from AI benefit from this association. M.C. Elish and danah boyd use the idea of magic to analyze "the manufacturing of hype and promise," which allows businesses to "produce a rhetoric around these technologies that extends far past the current methodological capabilities" (2018, 58). Similarly, Emmanuel Moss and Friederike Schüür show how mythic metaphors build an understanding of machine learning systems as "superhuman" in ways that implicitly separate them from the human capabilities and practices needed for their implementation (2018, 278). There is no doubt that discourses of magic contribute to the intense contemporary hype around AI in this wider sense.

The discourse of enchanted determinism goes beyond marketing or press hype that covers over technological shortcomings of deep learning and its social applications. Instead it operates when these systems *succeed*, at least according to the narrow engineering criteria selected by their creators, when magical mystery and technical mastery curiously work together.

Max Weber's theory of disenchantment allows us to draw out epistemological and political issues at play in the social application of deep learning systems.[4] Disenchantment—a more literal translation of his German phrase "*Entzauberung*" would be "de-magification"—is an epochal diagnosis of Western modernity, encompassing a widespread decline in mystical or religious forces[5] and their replacement by processes of "rationalization and intellectualization" (Weber 1946, 139). This social process encompasses the rise of modern science, whose concepts and experiments contrast with magical ways of understanding the world.[6] Disenchantment

[4] We are not the only scholars to have recently returned to classical Weberian concepts to analyze contemporary technological developments. Morgan Ames (2014, 2015) has used the Weberian notion of charisma—often associated with magic in his sociology of religion (Riesebrodt 1999)—to analyze the ways that technology operates ideologically to both promise solutions to social problems while simultaneously working to conserve an existing social order.

[5] While there is an extensive Weberian literature dedicated to the broader relationship between science and religion in modernity (Asad 2003, Taylor 2007, Scott 2017), our purpose is different. We are interested in when and why themes of enchantment and disenchantment recur in specific historical situations. In other words, our argument is not that enchantment is a useful analytic because "we have never been modern" (Latour 1993) or even that have never been disenchanted, as some of Weber's critics have recently suggested (Bennett 2001, Josephson-Storm 2017). Instead the concept of enchantment gives allows us to grasp how new technologies challenge models of causality, mastery, and the social itself in historically and culturally specific ways.

[6] Magic, of course, is a topic with a rich history in the social sciences, whose study predates Weber. Anthropologists have been particularly attentive to the use of magic in human societies. Many of the discipline's most influential early practitioners, from E.B. Tylor to Sir James Frazer, conceived of magic in terms similar to those with which Dreyfus characterized alchemy—as essentially mistaken premodern practices that science would replace on a historical path of modernization. Subsequent work in anthropology has taken ethnographic data on magic more seriously in order to understand models of causality (Winkelman 1982).

"means that principally there are no mysterious incalculable forces that come into play, but rather that one can, in principle, master all things by calculation" (Weber 1946, 139). Rationalization allows us to control the world in ways that were previously unimaginable, producing a calculative confidence in public life. Weber's thesis also had affective dimensions; he used disenchantment to express the feelings of alienation and lack of shared meaning that he felt as Germany transformed to an industrial mass democracy in the early twentieth century.

What makes contemporary deep learning systems interesting is their ambivalent position with respect to Weber's larger thesis. They certainly embody aspects of a disenchanted world in that they work to master or control new domains of social life through technical forms of calculation. Furthermore, they tend to emerge from the same scientific domains that are strongly associated with disenchantment. In social settings, deep learning systems promise to identify new efficiencies, produce more accurate classifications, or make more rational decisions, even allowing us to avoid irrational human biases (Kleinberg et al. 2018). At the same time, these systems seem to violate the epistemology of disenchantment, the idea that there are no longer "mysterious" forces acting in the world.[7] Paradoxically, when the disenchanted predictions and classifications of deep learning work as hoped, we see a profusion of optimistic discourse that characterizes these systems as magical, appealing to mysterious forces and superhuman power. This mystification covers over the ways in which deep learning systems can reproduce and intensify discriminatory or harmful processes of prediction and categorization when applied to humans and social institutions (Bowker and Starr 1999). Deep learning thus both intensifies and challenges diagnoses of disenchantment. When applied in critical social domains, it may deepen existing power imbalances between those who create the technologies, and those on whom they act. It is a form of power without knowledge.

3. Deep Learning's Enchanted Epistemology

Understanding some basic principles behind deep learning techniques can further illustrate this ambivalence. Instead of relying on domain specialists to explicitly encode rules, these systems rely on algorithmic models to identify patterns from vast amounts of data. A deep learning textbook authored by Ian Goodfellow, Yoshua Bengio, and Aaron Courville describes the approach in general terms:

> [Deep learning allows] computers to learn from experience and understand the world in terms of a hierarchy of concepts, with each concept defined through its relation to simpler concepts. By gathering knowledge from experience, this approach avoids the need for human operators to formally specify all the knowledge that the computer needs. The

[7] The political theorist Jane Bennett describes the epistemology of disenchantment as the belief that "in principle," or in the last instance, *experts* can explain all phenomena through calculation, even if Weber and Bennett both take pains to underscore that the average person may not be able to provide such explanations (2001, 59). The inability of experts to fully explain why deep learning systems achieve their results sits uneasily with this diagnosis.

> hierarchy of concepts enables the computer to learn complicated concepts by building them out of simpler ones. If we draw a graph showing how these concepts are built on top of each other, the graph is deep, with many layers. For this reason, we call this approach to AI deep learning (Goodfellow, Bengio, and Courville 2016, 1-2).

Deep learning works to identify patterns in data, with lighter human supervision than its expert system predecessors, testing these inferences against examples to tune models, often automatically. It extends machine learning techniques by layering more abstract representations of data on top of each other, from the coarse to the refined, nesting each step within a network. This layering process increases the "depth" of the network. The network's outputs can be tested in real-world contexts, usually in the form of classifications or predictions, such as deciding whether an email should be classified as spam. This "empirical" orientation differentiates deep learning from predecessors that relied on an initial, a priori specification of a task.

To recognize a face, for example, a deep learning model such as a convolutional neural network (LeCun, Bengio, and Hinton 2015, 439) would not begin with a set of instructions that formally define facial features in a digital image. Instead, a researcher would train their network inductively, with few a priori assumptions, often on data large datasets produced in networked environments, such as facial images from a security camera feed, or photos harvested from an online platform like Flickr. The system might first identify localized motifs—in the form of statistical correlations of pixel values—that occur a number of different images (LeCun, Bengio, and Hinton 2015, 438).

Moving toward a more complex, hierarchical representation of this data, the next layers might identify correlations between local motifs and build a composite, working from parts of a face toward a whole. It is common for such systems to pick up on spurious patterns or noise, such as finding correlations between the backgrounds of images, demanding further tuning. Finally, once useful patterns in the training data set have been identified and the network's parameters have been weighted, researchers may test the network's ability to *generalize*, or how well it classifies unfamiliar faces, not included in the training data set. The resulting network's model of "the face" may produce accurate recognition results without any reference to facial features that humans find meaningful, such as eyes, noses, or mouths (Goodfellow, Bengio, and Courville 2016, 4).

The challenges associated with using complex, high-dimensional data in contemporary deep learning are assimilated in an older magical discourse in machine learning, "the *curse* of dimensionality" (Bellman 1957, ix emphasis added). Representations are layered in a hierarchical order, building in complexity, to form "meaningful" models—although the question of meaningful to whom is often left unanswered. Jenna Burrell describes this situation as a fundamental "mismatch between mathematical optimization in high-dimensionality settings…and demands of human-scale reasoning and styles of interpretation" (2016, 2).

Questions of complexity in contemporary deep learning also relate to a broader sociotechnical development: the production of large digital data sets. Writing about the nineteenth century, Ian Hacking describes an "avalanche of printed numbers" produced by emerging administrative and governmental institutions (1990, 2). Today, this avalanche has

intensified to include millions of photographs, texts, and location points being uploaded every minute, collected by companies that amass and analyze those digital traces at scales that were unfeasible only a decade ago. These massive datasets have played a central role in the development of deep learning techniques. Goodfellow, Bengio, and Courville note that many of the algorithms used in deep learning have been known since the 1980s (2016, 19). Part of what accounts for their improved results in the present is the fact that training data is both more plentiful and that it is produced not in artificial laboratory settings but in the everyday lives of millions of people.[8] This abundance of data affords deep learning systems a much wider range of "real world" information on which to train and test their systems but also is used to justify *forgoing* the types of explanations we have come to expect in a disenchanted world. Proponents of these changes argue that we need to focus less on causality and theoretical mechanisms and more on simply identifying useful correlations (Mayer-Schonberger and Cukier 2013, 14; Anderson 2008). When data, defying our disenchanted intuition, is *unreasonably* effective (Halevy, Norvig, and Pereira 2009), theorization or causal explanation can be jettisoned. In these enchanted worlds, predictive efficacy trumps causal explanation.

4. Producing Optimism: The "Black Art"

This gap—between the ability of an observer to interpret the principles behind predictions made by deep learning models and their accuracy or efficacy in certain contexts— has produced many examples of enchanted determinism. AI researchers themselves offer some of the best characterizations of their models as brilliant but mysterious. As early as 2012, before deep learning techniques were widely used, the machine learning researcher Pedro Domingos wrote, "developing successful machine learning applications requires a substantial amount of 'black art' that is difficult to find in textbooks" (2012, 78). More recent articles in the deep learning literature continue to express a tension between performance or efficacy and lack of knowledge. To quote a typical example, "deep neural networks have proved *astoundingly* effective at a wide range of empirical tasks...Despite these successes, *understanding of how and why neural network architectures achieve their empirical successes is still lacking*." (Raghu et al. 2017, 1 emphasis added). Understanding and technological progress is uncoupled; researchers admit that they don't fully understand how or why deep learning works as well as it does. Another paper opens, "Deep neural networks have achieved state-of-the-art performance in a wide range of tasks…Despite their promising results in applications, *our theoretical understanding of neural networks remains limited*" (Lu et al. 2017 1, emphasis added). This is a form of enchantment—empirical accuracy and predictive success defy the intuitions of even the most knowledgeable experts, who admit

[8] The production and standardization of benchmark datasets has taken on a social life of its own. If a researcher is interested in handwritten digits, they can use MNIST, with 60,000 examples for training. Or if researchers are interested in faces, there is the CelebA dataset features more than 200,000 thousand celebrity faces with 40 attribute annotations each. Of course, these seem quite small compared to Google's Open Images dataset, which comprises roughly nine million images and labels for 6000 categories.

that they don't fully understand the theoretical basis for why deep learning works as well as it does.

In the following section, we analyze two aspects of this discursive tension between performance and understanding, enchantment and disenchantment in social contexts. We begin with one of the classic problems in AI: games. Our second example addresses the dangers of enchantment when it becomes deterministic, when socially-oriented systems reproduce social inequalities through new techniques of prediction or classification. Deep learning systems do not simply reflect the world. They also shape it, deepening and naturalizing socially contested classifications and hierarchies and foreclosing contestation or political discussion (Noble 2018; Eubanks 2018). This perilous combination of diminished agency for the classified and expansive social power for system builders lies at the heart of enchanted determinism.

4a. Surprising Deep Learning

In 2015 AlphaGo, a program that uses deep neural networks combined with human training, defeated the Go champion Lee Sedol, an event then seen as an AI landmark. Developed by DeepMind, a Google subsidiary, the program mastered a complex, open-ended game that is significantly more challenging to model than games like chess (Borowiec 2015). Both Elish and boyd (2018, 63) and Moss and Schüür (2018, 277) analyze this event as a cultural spectacle that generates confidence in deep learning systems and awe at their superhuman prowess. In addition, it invokes the epistemological dimension of enchantment, an idea that deep learning produces novel and strange playing strategies that are difficult for humans, even experts to understand. The match against Sedol was notable not just for the result, but also for the unusual moves that AlphaGo made during its gameplay. A reporter from *Wired* wrote that it showed "machines are now capable of moments of genius." He continued, "in Game Two, the Google machine made a move that no human ever would. And it was *beautiful*. As the world looked on the move so perfectly demonstrated the enormously powerful and *rather mysterious* talents of modern artificial intelligence" (Metz 2016, emphasis added).

Speaking through an interpreter, the expert Go player Sedol described the 37th move of game two in no less enchanted terms: "Yesterday, I was surprised. But today I am speechless" (Metz 2016). So, an AI system is described using aesthetic categories: beauty, mystery, surprise, and virtuosic genius. This is enchanted determinism at work: a technological system described in terms of the sublime (Ames 2018), while operating within a predetermined set of rules and outcomes. But something very different is going on if we look at how a system is actually producing these kinds of moves. In the case of AlphaGo, it was trained on a massive set of former games, conducting a pattern analysis of every possible move: something that is neither the work of individual genius nor the product of transcendental intelligence. It is instead doing mathematical optimization at scales far beyond expert human play. These disenchanted means nonetheless are discursively rendered as enchanted, defying expert intuitions about strategy and performance.

In 2017, DeepMind unveiled AlphaZero—the successor to AlphaGo—that extended their approach through the use of a "pure reinforcement learning" that dropped even high-level human instructions, simply playing against itself with the positions on the board as inputs. The researchers from DeepMind characterized AlphaZero's performance as "superhuman," purporting to "master" the game "without human knowledge" (Silver et al. 2017, 354). Later, DeepMind CEO Demis Hassabis compared the system's performance to a chess-playing alien or "chess from another dimension" (Knight 2017). The discourse of exceptional, enchanted, otherworldly and superhuman intelligence shapes our understanding and expectations of deep learning systems. It also has social and political effects, often serving the interests of their powerful creators. Most important among these is that it situates deep learning applications outside of understanding, outside of regulation, outside of responsibility, even as they sit squarely within systems of capital and profit. A massively large and expensive computing infrastructure doing statistical analysis is discursively conferred the status of an enchanted object, closing them off to other forms of critique. In these contexts, the ambivalent discourse of enchanted determinism—systems that are both mystical yet profoundly accurate predictive engines—can create a blindness to forms of risk. This is a continuation of another dynamic of modernization, where technologies hailed for producing progress result in unintended consequences, new forms of unmanageable risk (Beck 1992, 13).

This failure to understand mechanisms underlying classifications has already produced such hazards, both in deep learning and more familiar linear classifiers. So-called "adversarial examples" in machine learning work by making small changes to inputs designed "to maximize the prediction error" of a model (Szegedy et al. 2013, 1). Importantly, adversarial examples do not work in ways that are easy for humans to perceive because they do not rely on the causal models that we find intuitive or familiar (Selbst and Barocas 2018). It is less a matter of creating disguised or ambiguous images to fool human observers than targeting the optimization models at the heart of deep learning algorithms, exploiting their counter-intuitive mathematical properties to make detection difficult (Szegedy et al. 2013, 5).

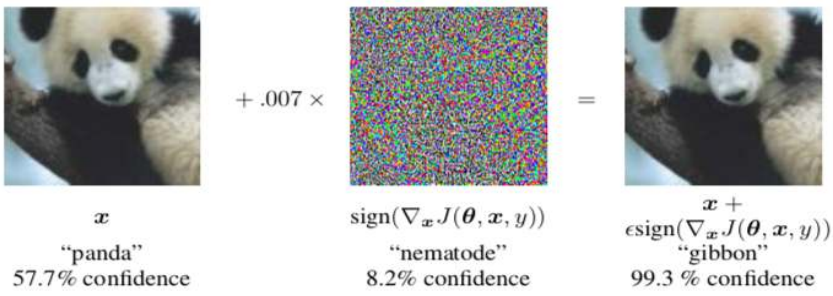

Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

Figure 1. An adversarial example developed by Goodfellow Shlens, and Szegedy (2014, 3).

As the illustration above shows, relatively small changes to inputs, otherwise meaningless to human observers, can lead to huge increases in classification error rates. In this case, the addition of a specially designed, imperceptible vector can cause an image recognition system to misclassify an image of a panda as a gibbon, with high confidence in the incorrect classification.

Adversarial examples reveal a darker side of machine learning's enchanted predictive powers; they cause machines to be misled in ways that are difficult for humans to identify, understand, and control. As Goodfellow, Shlens, and Szegedy argue, "the existence of adversarial examples suggests that being able to explain the training data or even being able to correctly label the test data *does not imply that our models truly understand the tasks we have asked them to perform*. Instead, their linear responses are overly confident at points that do not occur in the data distribution, and these confident predictions are often highly incorrect" (2014, 9 emphasis added). This *divergence* between calculation and understanding raises important questions about the application of deep learning in social domains. To take just one example, an attacker could wreak havoc on roads populated by self-driving cars whose object-recognition systems were manipulated to misrecognize stop signs (Vincent 2017).

4b. Flatland

A further difficulty emerges as enchanted deep learning is applied in the disenchanted world that Weber diagnosed. If we go back to the historical roots of AI in the U.S., particularly in military research in the areas of signal processing and optimization during the Second World War, we see how the social itself was rendered as a type of chaotic, uncontrollable set of forces that had to be managed. A strong emphasis on prediction emerged from applications like fire-control systems (Galison 1994). Their implicit theory of the social came out of the belief that accurate prediction is fundamentally about extracting the right information from chaotic or "noisy" social situations. The aim was to "tame chance" and rationalize the world in the manner theorized by Weber (Hacking 1990).

This epistemological "flattening" of complex social contexts into clean "signal" for the purposes of prediction also has a bearing on the social applications of machine learning. The tensions of enchanted determinism become acute when deep learning techniques promise to extract useful signals without the epistemological modeling or hypothesis formation of a classical probabilistic worldview—when efficacy and explanation are decoupled. Instead, cause and effect questions are bracketed in preference to detecting complex patterns in nonlinear ways from large data sets.

Deep learning systems are at their most deterministic when they are applied to ascribe identity or other social characteristics from a set of inputs understood as signals. That includes predicting sexuality from a photograph of a face (Wang & Kosinski 2018), whether a person will or will not commit a crime after being released on bail (Kleinberg et al. 2018), whether a person is a credit risk (Crosman 2017), or whether a crime was "gang-related" (Seo et al. 2018), to name

only a few applications. These are profoundly relational, contextual, and socially determined identities that are not fixed but change over time and context. To make such predictions, deep learning systems are being trained on highly contingent data, such as people's self-identifications of their sexuality on sites designed for dating, already racially skewed crime data, or gang-databases which are notoriously inaccurate and biased data sources (Sweeney and Fry 2018). When these systems are mystified through discourses of enchantment, they displace the technical realities (and biases) of how the systems were trained, optimized and commercialized, and the political processes of social categorization that underlie them. This works to place them above critical questioning, and paints them as free from subjective human decision-making, which is discursively positioned as arbitrary and biased by comparison (Miller 2018).

For example, in 2018 Yilun Wang and Michal Kosinski of Stanford University published an article titled "Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images." The study followed a familiar pattern for applying deep learning techniques in social settings. First, they used Face++ (a popular commercial face-detection software) to extract faces from images originally posted on an unnamed U.S. dating website that included self-identified data on sexual orientation that could be used for validation. Next, human workers on Amazon Mechanical Turk cleaned the face data, verifying gender and race—the study only looked at Caucasian faces—and a few other parameters (2018, 248). They only considered gender as a binary category: men or women. Next, the researchers extracted facial features from the set of cleaned images using a deep neural network called VGG-Face, which translates each facial image into 4,096 scores. Finally, the authors used a logistic regression model to classify each face's sexual orientation based on the VGG-Face scores. The results of this model were then compared to classifications made by human judges, once again recruited from Amazon Mechanical Turk (2018, 253), leading to the study's headline finding that the deep neural networks could correctly identify sexuality from a facial image 81% of the time for men and 71% of the time for women (2018, 250), an accuracy rate higher than the human judges, who scored 61% for male and 54% for female images (2018, 253).

Despite the seemingly clear-cut percentages, the social implications of these results are not easily interpretable. Even the authors find it difficult to explain the reasons why their model produced these higher accuracy scores, suggesting that it is due to the ability of deep learning to somehow process social signals at superhuman levels: "the findings reported in this work show that our faces contain more information about sexual orientation than can be perceived or interpreted by the human brain" (2018, 254). They speculate that this apparent link between facial morphology and sexual orientation might be explained by a disputed theory that sexual orientation is determined by the presence of prenatal hormones (PHT), even though prior studies have found limited evidence linking facial features and sexual orientation.[9]

[9] The determination of sexual orientation and its physical expression is obviously too complicated an issue to legislate here. What we can say is that Wang and Kosinski do not present this issue at adequate depth, a problem that is endemic in the socially-oriented deep learning literature. For example, they cite a number of papers in support of what they term "widely accepted prenatal hormone theory of sexual orientation" (2018, 247). Here is the conclusion of one of the articles that the authors cite repeatedly (2018, 247, 254): "At the

In discourses of enchanted determinism, claims about rates of accuracy tend to displace causal scientific explanations. The authors present their findings not as direct evidence in support of a PHT theory of sexuality, which would require evidence of biological mechanisms but merely that their work is indirectly "consistent" with such a theory. The *consistency* of uninterpretable correlations replaces a causal epistemology or theoretical explanation of sexuality and is deployed retroactively to justify the methodological choices. Their decision to do the study at all, despite the evident risk to people living in countries where homosexuality is illegal, is justified by the authors in terms of the fact that if it is possible, then it represents a risk and should be public. Wang and Kosinski argue that "as governments and companies seem to be already deploying face-based classifiers aimed at detecting intimate traits, there is an urgent need for making policymakers, the general public, and gay communities aware of the risks that they might be facing already" (2018, 255). Furthermore, the fact that one particular deep learning system is more accurate than a particular group of human beings recruited from Amazon Mechanical Turk tells us very little about human capacities in general and even less about the significant normative questions regarding whether facial images *should* be used to judge sexuality at all. The focus on the "technical success" of a system in a discourse of enchanted determinism works to seal off its epistemological shortcomings and ethical problems.

5. Enchantment and Disenchantment: The Worst of Both Worlds?

In each of these cases we see a type of mystification and myth-making about how deep learning systems work, with two strands that sit in tension: claims of high levels of accuracy and objectivity for systems that are simultaneously beyond human understanding or explanation. This discourse has the effect of placing deep learning systems beyond responsibility and liability ("how could we have known?"). Their accuracy measures ascribe to them a power of social prediction and categorization without assessing the underlying social and political risks, hidden behind counter-intuitive properties or uninterpretable mechanisms. When these techniques are introduced in social domains, they have the potential to intensify hierarchies and differences while closing them off to political debate, visibility, or accountability. The discourse of enchanted determinism values the technical "accuracy" of predicting social identities over a deeper contextual knowledge of how those predictions are made, their limitations, or the impacts those predictions have in the world.

---

present time, therefore, we have no clear evidence of a specific determinant of homosexuality, but indications that a number of factors, varying in importance across individuals, can interact to make same-sex interaction and attraction more likely, followed by the impact of sociocultural "constructionism" on sexual identity formation" (Jannini et al. 2010, 3252). Here is the conclusion of another: "Finally, the relationship between genes and homosexual behaviour remains unexplored in this paper. The problem is that researchers cannot consistently demonstrate genetic linkage between markers for homosexuality in males, and no genetic traces have shown up for behavioral genetic methods with females" (Udry and Chantala 2006).

Dreyfus's critical analogy between alchemy and AI pointed out the dangers of simplistic analogies between human cognition and artificial intelligence. Today we see a different danger in deep learning, as these systems are portrayed as *exceeding* human intelligence or performance, as forms of "genius" or even alien cognition. In fact, their insights depend on their ability to extract patterns from vast amounts of data harvested from the patterns of everyday life, at a granular scale made possible by datasets drawn from smart phones, sensor networks, and social media. We are not being confronted with a sublime form of superhuman intelligence, but a form of complex statistical modeling and prediction that has extraordinarily detailed information about patterns of life but lacks the social and historical context that would inform such predictions responsibly—an irrational rationalization.

Of course, deep learning does not have a monopoly on counter-intuitive or unexplainable properties. A surprising amount of our knowledge, scientific and otherwise, is "tacit" or difficult to explain in explicit terms (Collins 2010). Deep learning researchers also point to the relative novelty of the field to assert that we will soon better understand why it appears to perform so, although at the moment such assumptions are difficult to distinguish from McCarthy's original and still unrealized optimism. The most productive response to this epistemological puzzle is a new interdisciplinary literature on fairness, accountability, and transparency (FAT*) in algorithmic systems, whose very existence nonetheless underscores the problems of enchantment. The laudable effort to better understand and control these systems is *premised* on their complexity and counter-intuitive properties, the need to work around inherent "trade-offs between accuracy, transparency, and interpretability" (Zeng, Ustun, and Rudin 2017, 690).

Researchers in the area of interpretability in machine learning attempt to produce the types of rational explanations for classifications and predictions that enchanted discourses withhold. They have devised ingenious approaches, even for techniques like deep neural networks that seem to defy statistical intuitions and explanations. Some, such as saliency masks (Guidotti et al. 2019) or feature visualization (Olah, Mordvintsev, and Schubert 2017) identify the features or parts of an input that most informs its eventual classification. Implicit in this work is the idea that if we are able to identify these most salient aspects, we can (re)construct theoretical or causal explanations for why individual predictions were made. At the moment, however, such explanatory methods are novel, and there are no guarantees that identifying the features that are most mathematically informative to machines will produce explanations that are informative to humans, as the case of adversarial examples suggests. Furthermore, these and other explanatory techniques tend to be backwards-looking, providing post-hoc rationales for predictions, potentially only after harms have been recognized (Lipton 2018, 21-22). Deep learning once again takes on an enchanted quality when we are forced to cast about for explanations—or myths—for how accurate predictions might have been made.

And there are deeper issues with interpretability that suggest that enchanted determinism might not give way so easily to disenchanted scientific rationality. A major issue is that ideas like interpretability and explanation entail ethical and political choices—"interpretability for whom?" Experts in the field have identified a set of basic, unresolved issues. The computer scientist Zachary C. Lipton observes,

> The academic literature [on interpretability] has provided diverse and sometimes non-overlapping motivations for interpretability and has offered myriad techniques for rendering interpretable models. Despite this ambiguity, many authors proclaim their models to be interpretable axiomatically, absent further argument. Problematically, it is not clear what common properties unite these techniques" (2018, 1).

Similarly, in their review of the field, Guidotti et al. argue that one of the limitations of explaining decisions made by machine learning is that there has been "no agreement on what an *explanation* is"—a problem that they begin to address by classifying different approaches (2019, 36). This issue, that interpretability and explanation might mean different things to different people, becomes more acute when deep learning systems are applied in contested social situations where tradeoffs and other ethical and political conflicts are unavoidable. People tend to change their behavior when classifications have high stakes, and their ability to do so will depend on both their knowledge of these systems and their own social positions (Hu, Immorlica, and Vaughan 2019; Espeland and Sauder 2007).

As researchers work to formulate the field's definitions and guiding problems, they have come to realize that "interpretability" it is not a simple panacea that will inevitably displace discourses like enchanted determinism. Indeed, the most promising research in this area forces us to confront the fact that interpretability is not simply a property of any model or technique but rather only emerges with deep contextual knowledge of the social structures and even histories where they are applied. These dilemmas point to a final ambivalence of enchanted determinism—that it embodies the dangers of both disenchanted and enchanted worlds. Weber used the image of the iron cage to suggest that, once set in motion, processes of rationalization and calculation are very difficult to resist and control. They tend to intensify without regard to our intentions. (Weber 1992, 123). If deep learning techniques are deemed effective or accurate enough to use without understanding how they work or whether they deepen social stratification and hierarchies, this will help to inoculate them from political contestation and debate and make their predictions more difficult to dispute. The enchanted nature of this determinism exacerbates this situation by discursively rendering deep learning systems transcendent, inaccessible to human interpretation, governed by tradeoffs at their core. The efficiency of disenchanted rationalization and the magical transcendence of enchantment work together to defuse critique, celebrating predictive accuracy while erasing the social and political work that goes into a system's design or its ongoing effects. The work of contesting hierarchies and injustices begins with *understanding* and demonstrating how social discrimination affects these different groups. If these differences are hidden through uninterpretable mechanisms of classification and subsequently justified by their accuracy or better-than-human classification performance, it is difficult to know where or how to make demands for justice.

One critical possibility involves contesting the unspoken normalization of dominant *categories* at the heart of these technologies of classification. Simone Browne has argued that a "prototypical whiteness (as well as proto-typical maleness, youth, and able-bodiedness)" inscribes racial categories into surveillance technologies like facial recognition (2015, 110). In

other words, "there is a certain assumption with these technologies that categories of gender identity and race are clear cut, that a machine can be programmed [or in the case of deep learning learn] to assign gender categories or determine what bodies and body parts should signify" (Browne 2015, 114). Such prototypical whiteness pervades machine learning; recall that in Wang and Kosinski's study, not only were there only two genders and two sexual orientations recognized, only facial images of Caucasian users were considered. Amazon Mechanical Turk workers were hired to remove all other races from the dataset—data cleaning as racial exclusion (2018, 248). The question of generalization becomes a political one when training datasets are assumed to be representative of our social worlds, when AI's proponents use accuracy as justification for placing limits on human discretion, judgment, and agency. Technical questions of accuracy and performance are shot through with political choices about categories and norms (Benthall and Haynes 2019). But they are rarely acknowledged as such.

The discourse of enchanted determinism has the effect of covering over these problems, suggesting that if we want "superhuman" accuracy and performance from deep learning systems, we may need to give up the types of rational, causal explanations that Weber associated with disenchanted modernity. We are left with ever-intensifying social processes of classification and prediction that are resistant to inherited political strategies for combatting discrimination and inequality. The work of classifying and predicting identities, credit risks and many other social characteristics is not done in a social vacuum, neither does it reflect a set of underlying social signals that are perceptible only by the "genius" of deep learning systems. It is a process that is actively shaped by system designers and the data used to reflect the world. When we see discourses of enchanted determinism at work, we should ask whose interests they serve and where the responsibility for the impacts of that system will ultimately rest.

Author Biography

Alexander Campolo was a research assistant at the AI Now Institute where he worked on a range of critical and epistemological topics in AI. He is currently a postdoctoral researcher at the Stevanovich Institute on the Formation of Knowledge at the University of Chicago, where he continues to work at the intersection of data, knowledge, and politics.

Author Biography

Kate Crawford is a Distinguished Research Professor at New York University, where she co-founded the AI Now Institute. Her work over the last decade has centered on the social and political implications of large-scale data and machine learning. She is also a Principal Researcher at Microsoft Research, and the inaugural Visiting Chair in AI and Justice at the École Normale Supérieure in Paris.

Acknowledgements

We would like to thank Meredith Whittaker, Trevor Paglen, and the research community at the AI Now Institute as well as Laura Forlano, Caroline Kao, and Nick Proferes of the "Techno-Optimism Within and Beyond Silicon Valley" conference, organized by Damien Droney, Morgan Ames, & Mark Gardiner, for their invaluable insights and feedback on this essay. The article was further improved by comments from ESTS editors and an anonymous reviewer.

References

Ames, Morgan G. 2014. "Translating Magic: The Charisma of One Laptop per Child's XO Laptop in Paraguay." In *Beyond Imported Magic: Essays on Science, Technology, and Society in Latin America*, edited by Eden Medina, Ivan da Costa Marques and Christina Holmes, 207–224. Cambridge, MA: MIT Press.

_____. 2015. "Charismatic Technology." In *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives*, 109–120. AA '15. Aarhus, Denmark: Aarhus University Press. https://doi.org/10.7146/aahcc.v1i1.21199.

_____. 2018. "Deconstructing the Algorithmic Sublime." *Big Data & Society* 5 (1): 1–4. https://doi.org/10.1177/2053951718779194.

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired*. June 23. https://www.wired.com/2008/06/pb-theory/.

Arora, Sanjeev. 2019. "Brief Introduction to Deep Learning and the 'Alchemy' Controversy." presented at *Deep Learning: Alchemy or Science?*, Institute of Advanced Study, Princeton University, February 22. https://www.youtube.com/watch?v=kqhg-o-KEns.

Asad, Talal. 2003. *Formations of the Secular: Christianity, Islam, Modernity*. Stanford: Stanford University Press.

Beck, Ulrich. 1992. *Risk Society: Towards a New Modernity*, translated by Mark Ritter. London: Sage.

Bellman, Richard. 1957. *Dynamic Programming*. Princeton: Princeton University Press.

Bennett, Jane. 2001. *The Enchantment of Modern Life: Attachments, Crossings, and Ethics*. Princeton: Princeton University Press.

Benthall, Sebastian, and Bruce D. Haynes. 2019. "Racial Categories in Machine Learning." In *FAT* '19: Conference on Fairness, Accountability, and Transparency*, 10. Atlanta, GA: ACM. https://doi.org/10.1145/3287560.3287575.

Borowiec, Steven. 2016. "AlphaGo Seals 4-1 Victory over Go Grandmaster Lee Sedol." *The Guardian*, March 15. https://www.theguardian.com/technology/2016/mar/15/googles-alphago-seals-4-1-victory-over-grandmaster-lee-sedol.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

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

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 1–12. https://doi.org/10.1177/2053951715622512.

Citron, Danielle Keats, and Frank A. Pasquale. 2014. "The Scored Society: Due Process for Automated Predictions." *Washington Law Review* 89 (1): 1–33.

Collins, Harry. 2010. *Tacit and Explicit Knowledge*. Chicago: University of Chicago Press.

Crosman, Penny. 2017. "Is AI Making Credit Scores Better, or More Confusing?" *American Banker*. February 14. https://www.americanbanker.com/news/is-ai-making-credit-scores-better-or-more-confusing.

Domingos, Pedro. 2012. "A Few Useful Things to Know About Machine Learning." *Communications of the ACM* 55 (10): 78–87. https://doi.org/10.1145/2347736.2347755.

Dreyfus, Hubert L. 1965. "Alchemy and Artificial Intelligence." Santa Monica: RAND. http://www.rand.org/pubs/papers/P3244.html.

Edwards, Helen, and Dave Edwards. 2018. "Google's Engineers Say That 'Magic Spells' Are Ruining AI Research." *Quartz*. October 5. https://qz.com/1274131/googles-engineers-say-that-lack-of-rigor-is-ruining-ai-research/.

Espeland, Wendy Nelson, and Michael Sauder. 2007. "Rankings and Reactivity: How Public Measures Recreate Social Worlds." *American Journal of Sociology* 113 (1): 1–40.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Ford, Martin. 2018. *Architects of Intelligence: The Truth about AI from the People Building It*. Birmingham, UK: Packt Publishing.

Foucault, Michel. 1972. *The Archaeology of Knowledge*. Translated by A.M. Sheridan Smith. New York: Pantheon.

Galison, Peter. 1994. "The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision." *Critical Inquiry* 21 (1): 228–66.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2014. "Explaining and Harnessing Adversarial Examples." ArXiv:1412.6572 [Cs, Stat], December. http://arxiv.org/abs/1412.6572.

_____, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press.

Greenberger, Martin. 1962. *Management and the Computer of the Future*. New York: Wiley.

Guidotti, R., Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51 (5): 1–42. https://doi.org/10.1145/3236009.

Hacking, Ian. 1990. *The Taming of Chance* Cambridge: Cambridge University Press, 1990.

Hu, Lily, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. "The Disparate Effects of Strategic Manipulation." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 259–268. FAT* '19. New York, NY, USA: ACM. https://doi.org/10.1145/3287560.3287597.

Hutson, Matthew. 2018. "AI Researchers Allege That Machine Learning Is Alchemy." *Science*. May 3. http://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy.

Jannini, Emmanuele A., Ray Blanchard, Andrea Camperio-Ciani, and John Bancroft. 2010. "Male Homosexuality: Nature or Culture?" *The Journal of Sexual Medicine* 7 (10): 3245–53.

Josephson-Storm, Jason A. 2017. *The Myth of Disenchantment: Magic, Modernity, and the Birth of the Human Sciences*. Chicago: University of Chicago Press.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133 (1): 237–93. https://doi.org/10.1093/qje/qjx032.

Knight, Will. 2017. "Alpha Zero's 'Alien' Chess Shows the Power, and the Peculiarity, of AI." *MIT Technology Review*. December 8. https://www.technologyreview.com/s/609736/alpha-zeros-alien-chess-shows-the-power-and-the-peculiarity-of-ai/.

Latour, Bruno. 1993. *We Have Never Been Modern*, translated by Catherine Porter. Cambridge, MA: Harvard University Press.

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

Lipton, Zachary. C. 2018. "The Mythos of Model Interpretability." *Queue* 16 (3): 30:31–30:57. https://doi.org/10.1145/3236386.3241340.

Lu, Zhou, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. 2017. "The Expressive Power of Neural Networks: A View from the Width." ArXiv:1709.02540 [Cs], September. http://arxiv.org/abs/1709.02540.

Metz, Cade. 2016. "In Two Moves, AlphaGo and Lee Sedol Redefined the Future." *Wired*. March 16. https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/.

Miller, Alex P. 2018. "Want Less-Biased Decisions? Use Algorithms." *Harvard Business Review*. July 26. https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms.

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

Principe, Lawrence M., and William R. Newman. 2001. "Some Problems with the Historiography of Alchemy." In *Secrets of Nature: Astrology and Alchemy in Early Modern Europe*, 385–431. Cambridge, MA: MIT Press.

Raghu, Maithra, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. 2017. "On the Expressive Power of Deep Neural Networks." In *International Conference on Machine Learning*, 2847–54. http://proceedings.mlr.press/v70/raghu17a.html.

Riesebrodt, Martin. 1999. "Charisma in Max Weber's Sociology of Religion." *Religion* 29 (1): 1–14. https://doi.org/10.1006/reli.1999.0175.

Scott, Joan. 2017. *Sex and Secularism*. Princeton: Princeton University Press.

Selbst, Andrew, and Solon Barocas. 2018. "The Intuitive Appeal of Explainable Machines." *Fordham Law Review* 87 (3): 1085.

Silver, David. Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, et al. 2017. "Mastering the Game of Go without Human Knowledge." *Nature* 550 (7676): 354. https://doi.org/10.1038/nature24270.

Sweeney Annie, and Paige Fry. 2018. "Nearly 33,000 Juveniles Arrested over Last Two Decades Labeled as Gang Members by Chicago Police." *Chicago Tribune*, August 9. https://www.chicagotribune.com/news/local/breaking/ct-met-chicago-police-gang-database-juveniles-20180725-story.html.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. "Intriguing Properties of Neural Networks." ArXiv:1312.6199 [Cs], December. https://arxiv.org/abs/1312.6199.

Taylor, Charles. 2007. *A Secular Age*. Cambridge, MA: Belknap Press.

Udry, J. Richard, and Kim Chantala. 2006. "Masculinity-Femininity Predicts Sexual Orientation in Men but Not in Women." *Journal of Biosocial Science* 38 (6): 797–809. https://doi.org/10.1017/S002193200500101X.

Vincent, James. 2017. "Magic AI: These Are the Optical Illusions That Trick, Fool, and Flummox Computers." *The Verge*. April 12. https://www.theverge.com/2017/4/12/15271874/ai-adversarial-images-fooling-attacks-artificial-intelligence.

Weber, Max. 1946. "Science as a Vocation." In *From Max Weber: Essays in Sociology*, translated by H.H. Gerth and C.W. Mills, 129–56. New York: Oxford University Press.

_____. 1992. *The Protestant Ethic and the Spirit of Capitalism*, translated by Talcott Parsons. London: Routledge.

Winkelman, Michael. 1982. "Magic: A Theoretical Reassessment." *Current Anthropology* 23 (1): 37–66.

Zeng, Jiaming, Berk Ustun, and Cynthia. Rudin. 2017. "Interpretable Classification Models for Recidivism Prediction." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180 (3): 689–722. https://doi.org/10.1111/rssa.12227.