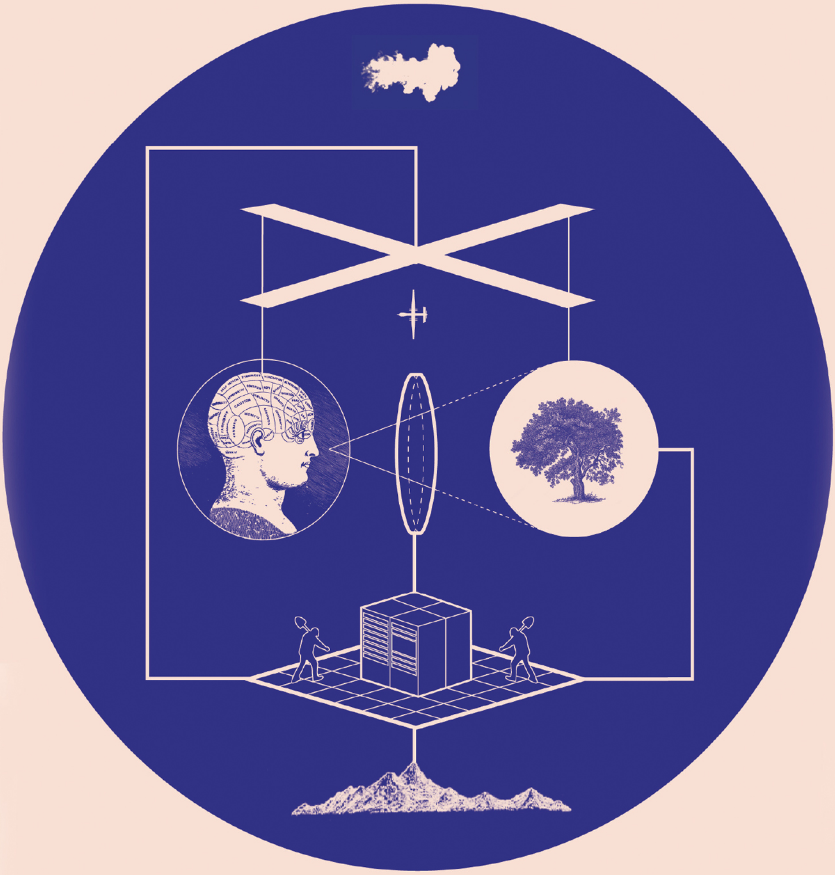
KATE CRAWFORD



# ATLAS OF AI

# Atlas of AI

*Power, Politics, and the Planetary Costs of Artificial Intelligence*

KATE CRAWFORD

Yale UNIVERSITY PRESS

New Haven and London

Yale University Press books may be purchased in quantity for educational, business, or promotional use. For information, please e-mail sales.press@yale.edu (U.S. office) or sales@yaleup.co.uk (U.K. office).

Cover design and chapter opening illustrations by Vladan Joler.
Set in Minion by Tseng Information Systems, Inc.
Printed in the United States of America.

Library of Congress Control Number: 2020947842
ISBN 978-0-300-20957-0 (hardcover : alk. paper)

A catalogue record for this book is available from the British Library.

This paper meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

10 9 8 7 6 5 4 3 2 1

# Contents

# 5
# Affect

In a remote outpost in the mountainous highlands of Papua New Guinea, a young American psychologist named Paul Ekman arrived with a collection of flashcards and a new theory.[1] It was 1967, and Ekman had heard that the Fore people of Okapa were so isolated from the wider world that they would be his ideal test subjects. Like many Western researchers before him, Ekman had come to Papua New Guinea to extract data from the indigenous community. He was gathering evidence to bolster a controversial hypothesis: that all humans exhibit a small number of universal emotions or affects that are natural, innate, cross-cultural, and the same all over the world. Although that claim remains tenuous, it has had far-reaching consequences: Ekman's presuppositions about emotions have grown into an expanding industry worth well over seventeen billion dollars.[2] This is the story of how affect recognition came to be part of artificial intelligence and the problems this presents.

In the tropics of Okapa, guided by medical researcher D. Carleton Gajdusek and anthropologist E. Richard Sorenson, Ekman hoped to run experiments that would assess how the Fore recognized emotions conveyed by facial expressions. Because the Fore had minimal contact with Westerners or mass media, Ekman theorized that their recognition and display of core expressions would prove that such expressions were universal. His methods were simple. He would show them flashcards of facial expressions and see if they described the emotion as he did. In Ekman's own words, "All I was doing was showing funny pictures."[3]

But Ekman had no training in Fore history, language, culture, or politics. His attempts to conduct his flashcard experiments using translators floundered; he and his subjects were exhausted by the process, which he described as like pulling teeth.[4] Ekman left Papua New Guinea, frustrated by his first attempt at cross-cultural research on emotional expression. But this would just be the beginning.

Today affect recognition tools can be found in national security systems and at airports, in education and hiring startups, from systems that purport to detect psychiatric illness to policing programs that claim to predict violence. By looking at the history of how computer-based emotion detection came to be, we can understand how its methods have raised both ethical concerns and scientific doubts. As we will see, the claim that a person's interior state of feeling can be accurately assessed by analyzing their face is premised on shaky evidence.[5] In fact, a comprehensive review of the available scientific literature on inferring emotions from facial movements published in 2019 was definitive: there is *no reliable evidence* that you can accurately predict someone's emotional state from their face.[6]

How did this collection of contested claims and experimental methodologies resolve into an approach that drives many parts of the affect AI industry? Why did the idea that there is a small set of universal emotions, readily interpreted from the face, become so accepted in the AI field, despite considerable evidence to the contrary? To understand that requires tracing how these ideas developed, long before AI emotion detection tools were built into the infrastructure of everyday life.

Ekman is just one of many people who have contributed to the theories behind affect recognition. But the rich and surprising history of Ekman's research illuminates some of the complex forces driving the field. His work is connected to U.S. intelligence funding of the human sciences during the Cold War through foundational work in the field of computer vision to the post-9/11 security programs employed to identify terrorists and right up to the current fashion for AI-based emotion recognition. It is a chronicle that combines ideology, economic policy, fear-based politics, and the desire to extract more information about people than they are willing to give.

## Emotion Prophets: When Feelings Pay

For the world's militaries, corporations, intelligence agencies, and police forces, the idea of automated affect recognition is as compelling as it is lucrative. It holds the promise of reliably filtering friend from foe, distinguishing lies from truths, and using the instruments of science to see into interior worlds.

Technology companies have captured immense volumes of surface-level imagery of human expressions—including billions of Instagram selfies, Pinterest portraits, TikTok videos, and Flickr photos. One of the many things made possible by this profusion of images is the attempt to extract the so-called hidden truth of interior emotional states using machine learning. Affect recognition is being built into several facial recognition platforms, from the biggest tech companies to small startups. Whereas facial recognition attempts to identify a *particular* individual, affect detection aims to detect and classify emotions by analyzing *any* face. These systems may not be doing what they purport to do, but they can nonetheless be powerful agents in influencing behavior and training people to perform in recognizable ways. These systems are already playing a role in shaping how people behave and how social institutions operate, despite a lack of substantial scientific evidence that they work.

Automated affect detection systems are now widely deployed, particularly in hiring. A startup in London called Human uses emotion recognition to analyze video interviews of job candidates. According to a report in the *Financial Times*, "The company claims it can spot the emotional expressions of prospective candidates and match them with personality traits"; the company then scores subjects on such personality traits as honesty or passion for a job.[7] The AI hiring company HireVue, which lists among its clients Goldman Sachs, Intel and Unilever, uses machine learning to assess facial cues to infer people's suitability for a job. In 2014, the company launched its AI system to extract microexpressions, tone of voice, and other variables from video job interviews, which they used to compare job applicants against the company's top performers.[8]

In January 2016, Apple acquired the startup Emotient, which claimed to have produced software capable of detecting emotions from images of faces.[9] Emotient grew out of academic research conducted at the University of California San Diego and is one of a number of startups working in this area.[10] Perhaps the largest of these is Affectiva, a company based in Boston that emerged from academic work done at Massachusetts Institute of Technology. At MIT, Rosalind Picard and her colleagues were part of an emergent wider field known as affective computing, which describes computing that "relates to, arises from, or deliberately influences emotion or other affective phenomena."[11]

Affectiva codes a variety of emotion-related applications, primarily using deep learning techniques. These range from detecting distracted and "risky" drivers on roads to measuring the emotional responses of consumers to advertising. The company has built what they call the world's largest emotion database, made up of over ten million people's expressions from eighty-seven countries.[12] Their monumental collection of videos of people emoting was hand labeled by crowdworkers based primarily in Cairo.[13] Many more companies have now licensed Affectiva's products to develop everything from applications that assess job candidates to analyzing whether students are engaged in class, all by capturing and analyzing their facial expressions and body language.[14]

Beyond the start-up sector, AI giants like Amazon, Microsoft, and IBM have all designed systems for affect and emotion detection. Microsoft offers emotion detection in its Face API, which claims to detect what an individual is feeling across the emotions of "anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise" and asserts that "these emotions are understood to be cross-culturally and universally communicated with particular facial expressions."[15] Amazon's Rekognition tool similarly claims that it can identify "all seven emotions" and "measure how these things change over time, such as constructing a timeline of the emotions of an actor."[16]

But how do these technologies work? Emotion recognition systems grew from the interstices between AI technologies, military priorities, and the behavioral sciences—psychology in particular. They share a similar set of blueprints and founding assumptions: that there is a small number of distinct and universal emotional categories, that we involuntarily reveal these emotions on our faces, and that they can be detected by machines. These articles of faith are so accepted in some fields that it can seem strange even to notice them, let alone question them. They are so ingrained that they have come to constitute "the common view."[17] But if we look at how emotions came to be taxonomized—neatly ordered and labeled—we see that questions are lying in wait at every corner. And a leading figure behind this approach is Paul Ekman.

## "The World's Most Famous Face-Reader"

Ekman's research began with a fortunate encounter with Silvan Tomkins, then an established psychologist based at Princeton who had published the first volume of his magnum opus, *Affect Imagery Consciousness,* in 1962.[18] Tomkins's work on affect had a huge influence on Ekman, who devoted much of his career to studying its implications. One aspect in particular played an outsized role: the idea that if affect was an innate set of evolutionary responses, they would be universal and so recognizable across cultures. This desire for universality has an important bearing on why these theories are widely applied in AI emotion recognition systems today: it offered a small set of principles that could be applied everywhere, a simplification of complexity that was easily replicable.

In the introduction to *Affect Imagery Consciousness,* Tomkins framed his theory of biologically based universal affects as one addressing an acute crisis of human sovereignty. He was challenging the development of behaviorism and psychoanalysis, two schools of thought that he believed treated consciousness as a mere by-product of—and in service to—other forces. He noted that human consciousness had "been challenged and reduced again and again, first by Copernicus"—who displaced man from the center of the universe—"then by Darwin"—whose theory of evolution shattered the idea that humans were created in the image of a Christian God—"and most of all by Freud"—who decentered human consciousness and reason as the driving force behind our motivations.[19] Tomkins continued, "The paradox of maximal control over nature and minimal control over human nature is in part a derivative of the neglect of the role of consciousness as a control mechanism."[20] To put it simply, *consciousness tells us little about why we feel and act the way we do.* This is a critical claim for all sorts of later applications of affect theory, which stress the inability of humans to recognize both the feeling and the expression of affects. If we as humans are incapable of truly detecting what we are feeling, then perhaps AI systems can do it for us?

Tomkins's theory of affects was his way to address the problem of human motivation. He argued that motivation was governed by two systems: affects and drives. Tomkins contended that drives tend to be closely associated with immediate biological needs such as hunger and thirst.[21] They are instrumental; the pain of hunger can be remedied with food. But the primary system governing human motivation and behavior is that of affects, involving positive and negative *feelings*. Affects, which play the most important role in human motivation, amplify drive signals, but they are much more complex. For example, it is difficult to know the precise reason or causes that lead a baby to cry, expressing the distress-anguish affect. The baby might be "hungry or cold or wet or in pain or [crying] because of a high temperature."[22] Similarly, there are a number of ways that this affective feeling can be managed: "Crying can be stopped by feeding, cuddling, making the room warmer, making it colder, taking the diaper pin out of his skin and so on."[23]

Tomkins concludes, "The price that is paid for this flexibility is ambiguity and error. The individual may or may not correctly identify the 'cause' of his fear or joy and may or may not learn to reduce his fear or maintain or recapture his joy. In this respect the affect system is not as simple a signal system as the drive system."[24] Affects, unlike drives, are not strictly instrumental; they have a high degree of independence from stimuli and objects, meaning that we often may not know why we feel angry, afraid, or happy.[25]

All of this ambiguity might suggest that the complexities of affects are impossible to untangle. How can we know anything about a system where the connections between cause and effect, stimulus and response, are so tenuous and uncertain? Tomkins proposed an answer: "The primary affects . . . seem to be innately related in a one-to-one fashion with an organ system which is extraordinarily visible." Namely, the face.[26] He found precedents for this emphasis on facial expression in two works published in the nineteenth century: Charles Darwin's *The Expression of the Emotions in Man and Animals* (1872) and an obscure volume by the French neurologist Guillaume-Benjamin-Amand Duchenne de Boulogne, *Mécanisme de la physionomie humaine ou Analyse électro-physiologique de l'expression des passions applicable à la pratique des arts plastiques* (1862).[27]

Tomkins assumed that the facial display of affects was a human universal. "Affects," Tomkins believed, "are sets of muscle, vascular, and glandular responses located in the face and also widely distributed through the body, which generate sensory feedback. . . . These organized sets of responses are triggered at subcortical centers where specific 'programs' for each distinct affect are stored"—a very early use of a computational metaphor for a human system.[28]

But Tomkins acknowledged that the *interpretation* of affective displays depends on individual, social, and cultural factors. He admitted that there were very different "dialects" of facial language in different societies.[29] Even the forefather of affect research raised the possibility that recognizing affect and emotion depends on social and cultural context. The potential conflict between cultural dialects and a biologically based, universal language had enormous implications for the study of facial expression and later forms of emotion recognition. Given that facial expressions are culturally variable, using them to train machine learning systems would inevitably mix together all sorts of different contexts, signals, and expectations.

During the mid-1960s, opportunity knocked at Ekman's door in the form of the Advanced Research Projects Agency (ARPA), a research arm of the Department of Defense. Looking back on this period, he admitted, "It wasn't my idea to do this [affect research]. I was asked—pushed. I didn't even write the research proposal. It was written for me by the man who gave me the money to do it."[30] In 1965, he was researching nonverbal expression in clinical settings and seeking funding to develop a research program at Stanford University. He arranged a meeting in Washington, D.C., with Lee Hough, head of ARPA's behavioral sciences division.[31] Hough was uninterested in how Ekman described his research, but he saw potential in understanding cross-cultural nonverbal communication.[32]

The only problem was that, by Ekman's own admission, he did not know how to do cross-cultural research: "I did not even know what the arguments were, the literature, or the methods."[33] So Ekman understandably decided to drop pursuit of ARPA funding. But Hough insisted, and according to Ekman, he "sat for a day in my office, and wrote the proposal he then funded that allowed me to do the research I am best known for—evidence for the universality of some facial expressions of emotion, and cultural differences in gestures."[34] He got a massive injection of funds from ARPA, roughly one million dollars—the equivalent of more than eight million dollars today.[35]

At the time, Ekman wondered why Hough seemed so eager to fund this research, even over his objections and despite his lack of expertise. It turns out that Hough wanted to distribute his money quickly to avoid suspicion from Senator Frank Church, who had caught Hough using social science research as a cover for acquiring information in Chile that could be used to overthrow its left-wing government under President Salvador Allende.[36] Ekman later concluded that he was just a lucky guy, someone "who could do overseas research that wouldn't get him [Hough] into trouble!"[37] ARPA would be the first in a long line of agencies from defense, intelligence, and law enforcement that would fund both Ekman's career and the field of affect recognition more generally.

With the support of a large grant behind him, Ekman began his first studies to prove universality in facial expression. In general, these studies followed a design that would be copied in early AI labs. He largely duplicated Tomkins's methods, even using Tomkins's photographs to test subjects drawn from Chile, Argentina, Brazil, the United States, and Japan.[38] He relied on asking research participants to simulate the expressions of an emotion, which were then compared with expressions gathered "in the wild," meaning outside of laboratory conditions.[39] Subjects were presented with photographs of posed facial expressions, selected by the designers as exemplifying or expressing a particularly "pure" or intense affect. Subjects were then asked to choose among these affect categories and to label the posed image. The analysis measured the degree to which the labels chosen by subjects correlated with those chosen by the designers.

From the start, the methodology had problems. Ekman's forced choice response format would be later criticized for alerting subjects to the connections that designers had already made between facial expressions and emotions.[40] Further, the fact that these emotions were faked or posed would raise significant concerns about the validity of these results.[41]

Ekman found some cross-cultural agreements using this approach, but his findings were challenged by the anthropologist Ray Birdwhistell, who suggested that this agreement may not reflect innate affect states if they were culturally learned through exposure to such mass media as films, television, or magazines.[42] It was this dispute that compelled Ekman to set out for Papua New Guinea, specifically to study indigenous people in the highlands region. He figured that if people with little contact to Western culture and media could agree with how he had categorized posed affective expressions, then this would provide strong evidence for the universality of his schema.

After Ekman returned from his first attempt to study the Fore people in Papua New Guinea, he devised an alternative approach to prove his theory. He showed his U.S. research subjects a photograph, then asked them to choose one of six affect concepts: happy, fear, disgust-contempt, anger, surprise, and sadness.[43] The results were close enough to subjects from other countries that Ekman believed he could claim that "particular facial behaviors are universally associated with particular emotions."[44]

## Affect: From Physiognomy to Photography

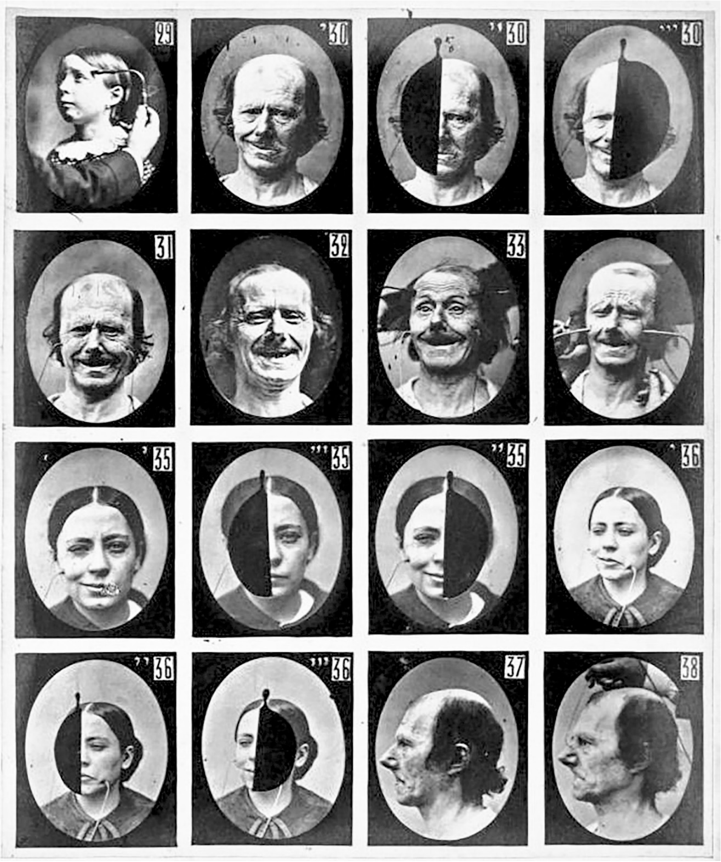
The idea that interior states can be reliably inferred from external signs stems in part from the history of physiognomy, which was premised on studying a person's facial features for indications of their character. In the ancient Greek world, Aristotle had believed that "it is possible to judge men's character from their physical appearance . . . for it has been assumed that body and soul are affected together."[45] The Greeks also used physiognomy as an early form of racial classification, applied to "the genus man itself, dividing him into races, in so far as they differ in appearance and in character (for instance Egyptians, Thracians and Scythians)."[46] They presumed a link between body and soul that justified reading a person's interior character based on their exterior appearance.

Physiognomy in Western culture reached a high point during the eighteenth and nineteenth centuries, when it was seen as part of the anatomical sciences. A key figure in this tradition was the Swiss pastor Johann Kaspar Lavater, who wrote *Essays on Physiognomy; For the Promotion of Knowledge and the Love of Mankind,* originally published in German in 1789.[47] Lavater took the approaches of physiognomy and blended them with the latest scientific knowledge. He tried to create a more "objective" comparison of faces by using silhouettes instead of artists' engravings because they were more mechanical and fixed the position of each face into the familiar profile form, allowing for a comparative viewpoint.[48] He believed that bone structure was an underlying connection between physical appearance and character type. If facial expressions were fleeting, skulls offered a more solid material for physiognomic inferences.[49] The measurement of skulls, as we saw in the last chapter, was used to support an emerging nationalism, racism, and xenophobia. This work was infamously elaborated on throughout the nineteenth century by phrenologists like Franz Joseph Gall and Johann Gaspar Spurzheim, as well as in scientific criminology through the work of Cesare Lombroso—all leading into the types of inferential classifications that recur in contemporary AI systems.

But it was the French neurologist Duchenne, described by Ekman as a "marvelously gifted observer," who codified the use of photography and other technical means in the study of human faces.[50] In *Mécanisme de la physionomie humaine,* Duchenne laid important foundations for both Darwin and Ekman, connecting older ideas from physiognomy and phrenology with more modern investigations into physiology and psychology. He replaced vague assertions about character with a more limited investigation into expression and interior mental or emotional states.[51]

Duchenne worked in Paris at the Salpetrière asylum, which housed up to five thousand people with a wide range of diagnoses of mental illness and neurological conditions. Some would become his subjects for distressing experiments, part of the long tradition of medical and technological experimentation on the most vulnerable and those who cannot refuse.[52] Duchenne, who was little known in the scientific community, decided to develop techniques of electrical shocks to stimulate isolated muscle movements in people's faces. His aim was to build a more complete anatomical and physiological understanding of the face. Duchenne used these methods to bridge the new psychological science and the much older study of physiognomic signs, or passions.[53] He relied on the latest photographic techniques, like collodion processing, which allowed for much shorter exposure times, allowing Duchenne to freeze fleeting muscular movements and facial expressions in images.[54]

Even at these very early stages, the faces were never natural or socially occurring human expressions but *simulations* produced by the brute application of electricity to the muscles. Regardless, Duchenne believed that the use of photography and other technical systems would transform the squishy business of representation into something objective and evidentiary, more suitable for scientific study.[55] In his introduction to *The Expression of the Emotions in Man and Animals,* Darwin praised Duchenne's "magnificent photographs" and included reproductions in his own work.[56] Because emotions were temporal, even fleeting occurrences, photography offered the ability to fix, compare, and categorize their visible expres-

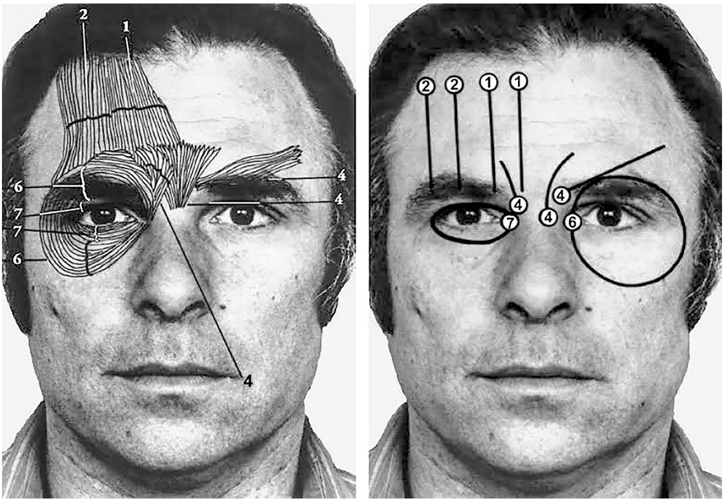


Plates from G.-B. Duchenne (de Boulogne),
*Mécanisme de la physionomie humaine, ou Analyse électro-physiologique de l'expression des passions.*
Courtesy U.S. National Library of Medicine

sion on the face. Yet Duchenne's images of truth were highly manufactured.

Ekman would follow Duchenne in placing photography at the center of his experimental practice.[57] He believed that slow motion photography was essential to his approach, because many facial expressions operate at the limits of human perception. The aim was to find so-called microexpressions—tiny muscle movements in the face. The duration of microexpressions, in his view, "is so short that they are at the threshold of recognition unless slow motion projection is utilized."[58] In later years Ekman also would insist that anyone could come to learn to recognize microexpressions, with no special training or slow motion capture, in about an hour.[59] But if these expressions are too quick for humans to recognize, how are they to be understood?[60]

One of Ekman's ambitious plans in his early research was to codify a system for detecting and analyzing facial expressions.[61] In 1971, he copublished a description of what he called the Facial Action Scoring Technique (FAST). Relying on posed photographs, the approach used six basic emotional types largely derived from Ekman's intuitions.[62] But FAST soon ran into problems when other scientists were able to produce facial expressions not included in its typology.[63] So Ekman decided to ground his next measurement tool in facial musculature, harkening back to Duchenne's original electroshock studies. Ekman identified roughly forty distinct muscular contractions on the face and called the basic components of each facial expression an Action Unit.[64] After some testing and validation, Ekman and Wallace Friesen published the Facial Action Coding System (FACS) in 1978; the updated editions continue to be widely used.[65] FACS was very labor intensive to use as a measurement tool. Ekman said that it took from seventy-five



Elements from the Facial Action Coding System.
Source: Paul Ekman and Wallace V. Friesen

to a hundred hours to train users in the FACS methodology and an hour to score a minute of facial footage.[66]

At a conference in the early 1980s, Ekman heard a research presentation that suggested a solution to the intense labor demands of FACS: the use of computers to automate measurement. Although in his memoir Ekman does not mention the researcher who gave the paper, he does state that the system was called Wizard and was developed at Brunel University in London.[67] This is likely Igor Aleksander's early machine learning object-recognition system, WISARD, which had used neural networks at a time when this approach was out of fashion.[68] Some sources report that WISARD was trained on a "database of known football hooligans," anticipating the widespread contemporary use of criminal mug shots to train facial recognition technologies.[69]

Because facial recognition emerged as a foundational application for artificial intelligence in the 1960s, it is not surprising that early researchers working in this field found common cause with Ekman's approach to analyzing faces.[70] Ekman himself claims to have played an active role in driving the automated forms of affect recognition through his old contacts in defense and intelligence agencies from his ARPA funding days. He helped to set up an informal competition between two teams working with FACS data, and this seems to have had lasting impact. Both of those teams have since gone on to feature prominently in the affective computing field. One team was composed of Terry Sejnowski and his student Marian Bartlett, who herself became an important figure in the computer science of emotion recognition and the lead scientist at Emotient, acquired by Apple in 2016.[71] The second team, based in Pittsburgh, was led by the psychologist Jeffrey Cohn of the University of Pittsburgh and the eminent computer vision researcher Takeo Kanade of Carnegie Mellon.[72] These two figures pursued affect recognition over the long term and developed the well-known Cohn-Kanade (CK) emotional expression dataset and its descendants.

Ekman's FACS system provided two things essential for later machine learning applications: a stable, discrete, finite set of labels that humans can use to categorize photographs of faces and a system for producing measurements. It promised to remove the difficult work of representing interior lives away from the purview of artists and novelists and bring it under the umbrella of a rational, knowable, and measurable rubric suitable to laboratories, corporations, and governments.

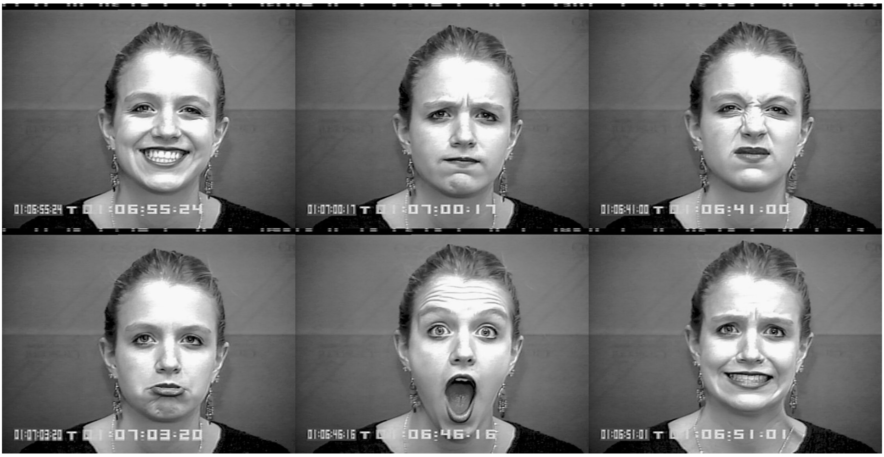
## Capturing Feeling: The Artifice of Performing Emotions

As work into the use of computers in affect recognition began to take shape, researchers recognized the need for a collection of standardized images to experiment with. A 1992 NSF report coauthored by Ekman recommended that "a readily accessible, multimedia database shared by the diverse facial research community would be an important resource for the resolution and extension of issues concerning facial understanding."[73] Within a year, the Department of Defense would begin funding the FERET program to collect facial photographs, as we saw in chapter 3. By the end of the decade, machine learning researchers had begun to assemble, label, and make public the datasets that drive much of today's machine learning research.

Ekman's FACS guidelines directly shaped the CK dataset.[74] Following Ekman's tradition of posed facial expressions, "subjects were instructed by an experimenter to perform a series of 23 facial displays," which FACS experts then coded, providing labels for the data. The CK dataset allowed laboratories to benchmark their results and compare progress as they built new expression recognition systems.

Other labs and companies worked on parallel projects, creating scores of photo databases. For example, researchers in a lab in Sweden created Karolinska Directed Emotional Faces. This database is composed of images of individuals portraying posed emotional expressions corresponding to Ekman's categories.[75] They make their faces into the shapes that accord with six basic emotional states. When looking at these training sets, it is difficult to not be struck by how extreme they are: *Incredible surprise! Abundant joy! Paralyzing fear!* These subjects are literally making machine-readable emotion.

As the field grew in scale and complexity, so did the types


01:06:55:24 T 01:06:55:24
01:07:00:17 T 01:07:00:17
01:06:41:00 T 01:06:41:00
01:07:03:20 T 01:07:03:20
01:06:46:16 T 01:06:46:16
01:06:51:01 T 01:06:51:01

Facial expressions from the Cohn-Kanade dataset: joy, anger, disgust, sadness, surprise, fear. Posed images from T. Kanade et al., *Yearbook of Physical Anthropology* (2000). © Cohn & Kanade

of photographs used in affect recognition. Researchers began using the FACS system to label data generated not from posed expressions but rather from spontaneous facial expressions, sometimes gathered outside of laboratory conditions. For example, a decade after the hugely successful release of the CK dataset, a group of researchers released a second generation, the Extended Cohn-Kanade (CK+) Dataset.[76] CK+ included the usual range of posed expressions but also began to include so-called non-posed or spontaneous expressions taken from videos where subjects made unprompted facial expressions.

By 2009, Affectiva emerged from the MIT Media Lab with the aim of capturing "naturalistic and spontaneous facial expressions" in real-life settings.[77] The company collected data by allowing users to opt into a system that would record their faces using a webcam as they watched a series of commercials. These images would then be hand-labeled using custom software by coders trained in Ekman's FACS.[78] But here we find another problem of circularity. FACS was developed from Ekman's substantial archive of posed photographs.[79] Even when images are gathered in naturalistic settings, they are commonly classified according to a scheme derived from posed images.

Ekman's work became a profound and wide-ranging influence on everything from lie detection software to computer vision. *The New York Times* described Ekman as "the world's most famous face reader," and *Time* named him one of the one hundred most influential people in the world. He would eventually consult with clients as disparate as the Dalai Lama, the FBI, the CIA, the Secret Service, and even the animation studio Pixar, which wanted to create more lifelike renderings of cartoon faces.[80] His ideas became part of popular culture, included in best sellers like Malcolm Gladwell's *Blink* and a television drama, *Lie to Me,* on which Ekman was a consultant for the lead character's role, apparently loosely based on him.[81]

His business also prospered: Ekman sold techniques of deception detection to security agencies such as the Transportation Security Administration, which used them in the development of the Screening of Passengers by Observation Techniques (SPOT) program. SPOT was used to monitor facial expressions of air travelers in the years following the September 11 attacks, attempting to "automatically" detect terrorists. The system uses a set of ninety-four criteria, all of which are allegedly signs of stress, fear, or deception. But looking for these responses meant that some groups are immediately disadvantaged. Anyone who was stressed, was uncomfortable under questioning, or had had negative experiences with police and border guards could score higher. This produced its own forms of racial profiling. The SPOT program has been criticized by the Government Accountability Office and civil liberties groups for its lack of scientific methodology and, despite its nine-hundred-million-dollar price tag, producing no clear successes.[82]

## The Many Critiques of Ekman's Theories

As Ekman's fame grew, so did the skepticism of his work, with critiques emerging from a number of fields. An early critic was the cultural anthropologist Margaret Mead, who debated Ekman on the question of the universality of emotions in the late 1960s, resulting in fierce exchanges not only between Mead and Ekman but also among other anthropologists critical of Ekman's idea of absolute universality.[83] Mead was unconvinced by Ekman's belief in universal, biological determinants of behavior rather than considering cultural factors.[84] In particular, Ekman tended to collapse emotions into an oversimplified, mutually exclusive binary: either emotions were universal or they were not. Critics like Mead pointed out that more nuanced positions were possible.[85] Mead took a middle ground, emphasizing that there was no inherent contradiction between "the possibility that human beings may share a core of innate behaviors . . . and the idea that emotional expressions could, *at the same time,* be highly-conditioned by cultural factors."[86]

More scientists from different fields joined the chorus over the decades. In more recent years, the psychologists James Russell and José-Miguel Fernández-Dols have shown that the most basic aspects of the science remain unsolved: "The most fundamental questions, such as whether 'facial expressions of emotion' in fact express emotions, remain subjects of great controversy."[87] Social scientists Maria Gendron and Lisa Feldman Barrett have pointed to the specific dangers of Ekman's theories being used by the AI industry because the automated detection of facial expressions does not reliably indicate an internal mental state.[88] As Barrett observes, "Companies can say whatever they want, but the data are clear. They can detect a scowl, but that's not the same thing as detecting anger."[89]

More troubling still is that in the field of the study of emotions, there is no consensus among researchers about what an emotion actually is. What emotions are, how they are formulated within us and expressed, what their physiological or neurobiological functions could be, their relation to stimuli, even how to define them—all of this in its entirety remains stubbornly unsettled.[90]

Perhaps the foremost critic of Ekman's theory of emotions is the historian of science Ruth Leys. In *The Ascent of Affect* she thoroughly pulls apart "the implications of the fundamental physiognomic assumption underlying Ekman's work . . . namely, the idea that a distinction can be strictly maintained between authentic and artificial expressions of emotion based on differences between the faces we make when we are alone and those we make when we are with others."[91] Leys sees a fundamental circularity in Ekman's method. First, the posed or simulated photographs he used were assumed to express a set of basic affective states, "already free of cultural influence."[92] Then, these photographs were used to elicit labels from different populations to demonstrate the universality of facial expressions. Leys points out the serious problem: Ekman assumed that "the facial expressions in the photographs he employed in his experiments must have been free of cultural taint because they were universally recognized. At the same time, he suggested that those facial expressions were universally recognized because they were free of cultural taint."[93] The approach is fundamentally recursive.[94]

Other problems became clear as Ekman's ideas were implemented in technical systems. As we've seen, many datasets underlying the field are based on actors simulating emotional states, performing for the camera. That means that AI systems are trained to recognize faked expressions of feeling. Although AI systems claim to have access to ground truth about natural interior states, they are trained on material that is inescapably constructed. Even for images that are captured of people responding to commercials or films, those people are aware they are being watched, which can change their responses.

The difficulty in automating the connection between facial movements and basic emotional categories leads to the larger question of whether emotions can be adequately grouped into a small number of discrete categories at all.[95] This view can be traced back to Tomkins, who argued that "each kind of emotion can be identified by a more or less unique signature response within the body."[96] But there is very little consistent evidence of this. Psychologists have conducted multiple reviews of the published evidence, which has failed to find associations among measurable responses to the emotional states that they assume to exist.[97] Finally, there is the stubborn issue that facial expressions may indicate little about our honest interior states, as anyone who has smiled without feeling truly happy can confirm.[98]

None of these serious questions about the basis for Ekman's claims have stopped his work from attaining a privileged role in current AI applications. Hundreds of papers cite Ekman's view of interpretable facial expressions as though it were unproblematic fact, despite decades of scientific controversy. Few computer scientists have even acknowledged this literature of uncertainty. The affective computing researcher Arvid Kappas, for example, directly names the lack of basic scientific consensus: "We know too little regarding the complex social modulators of facial and possibly other expressive activity in such situations to be able to measure emotional state reliably from expressive behavior. *This is not an engineering problem that could be solved with a better algorithm.*"[99] Unlike many in the field who confidently support affect recognition, Kappas questions the belief that it's a good idea for computers to be trying to sense emotions at all.[100]

The more time researchers from other backgrounds spend examining Ekman's work, the stronger the evidence against it grows. In 2019, Lisa Feldman Barrett led a research team that conducted a wide-ranging review of the literature on inferring emotions from facial expressions. They concluded firmly that facial expressions are far from indisputable and are "not 'fingerprints' or diagnostic displays" that reliably signal emotional states, let alone across cultures and contexts. Based on all the current evidence, the team observed, "It is not possible to confidently infer happiness from a smile, anger from a scowl, or sadness from a frown, as much of current technology tries to do when applying what are mistakenly believed to be the scientific facts."[101]

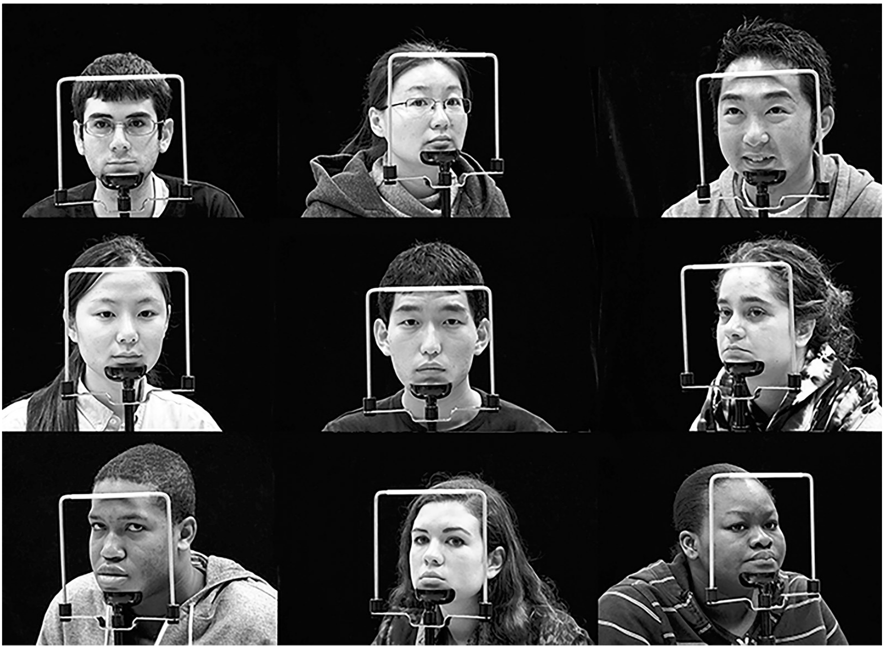
Barrett's team was critical of AI companies claiming to be able to automate the inference of emotion: "Technology companies, for example, are spending millions of research dollars to build devices to read emotions from faces, erroneously taking the common view as a fact that has strong scientific support. . . . In fact, our review of the scientific evidence indicates that very little is known about how and why certain facial movements express instances of emotion, particularly at a level of detail sufficient for such conclusions to be used in important, real-world applications."[102]

Why, with so many critiques, has the approach of "reading emotions" from the face endured? By analyzing the history of these ideas, we can begin to see how military research funding, policing priorities, and profit motives have shaped the field. Since the 1960s, driven by significant Department of

Defense funding, multiple systems have been developed that are increasingly accurate at measuring movements on faces. Once the theory emerged that it is possible to assess internal states by measuring facial movements and the technology was developed to measure them, people willingly adopted the underlying premise. The theory fit what the tools could do. Ekman's theories seemed ideal for the emerging field of computer vision because they could be automated at scale.

There are powerful institutional and corporate investments in the validity of Ekman's theories and methodologies. Recognizing that emotions are not easily classified, or that they're not reliably detectable from facial expressions, could undermine an expanding industry. In the AI field, Ekman is commonly cited as though the issue was settled, before directly proceeding into engineering challenges. The more complex issues of context, conditioning, relationality, and cultural factors are hard to reconcile with the current disciplinary approaches of computer science or the ambitions of the commercial tech sector. So Ekman's basic emotional categories became standard. More subtle approaches, like Mead's middle ground, were largely overlooked. The focus has been on increasing the accuracy rates of AI systems rather than on addressing the bigger questions about the many ways we experience, show, and hide emotion and how we interpret the facial expressions of others.

As Barrett writes, "Many of the most influential models in our science assume that emotions are biological categories imposed by nature, so that emotion categories are *recognized,* rather than constructed, by the human mind."[103] AI systems for emotion detection are premised on this idea. Recognition might be the wrong framework entirely when thinking about emotions because recognition assumes that emotional categories are givens, rather than emergent and relational.



Columbia Gaze Dataset. From Brian A. Smith et al., "Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction," *ACM Symposium on User Interface Software and Technology (UIST),* October 2013, 271–80. Courtesy of Brian A. Smith

## The Politics of Faces

Instead of trying to build more systems that can group expressions into machine-readable categories, we should question the origins of those categories themselves, as well as their social and political consequences. Already, affect recognition tools are being deployed in political attacks. For example, a conservative blog claimed to create a "virtual polygraph system" to assess videos of Congresswoman Ilhan Abdullahi Omar.[104] By using face and speech analytics from Amazon's Rekognition, XRVision Sentinel AI, and IBM Watson, the blogger claimed that Omar's analytic lie score exceeded her "truth baseline" and that she was registering high on stress, contempt, and nervousness. Several conservative media outlets ran with the story, claiming that Omar is a "pathological liar" and a security threat to the nation.[105]

It's known that these systems flag the speech affects of women differently from men, particularly Black women. As we saw in chapter 3, the construction of the "average" from unrepresentative training data is epistemologically suspect from the outset, with clear racial biases. A study conducted at the University of Maryland has shown that some facial recognition software interprets Black faces as having more negative emotions than white faces, particularly registering them as angrier and more contemptuous, even controlling for their degree of smiling.[106]

This is the danger of affect recognition tools. As we've seen, they take us back to the phrenological past, where spurious claims were made, allowed to stand, and deployed to support existing systems of power. The decades of scientific controversies around the idea of inferring distinct emotions from human faces underscores a central point: the one-size-fits-all recognition model is not the right metaphor for identifying emotional states. Emotions are complex, and they develop and change in relation to our families, friends, cultures, and histories, all the manifold contexts that live outside of the AI frame. In many cases, emotion detection systems do not do what they claim. Rather than directly measuring people's interior mental states, they merely statistically optimize correlations of certain physical characteristics among facial images. The scientific foundations of automated emotion detection are in question, yet a new generation of affect tools is already making inferences across a growing range of high-stakes contexts from policing to hiring.

Even though evidence now points to the unreliability of affect detection, companies continue to seek out new sources to mine for facial imagery, vying for the leading market share of a sector that promises billions in profits. Barrett's systemic review of the research behind inferring emotion from people's faces concludes on a damning note: "More generally, tech companies may well be asking a question that is fundamentally wrong. Efforts to simply 'read out' people's internal states from an analysis of their facial movements alone, without considering various aspects of context, are at best incomplete and at worst entirely lack validity, no matter how sophisticated the computational algorithms. . . . It is premature to use this technology to reach conclusions about what people feel on the basis of their facial movements."[107]

Until we resist the desire to automate affect recognition, we run the risk of job applicants being judged unfairly because their microexpressions do not match other employees, students receiving poorer grades than their peers because their faces indicate a lack of enthusiasm, and customers being detained because an AI system flagged them as likely shoplifters based on their facial cues.[108] These are the people who will bear the costs of systems that are not just technically imperfect but based on questionable methodologies.

The areas of life in which these systems are operating are expanding as rapidly as labs and corporations can create new markets for them. Yet they all rely on a narrow understanding of emotions—grown from Ekman's initial set of anger, happiness, surprise, disgust, sadness, and fear—to stand in for the infinite universe of human feeling and expression across space and time. This takes us back to the profound limitations of capturing the complexities of the world in a single classificatory schema. It returns us to the same problem we have seen repeated: the desire to oversimplify what is stubbornly complex so that it can be easily computed, and packaged for the market. AI systems are seeking to extract the mutable, private, divergent experiences of our corporeal selves, but the result is a cartoon sketch that cannot capture the nuances of emotional experience in the world.