"Excavating AI" Re-excavated: Debunking a Fallacious Account of the JAFFE Dataset

Michael J. Lyons
Ritsumeikan University

Abstract

Twenty-five years ago, my colleagues Miyuki Kamachi and Jiro Gyoba and I designed and photographed JAFFE, a set of facial expression images intended for use in a study of face perception. In 2019, without seeking permission or informing us, Kate Crawford and Trevor Paglen exhibited JAFFE in two widely publicized art shows. In addition, they published a nonfactual account of the images in the essay "Excavating AI: The Politics of Images in Machine Learning Training Sets." The present article recounts the creation of the JAFFE dataset and unravels each of Crawford and Paglen's fallacious statements. I also discuss JAFFE more broadly in connection with research on facial expression, affective computing, and human-computer interaction.

Keywords: facial expression, machine learning, training sets, affective computing, artificial intelligence

"To make a mistake and not correct it: this is a *real* mistake."

Confucius, The Analects

In September of 2019, Kate Crawford and Trevor Paglen opened the art exhibition *Training Humans* at the Milan Osservatorio of the Fondazione Prada. The stated objective of the exhibition was to explore:

> . . . *how humans are represented, interpreted and codified through training datasets, and how technological systems harvest, label and use this material.*

Accompanying the exhibition, the essay "Excavating AI: The Politics of Images in Machine Learning Training Sets" was self-published online.[1] To our

---

[1] https://excavating.ai/

surprise and dismay, *Training Humans* exhibited more than a hundred facial images taken from the JAFFE dataset,[2] a set of posed facial expressions photographed by my colleagues and me for use in our research [39]. Though we provide the JAFFE dataset to other researchers for purposes of non-commercial scientific research, Crawford and Paglen's public exhibition of the images was a violation of the terms of use [36]. Meanwhile, "Excavating AI" spotlighted JAFFE as an 'anatomical model' of machine learning training data, using it to expound an 'archaeological method' of "cataloguing the principles and values by which [it] was constructed." The essay so misportrayed JAFFE as to render it nearly unrecognizable by those most familiar with the dataset: its creators.

A previous commentary [36] outlined the errors in Crawford and Paglen's account of JAFFE and explained that their unauthorized use of the images violated not only the terms of use but also the principle of informed consent. By misusing JAFFE, the authors of "Excavating AI" breached the very ethics they claimed to espouse. Despite the objections raised in my commentary, they continue to propagate a fallacious account of JAFFE by republishing the nearly unmodified text of "Excavating AI"—it has now been reprinted in a publication[3] accompanying the 11th edition of the Liverpool Biennial of Contemporary Art [24] and published in a special issue of the peer-reviewed scholarly journal *AI & Society* [1].[4] This has motivated the present in-depth exposition of Crawford and Paglen's misleading statements about JAFFE.

The present article describes the JAFFE dataset in some detail, summarizing information from the published documentation of the dataset. Each of Crawford and Paglen's claims regarding JAFFE is explicitly refuted. Analysis of the general pattern of the claims shows that "Excavating AI" makes liberal use of the common fallacy of informal logic known as the *Straw Man* argument [56]. Concurrently, the article situates JAFFE in the broader context of experimental psychology, neural computation, affective computing, and human-computer interaction.

What is JAFFE?

JAFFE—The Japanese Female Facial Expression dataset—is a set of images depicting facial expressions posed by Japanese women, accompanied by semantic ratings on nouns describing the expressions. JAFFE was created specifically for use in an experiment that compared a biologically-inspired representation of facial appearance [30, 31, 40] with facial expression perception by human observers [27, 28, 29, 39]. JAFFE is described as a dataset because it includes observer ratings for each facial expression image. Without this data, it would be more accurate to describe JAFFE as an image set or stimulus set.

---

[2]JAFFE is available for non-commercial scientific research via the Zenodo repository. Note that all JAFFE images in this article are subject to specific terms of use and may not be reused without permission, regardless of the license applied to the document as a whole.

[3]https://www.biennial.com/journal/issue-9/

[4]An editor of the special issue assured me that the latter publication had indeed undergone peer review, but the text of the *AI & Society* version is nearly identical to the original web essay, casting doubt on the thoroughness of the referees' evaluations. As of June/July 2021, the article has been retracted from the *AI & Society* website but is still available elsewhere online.
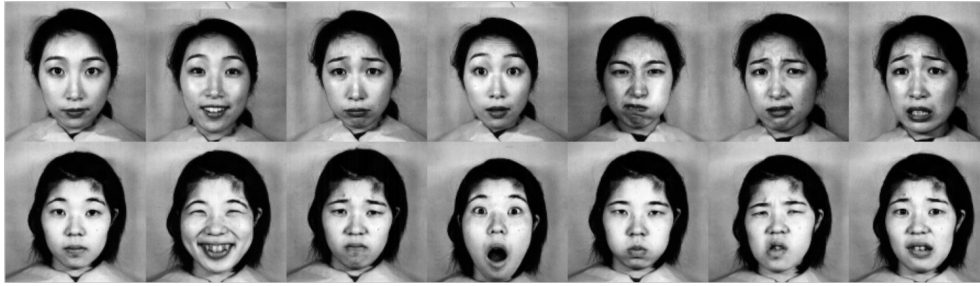
Figure 1: Posed facial expression images from the JAFFE dataset.

We planned and assembled JAFFE in 1996 in collaboration with Miyuki Kamachi and Jiro Gyoba of the Psychology Department at Kyushu University, Japan. Miyuki recruited volunteers in the Psychology Department and asked them to pose three or four examples of six facial expressions (happiness, sadness, surprise, anger, disgust, fear) and a neutral or expressionless face.[5] These six expressions are the *basic facial expressions* (BFE) studied by Paul Ekman and many others and thought by some of these scientists to be recognizable, to an extent, in all cultures [18].[6]

JAFFE volunteers posed the requested expressions while facing the camera through a transparent, semi-reflective plastic pane [39] and triggered the shutter remotely while viewing the reflection of their face—the JAFFE images are selfies. Unlike some other facial expression image sets (see, for example, [17, 42]), volunteers were not coached to pose specified facial configurations or to imitate prototypes, nor were they trained in the facial action coding system (FACS) [19]). The volunteers were given a Japanese adjectival noun for each of the six basic facial expressions and asked to photograph three or four poses for each of these while attending to their reflected expressions. JAFFE images did not undergo a selection procedure, as with, for example, Ekman's POFA (Pictures of Facial Affect) set, for which an iterative procedure selected the most prototypical images for each basic facial expression [17]. JAFFE includes nearly all of the photographs posed by all of the volunteers.[7] Because of this, the facial configuration for each expression category varies considerably between poses (figure 2) and between posers (figure 3).

We created JAFFE to experimentally test a mathematical model of facial expression representation derived from a simplified description of the receptive fields of neurons in the visual cortex [29, 30, 39, 40]. We aimed to compare the model with human perception. To reduce arbitrary assumptions, we

---

[5]Facial expression poses were requested using the Japanese adjectival nouns: 幸福、悲しみ、驚き、怒り、嫌悪、恐れ、and　無表情　for the expressionless pose. We used the same adjectival nouns to obtain semantic ratings by Japanese female observers.

[6]The universality hypothesis for facial expression is attributed to Charles Darwin, who thought that at least seven categories of facial expressions—the six BFE and additionally contempt—are recognizable by all humans [15]. The extent to which facial expressions are universal remains a matter of controversy [4, 5] and a topic of active research [11, 53].

[7]The publicly available version of JAFFE contains 213 images: ten expressers posing three (or in a few cases four) examples of six expressions and a neutral pose. The original, complete set contains 219 images, but the publicly available set omitted six of the images due to an oversight. There is nothing unusual about the six images missing from the published set.

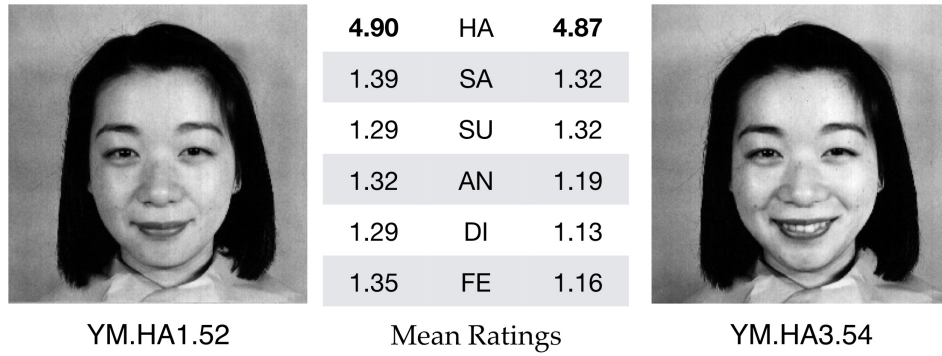| YM.HA1.52	Mean Ratings	YM.HA3.54
4.90	HA	**4.87**
1.39	SA	1.32
1.29	SU	1.32
1.32	AN	1.19
1.29	DI	1.13
1.35	FE	1.16

Figure 2: Two happy expressions posed by YM.

compared image similarity derived from the model with similarity measured from human observations of the facial expression images. We did not use categorical facial expression labels directly in the comparison. Instead, we obtained an empirical six-dimensional description for each image by asking observers to view expression images and rate the perceived expression using a five-level Likert scale for each of the six adjectival nouns used to request poses from the volunteers. Images were viewed and rated by 60 Japanese female observers (volunteers recruited at Kyushu University), and the responses averaged. Figure 2 shows sample mean ratings for two JAFFE images. The semantic ratings afford a more detailed and nuanced description of the images than that obtained with the more commonly used forced categorization approach [5]. Similarity calculated from the multidimensional ratings was correlated with similarity derived from our mathematical model. Model and empirical similarity measures were also analyzed using non-metric multidimensional scaling, revealing low dimensional structure resembling the affective circumplex [47, 50].

The approach outlined above does not require the JAFFE expressions to be prototypical. Strictly speaking, neither was it required that the volunteers pose authentic or felt expressions because we compared data on the *observers' perceptions* with the mathematical model, not the poser's *intended expression*. *Category labels are not directly used in our analysis* [39]. This does leave open the question of how posed expression images relate to facial expressions encountered 'in the wild,' and we were fully aware that our research project did not address this point.

A large number of scientific studies in the published literature make use of posed expressions [5]: our study and JAFFE are not unusual in this respect. As it happens, Lisa Feldman Barrett, whom Crawford and Paglen often cite as an Ekman critic, distributes the *IASLab Face Set*, a dataset of posed facial expression images categorized and labelled with emotion nouns.[8] IASLab does not include multidimensional semantic ratings data: would it not serve better than JAFFE as the 'Anatomical Model' for "Excavating AI"?

---

[8]The IASLab images and data are available for non-profit academic research at: https://www.affective-science.org/face-set.shtml.

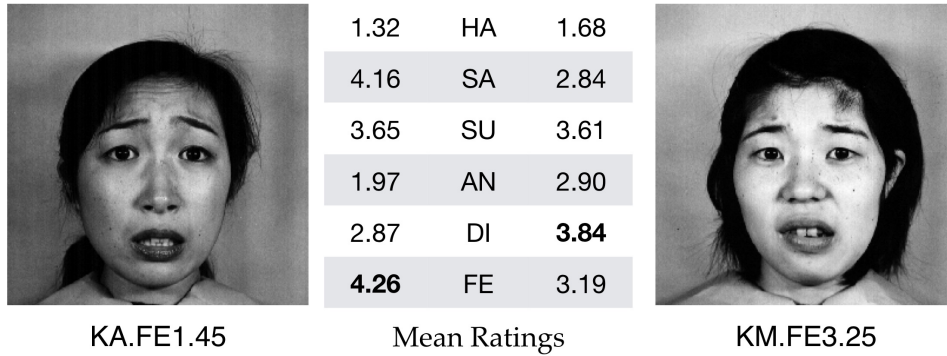|			
1.32	HA	1.68
4.16	SA	2.84
3.65	SU	3.61
1.97	AN	2.90
2.87	DI	**3.84**
4.26	FE	3.19

KA.FE1.45   Mean Ratings   KM.FE3.25

Figure 3: Fear expressions posed by KA and KM.

Cultural Specificity of JAFFE

As one might expect, JAFFE exhibits cultural specificity. At the outset of our project, when we were thinking through the design of JAFFE, Jiro Gyoba, an experienced perceptual psychologist, warned that we could not expect Japanese subjects to recognize images of fear poses reliably. Previous studies of facial expression recognition involving Japanese subjects [48] reported such findings. Most experiments have found that negative valence expressions are recognized less reliably in Japan than in Western cultures. This effect is most significant with fear expressions, where poser-observer agreement is measured to be just above chance.

Nevertheless, we included fear poses in the experiments. Our experimental observations confirmed Jiro's prediction (see Table 1 and figure 3): with fear poses, the agreement between the intended expression and the observer ratings is slightly above chance (17%). Fear poses are rated more strongly as surprise and disgust than fear itself (Table 1). These findings, however, did not change the conclusions of our study because we did not force categorization of the facial expressions. In parallel with the measurement on six expression words, we carried out independent measurements from a separate group of observers that did not include fear poses or ratings. Whether or not we

		RATED					
		HA	SA	SU	AN	DI	FE
P	HA	100%					
O	SA		97%			3%	
S	SU	3%		97%			
E	AN	3%			80%	17%	
D	DI	3%			6%	90%	
	FE		16%	31%		28%	25%

Table 1: JAFFE posed-rated agreement by facial expression. Agreement is reported here as the percentage of images for which the highest rated expression noun agrees with the requested pose. The average agreement is 81% over all poses and 93% if fear poses are not taken into account.

1.48	HA	1.45
2.48	SA	3.26
2.19	SU	1.77
4.48	AN	2.84
3.42	DI	**4.32**
2.16	FE	2.77

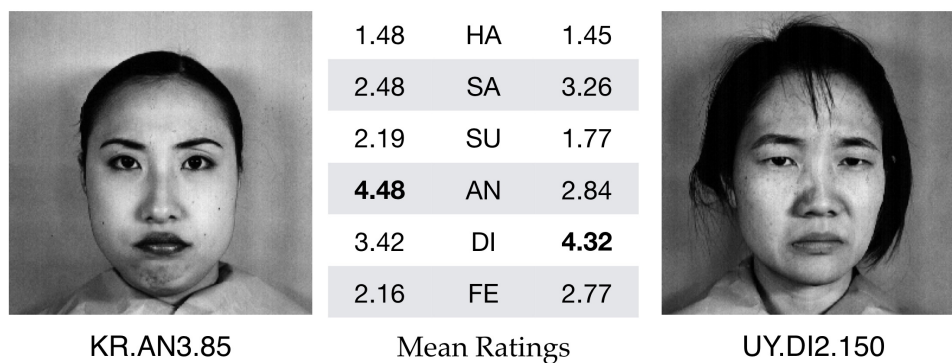| KR.AN3.85	Mean Ratings	UY.DI2.150

Figure 4: Anger and disgust expressions posed by KR and UY.

included fear poses and ratings, the visual-cortex-inspired model showed a significantly higher correlation with observer ratings than the control, a similarity measure derived from feature-point geometry.

We confirmed the cultural specificity of the JAFFE expressions a few years later in a collaborative cross-cultural study of facial expression judgments by Americans and Japanese [13, 14]. The study found, for example, that the anger pose shown in the left panel of figure 4 was classified as anger by 82% of Japanese study participants but only 34% of American participants. Similarly, the disgust pose shown in the right panel of figure 4 was classified as disgust by 66% of Japanese participants but only 18% of American participants [13].

JAFFE as a Training Set

We first reported the results of our project at the May 1997 edition of *ARVO*[9] a vision science conference held in Florida [29], and at a domestic cognitive science meeting held in Okinawa in June 1997 [28]. We learned that the *IEEE Face and Gesture* conference would take place not far from ATR,[10] and prepared an article for submission. The article [27] was shortlisted for a best paper award and invited to be published in a special issue of the Elsevier journal *Image and Vision Computing*.[11]

After hearing my talk, Zhengyou Zhang, a computer vision researcher who visited ATR in 1997, suggested we extend the visual-cortex-inspired model by training and testing a multi-layer neural network to classify the JAFFE facial expression images. By leveraging existing work, the project progressed quickly, and an article [58] was submitted to the same edition of the *IEEE Face and Gesture* conference. Zhang compared the classification rates obtained with the neural network directly with the poser-rater agreement levels (see table 1) to conclude that the neural network had exceeded human

---

[9]The Annual Meeting of the Association for Research in Vision and Ophthalmology.

[10]The Advanced Telecommunications Research Institute International (国際電気通信基礎技術研究所) an institute located in Kyoto devoted to basic research, where I worked from 1996-2007.

[11]The final manuscript that we prepared and submitted for publication in the *Image and Vision Computing* IEEE FG'98 special issue is available at the arXiv open access repository [39]. The special issue of *Image and Vision Computing* never materialized.

performance. However, this comparison was misleading because the neural network generalization rate measurement had not distinguished individual posers. Moreover, the human observers had rated the facial expression images without any initial instruction on poser intentions, whereas the neural network model was exposed to intended pose labels during training via supervised learning. In other words, the human observers were not 'trained' using the pose labels. In accord with my objections, the results of a subsequent study showed that we could equal or slightly exceed the performance of the multilayer neural network using a simple statistical model, while the cross-individual generalization rate was significantly lower [37].

In the interests of an *Open Data* policy [44], we decided to provide JAFFE images and ratings to others for use in non-commercial scientific research. To my surprise, JAFFE became a frequently used benchmark dataset in academic research on facial expression classification algorithms. By 2007, such studies had become so numerous that I publicly questioned the value of optimizing expression classification on benchmark datasets and cautioned that this was not the same research problem as emotion recognition [34].[12]

Despite the impression given by "Excavating AI," JAFFE is not used exclusively for machine learning research but also in many other fields, including visual psychology, cognitive science, and experimental neuroscience.

'Archaeology' of JAFFE: The Bones of Contention

Crawford and Paglen adopted JAFFE as their 'anatomical model' of machine learning training sets for the essay "Excavating AI: The Politics of Images in Machine Learning Training Sets" [12]. The relevant discussion is in the section titled "Anatomy of a Training Set," that they intend to serve as an exposition of what they describe as their 'dataset archaeology methodology.'

Nearly every statement about JAFFE in this important theoretical section of "Excavating AI" *is unequivocally false*. In the following, excerpts from Crawford and Paglen's account of JAFFE are quoted verbatim, in italics. My commentary, in plain font, follows each excerpt. The excerpts are quoted from the web version [12]. The text of the AI & Society article is nearly identical.

Emotional States

> *The dataset contains photographs of 10 Japanese female models making seven facial expressions that are meant to correlate with seven basic emotional states.*

We asked volunteers to freely pose six basic facial expressions as well as an expressionless face. We did not instruct or guide the volunteers to imitate specific prototypical facial action configurations. The volunteers were not 'models.'

---

[12]Invited talk given at Fechner Day 2007, the 23rd Annual Meeting of International Society for Psychophysics held in Tokyo, October 20–23, 2007.

Intended purpose

The intended purpose of the dataset is to help machine-learning systems recognize and label these emotions for newly captured, unlabelled images.

JAFFE was designed and photographed to compare a mathematical model of facial expression representation with data from human observers [27, 28, 29, 39]. Our project did not train a machine learning system, nor did it 'label emotions for newly captured images.' A central concern of "Excavating AI" is the critique of image taxonomy; ironically, Crawford and Paglen's mistaken description of JAFFE's origins as a machine-learning training set is itself a taxonomic error.

Depicting Emotions

The implicit, top-level taxonomy here is something like "facial expressions depicting the emotions of Japanese women."

We have never described JAFFE as anything but a set of images of posed facial expressions. The accompanying article [27, 39] describes in detail the procedure for photographing the JAFFE images. We made no explicit or implicit claim that these depict felt emotions.

Organizing Buckets

If we go down a level from taxonomy, we arrive at the level of the class. In the case of JAFFE, those classes are happiness, sadness, surprise, disgust, fear, anger, and neutral. These categories become the organizing buckets into which all of the individual images are stored. They are the distinct concepts used to order the underlying images.

The title of each JAFFE image file is composed of a two-letter volunteer identifier, two-letter pose identifier, pose number, and image number. The distribution includes a README_FIRST file containing multidimensional ratings for each image based on sixty Japanese observers' judgments. README_FIRST contains the following explicit caveat:

... it is important to realize that the expressions are never pure expressions of one emotion but always admixtures of different emotions. The image expression labels represent the predominant expression in that image—the expression that we asked the subject to pose.

Intended pose may not agree with the observer ratings, as in the right panel of figure 3. Crawford and Paglen's concept of categorical 'organizing buckets' ignores the nuanced description available from the subjective rating data and is a misleading characterization of JAFFE.

Emotions as Visual Concepts

> *There are several implicit assertions in the JAFFE set. First there's the taxonomy itself: that "emotions" is a valid set of visual concepts.*

JAFFE makes no such assertion, explicitly or implicitly. Our study modelled ratings by observers, not the emotions or mental states of those posing the expressions. Uninformed users of JAFFE might assume this about the image labels if they have not read and understood README_FIRST and the article that describes JAFFE [27].

Emotions applied to Photographs

> *Then there's a string of additional assumptions: that the concepts within "emotions" can be applied to photographs of people's faces specifically Japanese women) . . .*

We made no such claims about the emotions of the volunteers who posed the expressions. Our study modelled perceived expression based on observer ratings, not felt emotion. Crawford and Paglen have misunderstood this point. Our measurement protocol (semantic rating on expression nouns) is not unusual but widely used—the misunderstanding is symptomatic of a lack of basic familiarity with the published literature on the psychology of facial expression.

Six Emotions

> *. . . that there are six emotions plus a neutral state . . .*

Our work is concerned with the visual perception of facial expressions, not emotions of the posers. We have never stated that there are only six emotions or facial expressions, a nonsensical claim.

True Emotional State

> *... there is a fixed relationship between a person's facial expression and her true emotional state ...*

This premise has nothing to do with JAFFE, which is instead characterized by subjective ratings data and accompanied by our explicit caveat about the meaning of the pose labels.

Consistent and Uniform

> *. . . and that this relationship between the face and the emotion is consistent, measurable, and uniform across the women in the photographs.*

This premise cannot derive from actual observation of JAFFE images and data. Look again at the image samples and ratings in figures 1, 2, 3, and 4. Each volunteer posed several variations for each facial expression. No two images in the JAFFE set have identical facial expressions or rating values.

Neutral Facial Expression

> *At the level of the class, we find assumptions such as "there is such a thing as a "neutral" facial expression*

While the concept of a neutral facial expression is not meaningless, the ratings data indicate that none of the images depicting expressionless (無表情) poses were perceived by the observers as devoid of expression.

Significant Six Emotional States

> *and "the significant six emotional states are happy, sad, angry, disgusted, afraid, surprised."*

Crawford and Paglen have failed to understand a fundamental aspect of our research project: it is concerned with the *visual appearance of facial expressions* in images, not 'emotional states.' Regarding the significance of the facial expression categories chosen for the study, these are, of course, the *Basic Facial Expressions* (BFEs) used in many psychological studies of facial expression, not only by the Ekman camp. The use of these expression categories in research dates back at least to Charles Darwin's work [15]. Our project did not assume that these are the *only* expressions, nor did we make any such claim. We were well aware of criticism of BFE theory. Indeed one of the aims of our work was to compare categorical and dimensional descriptions of facial expressions [36, 39]. Furthermore, our project did not rely on the validity of the BFE paradigm.

Interior State

> *At the level of labelled image, there are other implicit assumptions such as "this particular photograph depicts a woman with an 'angry' facial expression," rather than, for example, the fact that this is an image of a woman mimicking an angry expression. These, of course, are all "performed" expressions—not relating to any interior state, but acted out in a laboratory setting.*

One wonders if Crawford and Paglen have looked at *any* of the documentation associated with JAFFE. We clearly describe the images as depicting *posed facial expressions*. Our article clearly describes the laboratory conditions under which we photographed JAFFE [27]. There is no claim or implication that JAFFE images depict natural expressions, and there is no discussion of 'interior state.'

Implicit Claims

> *Every one of the implicit claims made at each level is, at best, open to question, and some are deeply contested.*

An implication is a compound statement having the form *X implies Y.* But Crawford and Paglen's statements concerning JAFFE take the form '*Y is implied*,' without specifying a premise, *X*, and without providing the slightest

explanation about how these unknown factors imply the alleged 'claims.' All of the allegations about JAFFE's 'implicit claims' are furnished as self-evident conclusions made without evidence, elaboration, or reference to statements in the relevant publications.

Relative Modesty

> *The JAFFE training set is relatively modest as far as contemporary training sets go. It was created before the advent of social media, before developers were able to scrape images from the internet at scale,*

Crawford and Paglen add to the fiction that we intended JAFFE as a training set. The fabulation continues: JAFFE is a scale-challenged 'modest' training set, stunted by the impossibility of scraping for images. Images scraped online would not have been suitable for the project. JAFFE was designed and explicitly photographed for our experiment. The relatively small size of the JAFFE set had nothing to do with a lack of big data: we chose the numbers of volunteers and poses to permit a statistically significant comparison of two measures of facial expression similarity [39]. Had we wished to design an image dataset for facial expression machine learning, we would have recruited more volunteers.

The Straw Men

Each of the alleged 'implicit claims' about JAFFE is a misrepresentation of our work. As a member of the team that created JAFFE, I can state unequivocally that these allegations do not correspond with our views or intentions.

How did Crawford and Paglen arrive at such a mistaken account of our work? It appears that the veracity of their allegations about JAFFE was not a priority. Instead, these serve a mainly rhetorical function. Examining the list of statements reveals a pattern:

- Each statement is made without evidence (e.g. 'JAFFE is intended as a machine learning training set')
- Each of the 'implied claims' is easy to refute without specialized knowledge (e.g. 'posed images depict interior states')
- Some of the 'implied claims' are ridiculous (e.g. 'there are only six emotions')

A *straw man* argument is a common fallacy of informal reasoning that distorts or misrepresents a position in order to make it easier to refute [56]. This fallacy provides the key to understanding Crawford and Paglen's account of our work. As we have seen, "Excavating AI" does not critique factual attributes of JAFFE, but nonfactual 'implied claims'—misrepresentations posited without a shred of evidence. Each false allegation is simplistic and intuitively refutable. Some are ridiculous: what kind of deranged charlatan would propose that there are no more than 'six facial expressions and a neutral face' and that it is possible to 'mind read' from a single photograph of a posed expression? These ridiculous caricatures, fabulated by Crawford and Paglen

in "Excavating AI," appear entirely believable by readers who have limited knowledge of the research literature on the psychology of facial expression.

The straw man offers several incentives: it is easier to communicate simplistic, fallacious views to a non-specialist reader than faithful, nuanced ones. Imagined 'claims' can be constructed for ease of refutation. Distorted caricatures are more memorable than plain reality. An influential study has found that falsehood spreads more quickly on social networks than the truth [55]. Another recent study found that populist politicians employ informal fallacies to attract attention and persuade [6]. Does "Excavating AI" frame JAFFE using straw man arguments for rhetorical purposes? This is the most plausible explanation I have found for the consistently erroneous account of JAFFE given in the essay.

Cognitive Biases

"Excavating AI" has enjoyed a largely positive reception by humanities scholars who cite it as having identified problematic epistemology in machine learning training sets. We do not deny the importance of this issue—there is no doubt that there are machine learning datasets embodying questionable epistemology. Nonetheless, it is surprising that the incorrect description of JAFFE has slipped past the discernment of so many: how can the fallacies have escaped the critical faculties of these readers?

First, consider the context in which "Excavating AI" has appeared: the decade preceding its publication has witnessed accelerated interest and investment in artificial intelligence and machine learning and the intensification of ideology that favours technology as the primary driver of social and economic innovation. Resistance has also grown: increasingly, the critical study of technology is a valued scholarly activity that is impacting public discourse and policy. Readers of "Excavating AI" who are skeptical of AI hype may be susceptible to *confirmation bias* [10] that leads them to overlook invalid parts of Crawford and Paglen's argument unwittingly.

The *halo effect* [10] could be another factor: Crawford and Paglen are frequent, high-profile speakers at public events, sometimes portrayed in the media as heroic social activists: we can expect a level of positive bias.

Narrative bias [7] may also be a factor: "Excavating AI" offers readers a compelling narrative that can overrule judgments on the veracity of its elements. Readers may assume that the 'implied claims' linked to JAFFE are factual, even without being offered evidence, because this interpretation fits the essay's overall theme: false image taxonomy. Ironically, René Magritte's *La Trahison des Images*, a point of reference in Crawford and Paglen's thesis, is relevant here: a false ontology accompanies the JAFFE images—the straw man arguments offer the reader/viewer a deceptive *misrepresentation*.

Crawford and Paglen provoked a media spectacle involving: the viral selfie app *ImageNet Roulette*; the highly visible art exhibitions *Training Humans* and *Making Faces* sponsored by global fashion brand *Prada*; mass media exposure that included coverage by the Netflix tech-lite variety show *Connections*; and torrents of activity on the social networks [36, 54]. Public attention accrued to such a degree that the *bandwagon effect* [10] may significantly bias consumers'

judgments, even if widespread popularity is no evidence for validity.

To recognize the misinformation, a reader would need to examine and understand the documentation accompanying JAFFE. It is unlikely that most readers will go to such lengths. Taken together with the biases I have outlined, this helps us understand how the fallacies concerning JAFFE pass inconspicuously.

Mind Reading Machines?

One of the most striking of the 'implied claims' in "Excavating AI" concerns the purported ability to read 'true emotional state' or 'interior state' from images of posed facial expressions. As I have shown, our work made no claim to read internal states. Where did Crawford and Paglen get such ideas? Perhaps their imaginations were stimulated by media reports of emotion reading software and the websites of affective computing startups?
Here are some press headlines and excerpts from company web pages:

- *Apple Buys Emotient, a Company That Uses AI to Read Emotions* [8]
- *. . . Startup That Can Read Your Face To Tell If You're Angry* [52]
- *AI Can Read Your Emotions. Should it?* [26]
- *Human Perception AI. Our software detects all things human: nuanced emotions, complex cognitive states. . .* [13]

My first encounter with this genre of publicity dates to about 2006, when the New York Times reported [51] on a project by Rana el Kaliouby and Rosalind Picard of the MIT Media Lab, who would found the startup *Affectiva* a few years later [20]. The article described an "Emotional-Social Intelligence Prosthesis" that promised to augment an autistic user's ability to 'mind read' by providing facial expression recognition support. Some might naïvely misinterpret the mention of *mind reading* as an 'implied claim' hinting at telepathy or phrenology. It is nothing of the kind: the project and el Kaliouby's earlier Ph.D. thesis [21] drew on the *Theory of Mind* theory of cognitive science— the hypothesis that we use folk psychology to make inferences about the beliefs and intentions of other beings [3, 9, 16, 46]. The New York Times report mentioned Simon Baron-Cohen, who proposed a model of autism based on the *Theory of Mind* theory [2, 3] and was one of el Kaliouby's thesis advisors.

I was impressed by el Kaliouby's ambitious attempt to model facial expression perception more meaningfully than the image classification studies still commonplace. At the same time, I was uncomfortable with the anthropomorphic description of computer software. I expressed my reservations at conferences in Tokyo in 2007 [34][14] and in Amsterdam in 2008 [35].[15] My skepticism was twofold. First, as mentioned above, the term 'mind reading' has alternative quasi-magical connotations likely to mislead the popular imagination. The researchers could have avoided potential misunderstandings by describing their work as, for example, a computational model of the theory of

---

[13]*Affectiva* homepage: https://www.affectiva.com/ (accessed July 12, 2021).

[14]The Fechner Day, 2007 conference mentioned in footnote 12.

[15]IEEE Face and Gesture Conference, Workshop on Facial and Bodily Expressions for Control and Adaptation of Games, Amsterdam, NL, September 16, 2008.

mind, which would be accurate though less sensational. Second, there were practical difficulties involved in interpreting facial displays. My presentations in Tokyo and Amsterdam gave a simple demonstration[16] showing that the crucial role played by context would be difficult to take into account generally. Barring a solution of the hard AI problem, one did not expect 'mind reading' systems to function except in rather constrained situations. Indeed, as late as 2011, after several years of further development, *Affectiva*'s technology was still having difficulty distinguishing some types of grimace from smiles [20].

Affectiva evolved and is valued by investors and clients in the automotive, marketing, and advertising industries, but Shoshana Zuboff argues that, despite initial good intentions, *Affectiva* was drawn into "surveillance capitalism's powerful force field."[17] I wonder if the outcome would have been more benign had el Kaliouby continued to pursue the computational modelling of the *Theory of Mind* theory scientifically rather than commercially?

Challenging the Common View

In December 2000, I took part in a workshop on affective computing held as part of the Neural Information Processing Systems conference (aka NeurIPS).[18] The call for participation encouraged speakers

> . . . *to talk about challenges and controversial topics both in their prepared talks and in the ensuing discussions.*[19]

My contribution offered two challenges and a proposal [32].[20]

My first challenge to the workshop reported our experimental finding that head tilt influences facial expression perception. Our cross-cultural study of this phenomenon, in collaboration with Ruth Campbell at the University College London, was inspired by the variable expressions of masks used in the Japanese Noh theatre [38]. Tilting the mask toward or away from a viewer changed the perceived expression. The effect differed for Japanese and British viewers. Our experiments demonstrated that a related effect obtains with human faces. This revealed a challenge: most research on facial expression recognition had assumed frontal, untilted views of the face. If slight changes in head pose could change the apparent expression, how had this influenced psychological studies of expression perception? Moreover, our results showed clearly that automatic facial expression recognition algorithms would need to account for head pose variations.

My second challenge concerned the cultural variability of facial expressions. Our study of Noh masks had revealed a cultural difference between Japanese and British experimental subjects. Furthermore, we had already observed cultural specificity with the JAFFE images and ratings data [39]. I recommended that affective computing researchers should carefully re-examine the facial expression universality hypothesis.

---

[16]See the Amsterdam presentation slides: https://zenodo.org/record/4026982.

[17]Chapter 9, section III, "Machine Emotion," of *The Age of Surveillance Capitalism* [59].

[18]https://www.cs.cmu.edu/Groups/NIPS/NIPS2000/Workshops/

[19]https://www.kdnuggets.com/news/2000/n20/34i.html

[20]NeurIPS'2000 Workshop presentation slides: https://zenodo.org/record/5016197.

Next, I proposed an alternative paradigm for facial expression computing, suggesting that it would be worthwhile to pursue the development of systems permitting the use of *voluntary* facial gestures for *intentional* human-computer interaction. I had already been working on such *facial gesture interfaces* [33]. To illustrate the concept, I showed workshop attendees a video of my *mouthesizer* prototype: a wearable facial gesture interface allowing musical performers to control audio effects in real-time with their mouths [41].

My presentation concluded with the announcement of a workshop we were organizing at the upcoming ACM CHI'2001 conference[21] in Seattle, entitled "New Interfaces for Musical Expression," and encouraged interested listeners to get involved. The NIME workshop would explore prospects for interdisciplinary research involving human-computer interaction and musical expression [45].

I wonder what impact, if any, the unconventional presentation had on the workshop attendees. Several years later, three of the four workshop organizers founded the startups *Affectiva* and *Emotient* to commercialize technology to recognize 'emotion' displayed involuntarily by the face. In the audience was the director of the US Department of Defence sponsored Facial Recognition Technology (FERET) program.[22] He had organized competitive tests of automatic facial recognition algorithms and told me he was now looking into possibilities in the affective computing field. I recall asking the other speakers during the coffee break if they were not concerned about the potential exploitation of their research for surveillance purposes.

My presentation had at least one tangible outcome: after the workshop, discussion continued with the fourth workshop organizer, cognitive scientist Gary Cottrell: it was the beginning of our collaborative study of culturally dependent aspects of facial expression perception [13, 14].

The NeurIPS'2000 Affective Computing workshop is a personally significant landmark that signalled a shift in my research activities: I had already begun to prioritize the voluntary, intentional, and expressive aspects of action and enaction. It also marked the gestation of a new research community, NIME,[23] that would grow and confer every year after the first CHI'2001 workshop. NIME offers a leading venue for scientific research and artistic practice at the intersection of musical expression and technology [23]. In December 2000, NIME extended the possibility of a new locus of positive resistance against the infiltration of the surveillance paradigm into the human-machine relationship and an exploration of the potential of technology in the realm of artistic expression.

Algorithms of Inequity

If Crawford and Paglen's essay is an excavation, then JAFFE is its Piltdown Man, a hoax created by would-be archaeologists [25]. As we have seen, the JAFFE dataset was abused as fodder for rhetorical straw man arguments about

---

[21] ACM Conference on Human Factors in Computing Systems, March 31 - April 5, 2001.

[22] https://www.nist.gov/programs-projects/face-recognition-technology-feret

[23] NIME Community Portal: https://www.nime.org/.

machine learning training data. But what can we conclude about the essay in general? Is it not advisable to scrutinize other aspects of "Excavating AI"?

Consider the overall focus on training set taxonomy. Imagine, for purposes of argument, that collaboration between socially aware engineers and humanities scholars has managed to resolve the data taxonomy issues raised in "Excavating AI." Would this lead to the development of ethical facial recognition technology? Should the possibility of reducing bias ease concerns about the deployment of surveillance systems? Will dataset debiasing efforts by the surveillance capitalist behemoths amount to anything more than *digital ethics bluewashing* [22]?

Whether for policing or profit, surveillance technology affords mediated social relations that are, by definition, unequal.[24] Automated surveillance is a species of machine-mediated interaction for which the framing of *interaction* as an *exchange of information* lacking *detailed reciprocity* leads directly to algorithms for the generation of inequity. Taxonomy fixes will never repair the root issue: surveillance presupposes a bias that privileges the watcher over the watched. By contrast, ethical human-machine systems afford transparency, volition, and reciprocity. Such systems do not engender or enhance power asymmetries and social inequity. I embraced these views in the 1990s, when a scientific interest in face perception briefly brought me into contact with the field of surveillance technology.

Lessons Learned?

I want to share some of the insights gained from debunking Crawford and Paglen's fallacious account of the JAFFE dataset. These lessons may seem self-evident but will be stated explicitly in the general spirit of this commentary.

Foremost, before criticizing something, it is necessary to understand it correctly. If there is a README file, read it. If there is associated documentation, study it. Failing to do this, one generates misinformation and misleads readers.

If anything is unclear in the documentation, and even if there is not, get in touch with the authors. Find out what they think about your analysis.

If a target of your critique raises objections, listen to what they have to say and understand their viewpoint. To do so is both good research practice and a matter of basic respect for other human beings.

If some part of your critique is mistaken, admit it. Sooner is better than later: avoid escalating a commitment to beliefs that have already been proven to be incorrect.

Before using facial images for an art exhibition, or publicity, clearly understand the terms of use and confirm the informed consent of the persons depicted [36]. The pamphlet entitled "Making AI Art Responsibly: A Field Guide" by the Partnership on AI [49][25] offers good advice about reusing image

---

[24]From the French, the verb *surveiller* adds the prefix *sur-* (over) to *veiller* (to watch) meaning 'to over watch' implying unequal vigilance.

[25]https://www.partnershiponai.org/wp-content/uploads/2020/09/Partnership-on-AI-AI-Art-Field-Guide.pdf

data and other resources. Do not misjudge the informal presentation style: the guide is well researched and based on a solid understanding of its subject.

Finally, to critique affective computing technology, choose a worthy target rather than a convenient straw man. The leading companies in the field began as research projects that spawned startups: paper trails of academic publications and patent applications are readily available. Less well documented approaches should be amenable to fieldwork and some reverse engineering. The recently published investigation of the *Vibraimage* system is an excellent example of thorough and productive critical research [57].

Michael Lyons is Professor of Image Arts and Science at Ritsumeikan University

References

1. AZAR, M., COX, G., AND IMPETT, L., Eds. *AI and Society, Special Issue: Ways of Machine Seeing.* Springer, 2021. 2

2. BARON-COHEN, S. *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press, 1997. 13

3. BARON-COHEN, S., LESLIE, A. M., AND FRITH, U. Does the autistic child have a theory of mind? *Cognition 21*, 1 (1985), 37–46. 13

4. BARRETT, L. F. AI weighs in on debate about universal facial expressions. *Nature 589* (2021), 202–203. 3

5. BARRETT, L. F., ADOLPHS, R., MARSELLA, S., MARTINEZ, A. M., AND POLLAK, S. D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest 20*, 1 (2019), 1–68. 3, 4

6. BLASSNIG, S., BÜCHEL, F., ERNST, N., AND ENGESSER, S. Populism and informal fallacies: An analysis of right-wing populist rhetoric in election campaigns. *Argumentation 33*, 1 (2019), 107–136. 12

7. BORGIDA, E., AND NISBETT, R. E. The differential impact of abstract vs. concrete information on decisions. *Journal of Applied Social Psychology 7*, 3 (1977), 258–271. 12

8. BYFORD, S. Apple buys Emotient, a company that uses AI to read emotions. *The Verge* (2016). 13

9. CHURCHLAND, P. S. *Neurophilosophy: Toward a Unified Science of the Mind-brain.* MIT Press, 1986. 13

10. COLMAN, A. M. *A Dictionary of Psychology.* Oxford quick reference, 2015. 12

11. COWEN, A. S., KELTNER, D., SCHROFF, F., JOU, B., ADAM, H., AND PRASAD, G. Sixteen facial expressions occur in similar contexts worldwide. *Nature 589*, 7841 (2021), 251–257. 3

12. CRAWFORD, K., AND PAGLEN, T. Excavating AI: The politics of images in machine learning training sets. https://www.excavating.ai/, 2019. [Online; accessed 27-August-2020]. 7

13. DAILEY, M. N., JOYCE, C., LYONS, M. J., KAMACHI, M., ISHI, H., GYOBA, J., AND COTTRELL, G. W. Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion 10*, 6 (2010), 874. 6, 15

14. DAILEY, M. N., LYONS, M. J., KAMACHI, M., ISHI, H., GYOBA, J., AND COTTRELL, G. W. Cultural differences in facial expression classification. In *Cognitive Neuroscience Society, 9th Annual Meeting* (2002), vol. 153. 6, 15

15. DARWIN, C. *The Expression of the Emotions in Man and Animals.* John Murray, 1872. 3, 10

16. DENNETT, D. C. *The Intentional Stance*. MIT Press, 1987. 13

17. EKMAN, P. Pictures of Facial Affect. *Consulting Psychologists Press* (1976). 3

18. EKMAN, P. An argument for basic emotions. *Cognition & Emotion 6*, 3-4 (1992), 169–200. 3

19. EKMAN, P., AND FRIESEN, W. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press* (1978). 3

20. EL KALIOUBY, R. *Girl Decoded: My Quest to Make Technology Emotionally Intelligent–and Change the Way We Interact Forever*. Penguin UK, 2020. 13, 14

21. EL KALIOUBY, R. A. *Mind-reading Machines: Automated Inference of Complex Mental States*. PhD thesis, Cambridge University, 2005. 13

22. FLORIDI, L. Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology 32*, 2 (2019), 185–193. 16

23. JENSENIUS, A. R., AND LYONS, M. J. *A NIME Reader: Fifteen Years of New Interfaces for Musical Expression*. Springer, 2017. 15

24. KRYSA, J., AND MOSCOSO, M., Eds. *Stages #9: The Next Biennial Should Be Curated By A Machine*. Liverpool Biennial in partnership with DATA browser series., 2021. 2

25. LEWIN, R. *Bones of Contention: Controversies in the Search for Human Origins*. University of Chicago Press, 1997. 15

26. LEWIS, T. AI can read your emotions. should it? *The Guardian* (2019). 13

27. LYONS, M., AKAMATSU, S., KAMACHI, M., AND GYOBA, J. Coding facial expressions with Gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* (1998), pp. 200–205. 2, 6, 8, 9, 10

28. LYONS, M., KAMACHI, M., AND GYOBA, J. Gabor wavelet representation of facial expression. *Technical Report of IEICE 97*, 117 (1997), 9–15. 2, 6, 8

29. LYONS, M., KAMACHI, M., TRAN, P., AND GYOBA, J. V1 similarity measure recovers dimensions of facial expression perception. *Investigative Ophthalmology & Visual Science 38*, 4 (1997). 2, 3, 6, 8

30. LYONS, M., AND MORIKAWA, K. A model based on V1 cell responses predicts human perception of facial similarity. *Investigative Ophthalmology and Visual Science 37*, 910 (1996). 2, 3

31. LYONS, M., AND MORIKAWA, K. Gabor-based coding and facial similarity perception. *Perception 26*, 1_suppl (1997), 250. 2

32. LYONS, M. J. Facial expression recognition: From perception to application. *NIPS'2000 Workshop on Affective Computing* (2000). 14

33. LYONS, M. J. Facial gesture interfaces for expression and communication. In *IEEE International Conference on Systems, Man and Cybernetics* (2004), vol. 1, pp. 598–603. 15

34. LYONS, M. J. Dimensional affect and expression in natural and mediated interaction. *arXiv:1707.09599* (2007). 7, 13

35. LYONS, M. J. A mimetic strategy to engage voluntary physical activity in interactive entertainment. *IEEE FG'08 Workshop on Facial and Bodily Expressions for Control and Adaptation of Games* (2008). 13

36. LYONS, M. J. Excavating "Excavating AI": The elephant in the gallery. *arXiv:2009.01215* (2020). 2, 10, 12, 16

37. LYONS, M. J., AND BUDYNEK, J. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 21*, 12 (1999), 1357–1362. 7

38. LYONS, M. J., CAMPBELL, R., PLANTE, A., COLEMAN, M., AND KAMACHI, M. The Noh mask effect: Vertical viewpoint dependence of facial expression perception. *Proceedings of the Royal Society of London. Series B: Biological Sciences 267*, 1459 (2000), 2239–2245. 14

39. LYONS, M. J., KAMACHI, M., AND GYOBA, J. Coding facial expressions with Gabor wavelets (IVC special issue). *arXiv:2009.05938* (1998). 2, 3, 4, 6, 8, 10, 11, 14

40. LYONS, M. J., AND MORIKAWA, K. A linked aggregate code for processing faces. *Pragmatics & Cognition 8*, 1 (2000), 63–81. 2, 3

41. LYONS, M. J., AND TETSUDANI, N. Facing the music: a facial action controlled musical interface. In *CHI'01 Extended Abstracts on Human Factors in Computing Systems* (2001), pp. 309–310. 15

42. MATSUMOTO, D. Japanese and Caucasian facial expressions of emotion (JACFEE) and neutral faces (JACNeuF). *Intercultural and Emotion Research Laboratory, Department of Psychology* (1988). 3

43. MULLER, A. C. trans. The Analects of Confucius. http://www.acmuller.net/con-dao/analects.html, 2012.

44. MURRAY-RUST, P. Open data in science. *Nature Precedings* (2008), 1–1. 7

45. POUPYREV, I., LYONS, M. J., FELS, S., AND BLAINE, T. New Interfaces for Musical Expression. In *CHI'01 Extended Abstracts on Human Factors in Computing Systems* (2001), pp. 491–492. 15

46. PREMACK, D., AND WOODRUFF, G. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences 1*, 4 (1978), 515–526. 13

47. RUSSELL, J. A. A circumplex model of affect. *Journal of Personality and Social Psychology 39*, 6 (1980), 1161. 4

48. RUSSELL, J. A., SUZUKI, N., AND ISHIDA, N. Canadian, Greek, and Japanese freely produced emotion labels for facial expressions. *Motivation and Emotion 17*, 4 (1993), 337–351. 5

49. SALTZ, E., COLEMAN, L., AND LEIBOWICZ, C. *Making AI Art Responsibly: A Field Guide.* Partnership on AI, 2020. 16

50. SCHLOSBERG, H. The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology 44*, 4 (1952), 229. 4

51. SCHUESSLER, J. The social-cue reader. *The New York Times, December 10* (2006). 13

52. SOLOMON, B. Apple buys startup that can read your face to tell if you're angry. *Forbes* (2016). 13

53. SRINIVASAN, R., AND MARTINEZ, A. M. Cross-cultural and cultural-specific production and perception of facial expressions of emotion in the wild. *IEEE Transactions on Affective Computing* (2018). 3

54. TEDONE, G. From spectacle to extraction. and all over again. *unthinking photography* (2020). [Online; accessed 27-August-2020]. 12

55. VOSOUGHI, S., ROY, D., AND ARAL, S. The spread of true and false news online. *Science 359*, 6380 (2018), 1146–1151. 12

56. WALTON, D. *Informal Logic: A Pragmatic Approach.* Cambridge University Press, 2008. 2, 11

57. WRIGHT, J. Suspect AI: Vibraimage, emotion recognition technology and algorithmic opacity. *Science, Technology and Society* (2021). 17

58. ZHANG, Z., LYONS, M., AND SCHUSTER, M. Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Proceedings Third IEEE International Conference on Automatic face and gesture recognition* (1998), pp. 454–459. 6

59. ZUBOFF, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* Profile Books, 2019. 14