P&D
Philosophy & Digitality
An Open Access, Peer-Reviewed Journal

Labels and Emotions

Machine Learning's Scientific Objects

Alexander Campolo*

Abstract: One strategy for challenging the use of automated emotion recognition systems has been to contest the validity of the psychological research programs on which they are based, by critically examining their history. This article takes a related but distinctive approach, examining instead how machine learning and emotion recognition became connected in the recent past. It argues that machine learning transforms the theoretical categories that it operationalizes, the judgments that it purports to automate, and show more broadly how technical and scientific objects can interact. A shared emphasis on facial images and emotion labels made the psychology of emotional expression highly compatible with machine learning techniques. Emotional expressions, in turn, were interpreted within a machine learning worldview in which visible expressions can point to deeper underlying signals. Future normative work on emotion computing might draw on such accounts of how machine learning transforms its scientific objects.

Keywords: emotion; affect; machine learning; automation; labelling

1. A different genealogical perspective

1 A number of important positions have now crystallized in the normative study of machine learning. One focuses on the legitimacy of its purposes or ends. Is it right that machine learning should be used to infer sensitive aspects of our behaviors or identities? Who—what groups—are subject to discrimination (Eubanks 2018)? How should fairness and bias be assessed? Another strand focuses in a more immanent way on machine learning techniques themselves. This can range from epistemological critiques of its opacity or uninterpretability (Pasquale 2016) to the ways that machine learning seems to reproduce or "parrot" its potentially biased training data (Bender et al. 2021) or reflects the particular "truths" and norms of those who build models (Jaton 2017). Other critical scholarship attends to instances of colonialist exploitation of workers in the global south who label data (Couldry and Mejias 2019), or even qualitatively new logics of capitalist accumulation (Zuboff 2019), which more recently have led to ecological criticisms of the huge amounts of energy required to train and run contemporary generative models (Crawford 2024).

2 Affect and emotion recognition[1] form a charged nexus where many of these normative issues converge. Intuitively, there seems to be something illegitimate about the way machine learning objectifies us when it is used to detect our emotional states, which we may wish to conceal or not even fully understand ourselves. Is it exploitative to use machine learning to profit from our feelings? Who may use or claim ownership to these intimate aspects of our identities? If affect and emotion recognition can be automated, they will be applied to more and more people. Who will be subjected to the detection of emotions, and for what purposes (Stark 2018)? How can we trust classifications produced by machine learning systems when we do not know exactly how they have been made?

3 A basic question underlying many of these normative issues is that of scientific validity: a necessary condition to properly judge how emotion recognition systems should be used is confidence that they are built on sound theoretical and methodological assumptions, supported by scientific evidence. Such questions, however, lead to

1. I use these words more or less interchangeably throughout, trying to follow the usage of the historical actors involved, although I recognize that some are invested in theoretical distinctions between the two. Barrett & Bliss-Moreau (2009) provide a useful, historically-informed overview.

Contact: Alexander Campolo <alexander.campolo@durham.ac.uk>
https://doi.org/10.18716/ojs/pd.v2i2.11947

thorny problems regarding the nature of emotion itself, putting concepts such as reason and intentionality on the table. They force us to examine the deeper historical and philosophical underpinnings of emotion science.

4 To this end, a set of critical studies have shed light on the origins of an influential theory of emotional expression, Basic Emotion Theory, which posits that there is a small number of primary human emotions, associated with a set of recognizable facial expressions (Stark 2017; Crawford 2021). These studies ask where Basic Emotion Theory has come from (some dark places, it turns out) to problematize its fragile, contingent coherence in the present (Koopman 2013, page 4). The most comprehensive of these works, Ruth Leys' *The Ascent of Affect*, traces the genealogy of emotional expression to warn against the importation of "dubious" scientific theories into the humanities—as well as other contexts like machine learning, one presumes (Leys 2017, page 2).

5 The overall effect of these inquiries is to problematize emotion computing on scientific grounds. A complementary perspective, one that I elaborate below, is to focus on the computing side of the equation (Catanzariti 2023). Rather than granting primacy to scientific theories and relegating technologies to secondary status as mere applications of these theories, this perspective illuminates how machine learning techniques mediate between scientific concepts (here, in the form of basic emotion categories or labels) and experiences (here, data relating to emotional expressions).

6 Of course, it has been common in the history of science to attend to the material instruments and technical contexts in which scientific concepts emerge and stabilize. Working against a Kuhnian, "theory first," philosophy of science (Rheinberger 1997, page 26) the historian Hans-Jörg Rheinberger draws a distinction between open-ended "epistemic things" and the "technical" objects that contain and stabilize them (Rheinberger 1997, page 28–31). In many ways, the use of machine learning in emotion recognition is well-described by this dichotomy; these systems fulfill the "answering machine" function characteristic of Rheinberger's technical objects (Rheinberger 1997, page 32): given an input of a face, what emotion does it express? By repeatedly answering this question, these systems doubtless stabilize and even reify certain categories for emotional experience.

7 However, instead of telling another story of how open-ended scientific concepts are stabilized by technical instruments and practices, I would like to explore a different type of question, which is, how do the "answers" given by technical objects like machine learning *refract* these scientific objects—epistemic things. My argument is that in addition to stabilizing or constraining epistemic things, technical objects disclose different aspects of both these phenomena and themselves. Machine learning, in other words, remakes our scientific understanding of emotional expression in its own image, as a set of signals to be identified beneath the confusing noise of our bodily dispositions. This question is genealogical in its own way. Instead of asking about the contingent origins of a science that is then applied in technology, it is concerned with the contingencies that have made a scientific theory and a technology capable of working together, as well as the effects on both sides of their union.

8 To address this question, I trace connections between the psychologist Paul Ekman, a central figure in much of the existing critical literature, and the machine learning community, focusing on a moment of crystallization in the early 1990s. I argue that Ekman's Basic Emotion Theory and Facial Action Coding Scheme had a number of features that enabled a special compatibility with machine learning technologies. First, the two communities, albeit in their own ways, took photographs of the human face as one of the basic objects of their research. Second, Ekman and his colleagues developed a set of class labels that form the pre-requisite for creating a supervised machine learning model. Indeed, the experimental design for many of Ekman's studies involved asking subjects to attribute an emotion label to a photograph of a human face, a design that could be translated directly into an objective for machine learning systems.

9 The use of machine learning for classifying emotional expressions was often justified in terms of automation; it would make the once-time consuming process of judging and labeling faces, of moving from questions to answers, inputs to outputs, much more efficient. However, I argue that this apparently quantitative relationship also involved a qualitative revaluation of the relationship between the science of emotion expression and machine learning. It prompted technical practitioners to reflect on the nature of emotional expression and experience, to conceptualize emotions as a paradigmatic instance of a more general (historical) epistemology that regards phenomena in terms of signals that may be extracted from noisy environments. Machine learning, much more than human judgment, is capable of extracting significant signals, the true nature of our emotional

experience that may be otherwise scrambled by cultural and communicative conventions. This leads to different normative questions. Not so much, is emotion recognition scientifically valid? But rather, how does the historical relationship between emotion science and machine learning change how we think about the truth of our and others' feelings, and how might we use this predictive knowledge to govern behaviors?

2. Faces and Labels

10 To understand how the science of emotion recognition became connected to machine learning, it is necessary to briefly sketch the Basic Emotion Theory associated most notably with the psychologist Paul Ekman. This theory is the object of ongoing controversy in psychology, where competing theoretical paradigms contend, such as "appraisal models" of emotion in which emotional stimuli are cognitively interpreted and socially and culturally mediated (Barrett 2006). Rather than intervening in these debates or reconstructing the genealogical contexts from which these theories emerged, my purpose is to analytically identify features of Basic Emotion Theory that later made it attractive to machine learning researchers.

11 The development of Basic Emotion Theory has now been critically documented by historians of science, above all Leys. Paul Ekman built on the work of the psychologist Sylvain Tomkins who had developed a distinctive theory of affect, repurposing nineteenth-century sources such as Charles Darwin's *The Expression of Emotions in Man and Animals* (1872) (Darwin 1998) and the work of French neurophysiologist Emile Duchenne du Boulogne who photographed subjects interred at Salpêtrière asylum in Paris (1862) (see also: Delaporte 2008). In his book *Affect, Imagery, Consciousness* (2008), Tomkins proposed a theory of affects that emphasized their independent, inherited, and discrete character: "hardwired, reflex-like, subcortical, and hence noncognitive…physiological reactions" or responses to stimuli in Leys' presentation (Leys 2017, page 32). Unlike more intentional, cognitive, meaning-focused, appraisal theories of emotion, Tomkins theorized affects as shared, cross-cultural, evolutionary adaptations.

12 In Tomkins' theory, the reason that patterns of affective responses can be grouped into a small number of discrete categories is because they are produced by what he terms, using a computational metaphor, "innate affect programs," each comprised of a "subcortical structure which can instruct and control a variety of muscles and glands to respond with unique patterns of rate and duration of activity characteristic of a given affect" (Tomkins 2008, page 135). As Leys stresses, these affect programs have no necessary connections to any single object or set of stimuli; rather, they are activated in a contingent way by many different possible "triggers," replacing a more intentional account of emotional experience with a more contingent, even error prone pattern of stimuli and responses (Leys 2017, page 31–32).

13 Paul Ekman energetically pursued an experimental research program based on Tomkins' theory. Critically for its later connections with machine learning, this experimental program was based on research designs that asked subjects to label photographs of faces expressing (often simulated) emotional expressions, choosing from a short list of basic emotions. Why the face? This is a complicated historical question, but the most proximate source is Tomkins, who, in typically counter-intuitive style took aim at the usual understanding of emotions as essentially *interior* phenomena—that "inner bodily responses are the chief site of the emotions" (Tomkins and McCarter 1964, page 120). He speculated instead that the face, rather than internal organs has become, through evolutionary mechanisms, the maximally sensitive point of affective expression. This reversal is illustrated through an analogy: "just as the fingers respond more rapidly, with both precision and complexity than the grosser and slower moving arm to which they are attached, so the face expresses affect, both to others and to the self, via feedback, which is more rapid and more complex than any stimulation of which the slower moving [internal] visceral organs are capable" (Tomkins and McCarter 1964, page 120).

14 Ekman picked up on this theme, affirming: "We agree with Tomkins and with Darwin"—always a useful authority to invoke—"that there are distinctive movements of the facial muscles for each of a number of primary affect states, and these are universal to mankind," although he also cautioned that more intentional, culturally specific "display rules" also govern these displays (Ekman et al. 1969, page 71–73). Much of the critical commentary on Basic Emotion Theory has focused on these universalistic claims or whether emotions are "natural kinds" (Barrett 2006; Leys 2017; Crawford 2021). Here, rather than exploring these ultimately metaphysi-

cal questions, I would like to focus on something apparently more practical, even mundane: the experimental designs to test this theory, which relied almost exclusively on studying statistical patterns in the *attribution of emotion labels to photographs of faces* by research subjects. I will describe how these connections from the empirical—the face—to the conceptual—the emotion label—are made, first in the context of a psychological research program and later using machine learning techniques.

15 An exemplary study, "Pan-Cultural Elements in Facial Displays of Emotion," began with the construction of what might anachronistically be called a reference dataset. From an initial set of 3,000 posed photos of affective facial displays, the authors selected thirty in each of a set of six "primary affect categories:" "happiness, surprise, fear, anger, disgust-contempt, and sadness" (Ekman et al. 1969, page 87). The criteria for selection were explained at a high level of generality. The authors looked for photographs "which showed only the pure display of a single affect," and mention an in-progress scoring technique that could presumably guide such selections in an objective manner—a project they would further develop. Drawing on their previous work on learned, culturally specific display rules, the authors attempted to select photos free of such influences. The selected photographs "showed male and female Caucasians, adults and children, professional and amateur actors, and mental patients" (Ekman et al. 1969, page 87).

16 These photographs were cropped and presented to subjects as slides. For each slide, the subjects chose from a list of the six affect labels, translated into subjects' native language: "Japanese, Portuguese [for Brazilian subjects], Neo-Melanesian Pidgin, Fore, and Bidayh"—the latter languages for a group of subjects from Papua New Guinea, chosen on the assumption that an "isolated Neolithic material culture would not have culturally learned emotional displays from exposure to Western media" (Ekman et al. 1969, page 87). The authors report generally high levels of accuracy and agreement between the intended and chosen affect label in the three "literate" national groups of respondents from the United States, Brazil, and Japan, and lower but still encouraging rates of accuracy among the subjects from Papua New Guinea, a difference that the authors attribute to "the enormous obstacles imposed by language barriers and task unfamiliarity in preliterate cultures" (Ekman et al. 1969, page 87). In sum, the ability of subjects to correctly match a pre-selected label to an image expressing a discrete basic emotion was interpreted as evidence that facial displays of emotion correspond to a small (here six) set of underlying, cross-cultural emotion categories.

3. Learning to label

17 These conclusions have subsequently been criticized on a variety of grounds, including the artificiality of posed expressions and the validity of using forced-choice designs, where pre-populated lists of emotion labels greatly constrain possible attributions (Russell 1994). Rather than evaluating how these critiques should affect our evaluations of emotion science and affective computing systems based on them, I will look more concretely at how this set of labels served as a prerequisite for machine learning systems. In addition to direct personal links, I argue that shared historical interests in face as a scientific object, a design methodology in which labelling plays a critical role, and a deeper epistemology of interior or latent forces and visible, exterior effects account for why Ekman's Basic Emotion Theory became a paradigmatic machine learning application.

18 This process began with a refinement of the labelling process, which decomposed broader emotion labels for facial expressions into a more granular set of descriptions of muscular movements. A first version, the "Facial Action Scoring Technique," was developed in collaboration with Tomkins (Ekman et al. 1971). The face was divided into three regions, and characteristic gross facial movements were associated with six basic emotion categories: happiness, surprise, anger, sadness, disgust, and fear. However, both the choice of movements and their association with emotion categories was largely subjective, drawn from the researchers "own combined observations and intuitions. We considered for each emotion each of the facial areas and the possibilities for muscular movements within each facial area, checking our hypotheses with a mirror and by looking at each other" (Ekman et al. 1971, page 41).

19 By the end of the 1970s, Ekman and Friesen had developed a different system, based not on "judgments"—the intuitive, subjective attribution of some emotion label to a face or facial region—but rather on a presumably more objective, anatomically-grounded measurement of facial movements, which could later be correlated

with emotional attributions (Ekman 1982, page 46–47). By the late 1970s, Ekman and Friesen developed a comprehensive descriptive system for classifying facial movements: the Facial Action Coding System (FACS). This system associates a set of 44 elementary observable movements—Action Units—with muscular groups, a more minimal inference than ascribing an emotional state (Rosenberg 2005, page 13). Later, standardized dictionaries (EMFACS) and databases (FACSAID) were produced to match facial expressions coded using FACS to emotional states (Rosenberg 2005, page 16).

20 A major problem with this system was that it was prohibitively slow and labor-intensive, with trained analysts required to annotate individual frames of video-recorded facial movements. As the granularity of descriptions increased, so did the cost of producing them. In a 1992 report to the US National Science Foundation on a meeting between emotion and computer vision researchers, Ekman estimated optimistically that one minute of facial expressions required an hour of measurement, and coders needed approximately 100 hours of training to reach acceptable reliability levels. However, by this time, a new technical solution to this problem had appeared: "Recent developments in computer vision and neural networks, used primarily in lip reading, and the recognition of specific individuals from static faces, indicate the possibility of being able to automate facial measurement" (Ekman et al. 1992, page 5). This report marks a moment of crystallization between parts of the emotion recognition and machine learning communities.

21 A first indication of this meeting's significance was its participants. In addition to Ekman, key members of the computer science community contributed: Thomas S. Huang from computer vision, Alex Pentland (doctoral advisor of Rosalind Picard, widely credited with inventing the interdisciplinary field of affective computing), and Terrence J. Sejnowski, an important figure in the neural network renaissance of the 1980s whose work, especially with collaborator Geoffrey Hinton, continues to reverberate in deep learning (Ackley et al. 1985). In parallel to Ekman's now decades-old research program, the face also constituted a primordial scientific object for the computer vision community, dating to the 1960s and 1970s in the work of scholars like Woody Bledsoe (1965) and Takeo Kanade (1977). More recently, Pentland had developed an "eigenface" technique on large databases of mugshots (Turk and Pentland 1991), tracing direct lines to criminological techniques for standardizing images to improve recognition (Ellenbogen 2012).

22 Technical challenges remained formidable, especially for tracking what Ekman termed "rapid facial signals"—muscular movements lasting a few seconds, measurable in terms of action units and conveying emotions—happiness, sadness, anger, disgust, surprise, and fear—as well as cultural modifiers as "messages" (Ekman et al. 1992, page 11). The assembled experts agreed that measuring such fine-grained, dynamic movements would be difficult with existing computational techniques, but participants also expressed optimism due to recent technical advances, especially in neural networks. In a remarkable introduction, the authors describe a hypothetical example in which a set of 30x30 pixel facial images would be converted into a set of 900 grayscale point values, which would be passed through a network to produce an output of 0 for a label of "female" and 1 for the label "male"—a simplified scenario which could be extended to action unit or emotion labels. The weights of the network would initially be set in a random configuration, with the first output label having "no semblance to the desired output" (Ekman et al. 1992, page 34). The output error is calculated and then used to slightly adjust all the model's weights using the technique of backpropagation. This process is repeated until errors appear to level off or cross-validation techniques, in which the model is applied to examples held out at random from the training set, indicate an acceptable level of performance. The same basic description could be applied to many machine learning systems today.

23 At the time, however, this scenario was optimistic. Massive engineering challenges remained, above all creating labelled databases large enough so that the statistical dependencies associated with different action units could be learned. The authors concluded: "throughout this report, repeated references to the need for a database of facial information have emphasized the glaring lack of such a resource. The preceding sections have indicated many types of information that should be part of such a database, including images of the face (both still and motion), FACS scores, interpretations of facial signs, physiological data, digitized speech and sounds, and synthetic images" (Ekman et al. 1992, page 53). By sharing this standardized resource, the research community could not only save time and money but also create standardized benchmarks for evaluating models—an emerging norm of the machine learning community.

24 In due course, such resources were created such as the Cohn Kanade dataset and its successors (Takea Kanade et al. 2000; Lucey et al. 2010) and RU-FACS (Bartlett et al. 2006). These objects would be worth further study in order to understand more precisely how FACS labels were encoded or operationalized in machine learning training datasets, and how these standards enabled model evaluation and benchmarking. Indeed, these databases even addressed the major criticisms of posed expression research designs by incorporating "spontaneous expressions" (Bartlett et al. 2005). Without underestimating these engineering achievements or even, for the moment, evaluating the prior theoretical validity of Ekman's research program, I would like to emphasize the striking fact that it was possible, at least at a conceptual level, to envision a neural network-based machine learning model for understanding emotional expression in 1992, one whose form remains clearly legible, even if technology has advanced considerably in the interim.

4. Automating Emotion Recognition

25 More specifically, I am interested in what these technical objects (even in embryonic form) *do* to the theories of emotion expression and recognition—epistemic things— whose application they are meant to automate in what Rheinberger calls "stable subroutines" (Rheinberger 1997, page 81). This entails thinking further about the imperative of automation, so often both used as a self-evident justification for the development of machine learning systems and as a trope of stabilization in the history of science. It is becoming more common to critically observe that such talk of automation, implying a quantitative reduction in the amount of human labor to produce an equivalent output or judgment, in fact obscures reconfigurations or relocations of human labor in production processes (Taylor 2018)—often to labelers based in poor countries. Instead of thinking one-sidedly about the amount of labor that could be saved through automation—by replacing a trained human expert with a machine learning system for accurately labelling facial images—this critical perspective insightfully attends to the huge amount of human labor required to construct datasets, build, and evaluate machine learning models for emotion recognition systems.

26 A different critical perspective on automation would address qualitative transformations of the inputs and outputs themselves, contesting the quantitative relationship implied by the idea of automation in its instrumental sense, the idea that it involves only a more effective means for turning inputs into outputs or for stabilizing epistemic things. Here what matters is how both inputs and outputs themselves are transformed through automation processes. Not only do input images need to be standardized and processed in new ways so as to become legible to machine learning algorithms (hence the primordial importance of database construction for the NSF group) but our interpretation and evaluation of these outputs (label attributions) also changes, refracting the very theories and judgments that themselves were meant to be automated.

27 To further illustrate this idea, we can return to the work of Marian Bartlett, one of Sejnowski's doctoral students who actually built the kind of machine learning emotion recognition systems imagined in the 1992 workshop. Early results of this work are collected in her book *Face Image Analysis by Unsupervised Learning* (Bartlett 2001). The overall approach, with its emphasis on unsupervised learning and generative modeling, is striking in light of contemporary trends in machine learning. More specifically, Bartlett was experimenting with ways to automate the recognition of action units that did not rely on measuring distances between a small set of pre-determined features, such as measurements of noses, mouths, or eyes (Bartlett 2001, page 71). Instead, she used unsupervised machine learning algorithms to learn "holistic" representations, drawing on the entire facial image in order to subsequently classify an action using an AU label.

28 Often such unsupervised methods are justified in terms of moving towards a fuller automation of recognition or learning systems; by obviating the need for "manual," human specification of features or "priors" relevant to the learning task one presumably reduces the amount of human labor required to produce an output (Bartlett 2001, page 97). But these quantitative arguments are mixed with epistemological and even ethical evaluations. In this case, Bartlett speculates, reasonably, that manual specification of features may cause a system to overlook potentially relevant sources of information contained in the face, and, by analogy, that humans process faces in a holistic way, with the implication being machines should also leverage this holism. Furthermore, refraining from manually specifying features may also make it possible for classifiers to "discover" previously unknown

"relevant features" (Bartlett 2001, page 33), potentially *adding* to our theoretical knowledge of emotion rather than merely applying it. Here, technical objects are no longer merely question-answering machines but (may) produce unexpected, question-generating correlations.

29 This type of discourse is an instance of a more general epistemic norm in machine learning, which holds, perhaps counterintuitively, that by *refraining*, to the extent possible, of imposing some model—theoretical or otherwise—on data, we can discover things, say about how emotions are expressed and recognized, that we would not have expected, or at least will avoid being misled by our models. I stress that this case is decidedly not an example of the atheoretical caricature sometimes made of machine learning as an extreme form of empiricism (Breiman 2001). Ekman's theory played an absolutely central role: providing labels and more generally, through experimental design, a paradigm in which inferences are made based on facial images—which resonated in machine learning due to the field's own longstanding preoccupation with facial recognition. We can also note, at least at a formal level, that here machine learning shares the "non-intentionalist orientation of" Tomkins' original theory (Leys 2017, page 33): just as humans are frequently mistaken about the objects of our emotional experiences, we should also be wary of intentionally specifying relevant features for recognition tasks in advance. For both psychologists and the machine learning community, talk of automation has an ambiguous double sense. It was clearly used instrumentally to point to the impracticality of training expert labelers but also, through contact with machine learning practices, led to theoretical and indeed normative evaluations regarding how inputs and outputs should be transformed and interpreted: the more automated these processes are, the better they are able to cut through our misleading cognitive evaluations of emotions to underlying physiological truths.

5. Interpreting Emotional Outputs

30 In his foreword to Bartlett's book, Sejnowski makes a more general analogy, one typical of the interpretations that machine learning researchers make of their scientific objects, casting computing in terms of human experience and cognition: "Much of human learning," he writes, "is unsupervised; that is, without the benefit of an explicit teacher." Moving directly into the statistical schema of machine learning, he continues, "the goal of unsupervised learning is to discover the underlying probability distributions of sensory inputs." The next step is decidedly epistemological: "The identification of an object in an image nearly always depends on the physical causes of the image rather than the pixel intensities. Unsupervised learning can be used to solve the difficult problem of extracting the underlying causes…" (Bartlett 2001, page xiv). Of course, supervised forms of machine learning can be used—drawing on the theoretical framework of FACS in the case of emotion recognition—to subsequently classify responses or assign labels to these causes. The point is not to advocate for either supervised or unsupervised approaches but rather to understand their implications—the interpretations that these techniques generate.

31 Such statements illustrate how automation techniques (or our interpretation of these techniques) add to or transform rather than simply produce scientific "outputs" like emotion more efficiently. Action units or emotion labels result from complex, holistic, inductive means of processing image data—albeit by algorithms not human intuition—and may not be reducible to the types of facial features that conventionally matter to humans. Even further and more controversially, Sejnowski appears to endorse some sort of causal account of the statistical structures discovered by such models. They provide evidence for "underlying causes," of the structure of observed outputs—like action units or emotion labels. Indeed, his reflection implies the more "automated" a machine learning system is, the more it is capable of grasping the causes that generated the inputs—a type of technical realism where machines grasp causes unintelligible to humans. The admittedly loose and ambiguous language of "physical causes" points to the possibility of a less subjective identification of underlying emotional states than was possible, even for experts, in Ekman's human research subjects of the late 1960s and early 1970s.

32 All of this is of course speculative. In response, one legitimate critical strategy is to simply reject loose usage of causal language to characterize the use of machine learning techniques for emotion recognition and many other applications. From this perspective, a properly deflationary account of machine learning would point to

the illegitimacy of drawing any causal conclusions from the observational data used to train algorithms, even more so given their opacity. However, the price exacted by such strategies is to lose touch with the theoretical sources in psychology that did in fact inspire, however tenuously, developments in machine learning. For better or worse, this type of reflection or analogical thinking emerges as researchers unavoidably interpret their own engineering practices and the behaviors of technical objects that they design. My purpose in giving a more detailed account of the connection between a research program in emotion science and machine learning is to suggest another critical strategy, one that looks to compatibilities and shared epistemological assumptions that explains why these fields came together in the way that they did. This critical perspective seeks to better characterize the types of claims and even normative imperatives that become possible when they meet. The payoff of studying interpretations and evaluations of these techniques—even when they are dubious—is that they can reveal significant aspects of our values and culture.

33 The case of emotion science can shed special light on the normative structure of claims that are becoming commonplace as machine learning and AI are widely adopted. Returning to Sejnowski's original remarks on emotion recognition, now as an instance of more general style of reflection on machine learning techniques,[2] we can begin to characterize a wider epistemology or even worldview that results from the adoption and interpretation of these techniques. The first premise of this interpretation is relatively familiar, shared by much of statistical thought: the data gathered to train machine learning systems, its inputs, can be thought of as some sample from a notional probability distribution that generates them, to be estimated by the system and used to produce outputs, like emotion labels. These inputs might be images of faces for emotion recognition or text downloaded from the internet in the case of more recent large language models. Where these claims depart from the type of statistical modeling that predominated in the twentieth century is that it is not humans who specify a model—based on intuition, theoretical insight, or just convenience—or even relevant features of the data in order to analyze it. Rather, structures of the inputs are learned over the course of a training process (which may be supervised or unsupervised) governed by the iterative backpropagation of errors, moving towards better approximations of this "underlying distribution"—a kind of probabilistic ontology. These approximations are then evaluated not with traditional statistical tests of goodness of fit but more often through techniques like cross-validation, which measure the model's ability to correctly process inputs that have been artificially held out of the training set (e.g. Bartlett (2001, page 91)).

34 This renunciation of a certain level of "manual" human judgment in specifying features is often justified in the name of a fuller automation. I have argued here that such claims or imperatives can be further scrutinized for the ways that they not only reduce or reconfigure human labor (inputs), nor even as a process of technical stabilization of epistemic things, but also transform how we understand the outputs themselves. This is not, as some critics have suggested, a total renunciation to the point of a pure, positivist empiricism or an impenetrable black box. In the case of emotion recognition, theoretical constructs, not least the categorical emotion output labels are absolutely critical, as are neurophysiological action units in other systems. Rather, the capability of a system to be automated in this way, the ability of a model to identify some level of structure or pattern from inputs, is interpreted as evidence of some kind that the patterns in data correspond to the constructs used as labels or learning tasks—"underlying causes" in Sejnowski's words, although many other examples could be given (Bartlett 2001, page xiv). Emotion recognition constitutes a particularly compelling site for the emergence of this type of probabilistic worldview, as the unobservable, underlying affect or emotion programs, in themselves unintelligible to both psychologists and modelers, are nonetheless expressed as intelligible patterns on faces. Likewise, the ability of contemporary large machine learning models to produce fluent-sounding text is interpreted to mean that they grasp the fundamental (if not intelligible) structures of language for instance through compression (Delétang et al. 2023).

35 Diagnosing this kind of worldview, in which machine learning algorithms identify the underlying distributions from which the phenomenal world around us represents a sample or set of inputs does not offer any simple prescription for normative action or governance over these systems. Rather, by analyzing the concrete technical

2.A full account of the ways in which artificial intelligence and machine learning have prompted reflection on the nature of human cognition and even culture is of course outside the scope of this essay. For the views of an earlier generation see (McCorduck 2004). For an interesting philosophically informed view on more recent developments in machine learning see Buckner (2024). The more basic point is that it is obvious that these two have gone hand in hand.

and scientific contexts from which it emerged, we may better understand the normative implications of these probabilistic relationships between intelligibility and unintelligibility. Machine learning turns correlations into meaningful, actionable labels, which almost always have an evaluative and therefore normative dimension. One response to this is to reject these operations on scientific grounds as some critics contend. But another strategy might be to contest dominant or simply conventional evaluations of outputs—often articulated in terms of automation by the designers of these systems or those that wish to profit from them—in light of our own interpretations or values.

Note

36 The research has received funding from the European Research Council (ERC) under Horizon 2020, Advanced Investigator Grant ERC-2019-ADG-883107-ALGOSOC "Algorithmic Societies: Ethical Life in the Machine Learning Age."

References

Ackley, David H., Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. "A Learning Algorithm for Boltzmann Machines." *Cognitive Science* 9 (1): 147–69. https://doi.org/https://doi.org/10.1016/S0364-0213(85)80012-4.

Barrett, Lisa Feldman. 2006. "Are Emotions Natural Kinds?" *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 1 (1): 28–58. https://doi.org/https://doi.org/10.1111/j.1745-6916.2006.00003.x.

Barrett, Lisa Feldman, and Eliza Bliss-Moreau. 2009. "Affect as a Psychological Primitive." In *Advances in Experimental Social Psychology*, vol. 41. Elsevier Academic. https://doi.org/https://doi.org/10.1016/S0065-2601(08)00404-8.

Bartlett, Marian S. 2001. *Face Image Analysis by Unsupervised Learning*. Springer.

Bartlett, Marian S., Gwen C. Littlewort, Mark G. Frank, C. Lainscsek, I. Fasel, and Javier R. Movellan. 2006. "Automatic Recognition of Facial Actions in Spontaneous Expressions." *Journal of Multimedia* 1 (6): 22–35.

Bartlett, Marian S., Javier R. Movellan, Gwen C. Littlewort, Bjørn Braathen, Mark G. Frank, and Terrence J. Sejnowski. 2005. "Toward Automatic Recognition of Spontaneous Facial Actions." In *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, 2nd ed., edited by Paul Ekman and Erika L. Rosenberg. Oxford University Press.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery. https://doi.org/https://doi.org/10.1145/3442188.3445922.

Bledsoe, Woodrow W. 1965. "The Model Method in Facial Recognition." *Technical Report PRI* 15.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 26 (3): 199–231. https://doi.org/https://doi.org/10.1214/ss/1009213726.

Buckner, Cameron J. 2024. *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*. Oxford University Press.

Catanzariti, Benedetta. 2023. "Seeing Affect: Knowledge Infrastructures in Facial Expression Recognition Systems." University of Edinburgh. https://era.ed.ac.uk/handle/1842/40685.

Couldry, Nick, and Ulises A. Mejias. 2019. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford University Press.

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

Crawford, Kate. 2024. "Generative AI's Environmental Costs Are Soaring — and Mostly Secret." *Nature* 626 (8000): 693–693. https://doi.org/https://doi.org/10.1038/d41586-024-00478-x.

Darwin, Charles. 1998. *The Expression of Emotions in Man and Animals*. Edited by Paul Ekman. Oxford University Press.

Delaporte, François. 2008. *Anatomy of the Passions.* Translated by Susan Emanuel. Stanford University Press.

Delétang, Grégoire, Anian Ruoss, Paul-Ambroise Duquenne, et al. 2023. "Language Modeling Is Compression." Version 2. arXiv. https://doi.org/10.48550/ARXIV.2309.10668.

Duchenne de Boulogne, Emile. 1862. *Mécanisme de La Physionomie Humaine Ou Analyse Électro-Physiologique de l'expression Des Passions*. P. Asselin.

Ekman, Paul. 1982. "Methods for Measuring Facial Action." In *Handbook of Methods in Nonverbal Behavior Research*, edited by Klaus R. Scherer and Paul Ekman. Cambridge University Press.

Ekman, Paul, Wallace V. Friesen, and Silvan S. Tomkins. 1971. "Facial Affect Scoring Technique: A First Validity Study." *Semiotica* 3 (1): 37–58. https://doi.org/10.1515/semi.1971.3.1.37.

Ekman, Paul, Thomas Huang, Terrence J. Sejnowski, and Joseph C. Hager. 1992. *Final Report to NSF of the Planning Workshop on Facial Expression Understanding*. National Science Foundation. https://pdfs.semanticscholar.org/8725/2fdc7756af89f073bc4be5a0054d8dc74bfb.pdf.

Ekman, Paul, Richard Sorenson, and Wallace V. Friesen. 1969. "Pan-Cultural Elements in Facial Displays of Emotion." *Science* 164 (3875): 86–88. https://doi.org/https://doi.org/10.1126/science.164.3875.86.

Ellenbogen, Josh. 2012. *Reasoned and Unreasoned Images: The Photography of Bertillon, Galton, and Marey*. Pennsylvania State University Press.

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

Jaton, Florian. 2017. "We Get the Algorithms of Our Ground Truths: Designing Referential Databases in Digital Image Processing." *Social Studies of Science* 47 (6): 811–40. https://doi.org/10.1177/0306312717730428.

Kanade, Takea, J. F. Cohn, and Yingli Tian. 2000. "Comprehensive Database for Facial Expression Analysis." *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 46–53. https://doi.org/10.1109/AFGR.2000.840611.

Kanade, Takeo. 1977. *Computer Recognition of Human Faces*. Interdisciplinary Systems Research. Springer.

Koopman, Colin. 2013. *Genealogy as Critique: Foucault and the Problems of Modernity*. Indiana University Press.

Leys, Ruth. 2017. *The Ascent of Affect: Genealogy and Critique*. University of Chicago Press.

Lucey, Patrick, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression." *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 94–101. https://doi.org/10.1109/CVPRW.2010.5543262.

McCorduck, Pamela. 2004. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. AK Peters.

Pasquale, Frank. 2016. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Rheinberger, Hans-Jörg. 1997. *Towards a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press.

Rosenberg, Erika L. 2005. "Introduction: The Study of Spontaneous Facial Expressions in Psychology." In *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, edited by Paul Ekman and Erika L. Rosenberg. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195179644.001.0001.

Russell, James A. 1994. "Is There Universal Recognition of Emotion from Facial Expression? A Review of the Cross-Cultural Studies." *Psychological Bulletin* 115 (1): 102–41.

Stark, Luke. 2017. "Albert Ellis, Rational Therapy and the Media of 'Modern' Emotional Management." *History of the Human Sciences* 30 (4): 54–74. https://doi.org/10.1177/0952695117722720.

Stark, Luke. 2018. "Algorithmic Psychometrics and the Scalable Subject." *Social Studies of Science* 48 (2): 204–31. https://doi.org/10.1177/0306312718772094.

Taylor, Astra. 2018. "The Automation Charade." *Logic(s)* 5 (August). https://logicmag.io/failure/the-automation-charade/.

Tomkins, Silvan S. 2008. *Affect Imagery Consciousness: The Complete Edition*. I–II. Springer.

Tomkins, Silvan S., and Robert McCarter. 1964. "What and Where Are the Primary Affects? Some Evidence for a Theory." *Perceptual and Motor Skills* 18 (1): 119–58. https://doi.org/10.2466/pms.1964.18.1.119.

Turk, Matthew, and Alex Pentland. 1991. "Eigenfaces for Recognition." *Journal of Cognitive Neuroscience* 3 (1): 71–86. https://doi.org/10.1162/jocn.1991.3.1.71.

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