



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Seeing Affect

Knowledge Infrastructures
in Facial Expression Recognition Systems

Benedetta Catanzariti

PhD in Science, Technology and Innovation Studies
The University of Edinburgh
2023



Declaration

I declare that this thesis was solely undertaken by myself and it has not been submitted for any other degree or professional qualification. Except where otherwise acknowledged through references and citations, this work is my own in its entirety. Parts of this work have been published in the edited volume *Data Justice & The Right To The City* (Edinburgh University Press).

For Maria Adelaide and Federico Catanzariti

Abstract

Efforts to process and simulate human affect have come to occupy a prominent role in Human-Computer Interaction as well as developments in machine learning systems. Affective computing applications promise to decode human affective experience and provide objective insights into users' affective behaviors, ranging from frustration and boredom to states of clinical relevance such as depression and anxiety. While these projects are often grounded in psychological theories that have been contested both within scholarly and public domains, practitioners have remained largely agnostic to this debate, focusing instead on the development of either applicable technical systems or advancements of the field's state of the art. I take this controversy as an entry point to investigate the tensions related to the classification of affective behaviors and how practitioners validate these classification choices.

This work offers an empirical examination of the discursive and material repertoires – the infrastructures of knowledge – that affective computing practitioners mobilize to legitimize and validate their practice. I build on feminist studies of science and technology to interrogate and challenge the claims of objectivity on which affective computing applications rest. By looking at research practices and commercial developments of Facial Expression Recognition (FER) systems, the findings unpack the interplay of knowledge, vision, and power underpinning the development of machine learning applications of affective computing.

The thesis begins with an analysis of historical efforts to quantify affective behaviors and how these are reflected in modern affective computing practice. Here, three main themes emerge that will guide and orient the empirical findings: 1) the role that framings of science and scientific practice play in constructing affective behaviors as “objective” scientific facts, 2) the role of human interpretation and mediation required to make sense of affective data, and 3) the prescriptive and performative

dimensions of these quantification efforts. This analysis forms the historical backdrop for the empirical core of the thesis: semi-structured interviews with affective computing practitioners across the academic and industry sectors, including the data annotators labelling the models' training datasets.

My findings reveal the discursive and material strategies that participants adopt to validate affective classification, including forms of boundary work to establish credibility as well as the local and contingent work of human interpretation and standardization involved in the process of making sense of affective data. Here, I show how, despite their professed agnosticism, practitioners must make normative choices in order to 'see' (and teach machines how to see) affect. I apply the notion of *knowledge infrastructures* to conceptualize the scaffolding of data practices, norms and routines, psychological theories, and historical and epistemological assumptions that shape practitioners' vision and inform FER design.

Finally, I return to the problem of agnosticism and its socio-ethical relevance to the broader field of machine learning. Here, I argue that agnosticism can make it difficult to locate the technology's historical and epistemological lineages and, therefore, obscure accountability. I conclude by arguing that both policy and practice would benefit from a nuanced examination of the plurality of visions and forms of knowledge involved in the automation of affect.

Lay Summary

Over the past decade, we have witnessed a rise in the research and development of technologies designed to recognize, simulate, and respond to human emotions based on affective behaviors such as changes in facial expressions, speech intonation, or heart rate. Practitioners envisage these tools as a way to improve interactions between humans and computers and to support decision making processes in a variety of sectors, including healthcare, education, human resources, in-car safety, and marketing and advertising. However, the idea that computers can recognize human emotions based on affective behaviors is premised on the assumption that we can establish a meaningful and reliable correlation between physical expressions and mental and emotional states. This assumption has been widely contested but has, nonetheless, greatly influenced the development of affective technologies. In fact, many practitioners have remained largely peripheral to this debate and continued to focus on technical developments without engaging critically with this underlying assumption.

In this thesis, I investigate the research and development of Facial Expression Recognition (FER) systems, that is, machine learning systems designed to recognize changes in facial expressive behaviors and produce some form of meaningful knowledge about such expressions. I begin by tracing historical efforts to measure and quantify affect and the social and cultural assumptions underlying these quantification projects. I then analyze data from interviews with 22 practitioners working on the development of FER systems across the academic and industry sector. Findings from this historical as well as empirical investigation illustrate the various ways in which practitioners justify and validate their classification choices and the norms and assumptions that underpin the development of these systems. This analysis shows that, despite efforts to frame FER systems as neutral and objective tools, practitioners risk

reinforcing assumptions about affective behaviors that, over time, can influence and shape how we understand and express affect.

Acknowledgements

There would be no words on these pages but for the community of friends, colleagues, and mentors who supported and inspired me throughout my PhD journey. First, however, I wish to thank my research participants: the 22 practitioners who took time out of busy and sometimes unpredictable schedules to share their perspective and experience with me. This thesis would have not been possible without their contribution.

I will be forever indebted to my supervisors, Morgan Currie, Ewa Luger, and Robin Williams, for uplifting this research from its early days and for setting a positive example of how to be a great scholar: kind, caring, inspiring and, most importantly, joyous. I could have never asked for better mentors. Thank you for your patience and unwavering support.

To my friend and research partner, SJ Bennett, I owe more than words can describe. It is our friendship and collaboration that helped me through some of the toughest times of this PhD. I also wish to thank my AI Ethics and Society comrades, Yazmin Morlet Corti, Vassilis Galanos, Lara Dal Molin, and Aditi Surana. I am very grateful for being able to surround myself with such talented and compassionate humans.

I owe my gratitude to many members of the Science, Technology and Innovation Studies (STIS) community at the University of Edinburgh: during my MScRes, Fadhila Mazanderani helped me navigate numerous and sometimes kafkaesque funding applications; I will be forever grateful for her patience. Michael Barany has provided valuable advice and critique throughout my PhD, starting from my first year board review (for which he was the internal examiner) to post-doctoral job applications. Similarly, Rob Smith and Matjaz Vidmar have inspired me and helped me imagine and shape my life trajectory beyond the PhD. I also wish to thank James Stewart for making

life in the academia light and fun. Thank you to my extended PhD cohort for priceless chats, writing sessions, drinks, and laughs: James Besse, Jarmo De Vries, Mayline Strouk, Xiao Yang, Hadewych Honné, Sophie Stone, Rachel Simpson, Mari Serine Kannelønning, and Asli Ates. I also wish to thank my pre-pandemic office mates for always keeping the morale high: Louis Volkmer, Antonio Ballesteros, Rob Nayan, and Katie Howell.

Outside of STIS, I am indebted to many more friends, colleagues, and teachers: Karen Gregory (Digital Sociology) examined my first-year board paper as an external reviewer and has been an inspiring force ever since. I was lucky to work with Beverley Hood (Edinburgh College of Art) on a creative project that combined research, theatre, and artificial intelligence. On days where writing felt too cumbersome, I looked forward to our happy experimentations. I am also thankful to James Garforth (School of Informatics) for offering me his friendship and for welcoming me into the Informatics teaching team. Working closely with computer science and engineering students has taught me more than I could hope for. I also wish to thank Ryan Bowler (Design Informatics) for sharing hopes and fears, as well as marking frustrations. Thanks to Jamie Webb (Centre for Biomedicine, Self and Society/Centre for Technomoral Futures) for offering valuable comments on chapter 7 of this thesis.

Despite a global pandemic, I was lucky to meet many excellent scholars around the world and be part of broader research networks – both virtually and in person. Many thanks to Oliver Kayas (Liverpool John Moores), who I met through the Surveillance Studies Network Mentorship Programme. Even when the programme ended, Oliver continued to offer valuable career advice and support. In addition, I wish to thank Alex Taylor (City, University of London) for inviting me to be part of the Data Labelling Working Group. Here, I had the opportunity to discuss parts of chapter 6 of this thesis in front of many brilliant researchers. I also had the opportunity to present parts of chapter 5 at the AI and the Digitalized Society Workshop, organized by the STS

department of the University of Helsinki and the Finnish Center for Artificial Intelligence, as well as at the panel on Inference Worlds at the 2022 European Association for the Study of Science and Technology (EASST) conference in Madrid, organized by Louise Amoore, Alex Campolo (Durham University), and Nanna Bonde Thylstrup (Copenhagen Business School). I am grateful for all the feedback and critique I received in these occasions, as well as the friendships and professional relationships forged throughout.

I wish to acknowledge the care work of the non-academic staff that has made this research materially possible: Margaret Acton, Geraldine Debard, Toni Jenkins, Cathy Naughton, and Valentina Guerrieri.

Outside of academic spaces, I owe a debt of gratitude to all the friends who always rooted for me: Dena, Silvia, Robbie, Anna, Andrea, Valentina, Giovanni, Noemi, and Roberto. Thank you for always being my allies.

To my family – my mom and dad, Raffaele and Gilda, and my brother Alessandro: even though we now speak largely different languages (literally and, perhaps, metaphorically), our life trajectories are somehow similar. We all left something behind and reached, although painfully, for the unknown. I hope you are proud.

Finally, to Alec: thank you for reminding me every day of what really matters.

Table of Contents

Declaration	ii
Abstract.....	iv
Lay Summary	vi
Acknowledgements.....	viii
Table of Contents	xi
List of Figures	xiv
List of Abbreviations	xv
Introduction	1
“Determining indeterminacy”	1
Agnostic visions and the politics of affective classification	4
Situating algorithmic knowledge: Research questions and overview.....	6
Thesis outline.....	7
Chapter 2 On theory and methods	11
Introduction	11
Theoretical position: Knowledge, vision, and power	11
A view from “somewhere”	13
Knowledge infrastructures.....	15
Methodology and methods	17
Document analysis	18
Semi-structured interviews.....	19
Sampling and recruitment Strategy.....	21
Participant characteristics	24
Interview procedure	30
Data analysis: Interpreting interpretations	30
Research ethics and the politics of anonymity	32
My position as a researcher: Doing STS in Computer Science spaces.....	33
Chapter 3 Performing theories.....	37
Introduction	37

Reading faces: A tale as old as (D)ARPA.....	41
Anatomy of a lie: the Facial Action Coding System	45
Automating visions: FACS meets neural networks	51
“We don’t do mind reading”	55
Conclusion.....	62
Chapter 4 The making of affect	63
Introduction	63
Affects in the laboratory	64
Classification breeds performance	66
Cybernetic feelings: between feedback and control.....	70
Data-driven emotions?	74
Conclusion.....	77
Chapter 5 Mobilizing science	79
Introduction	79
Boundary work in affective computing practice	80
Expulsion: Truthful science and ill intents.....	82
Autonomy: Agnosticism and ambiguity.....	88
Applied science	88
Basic research.....	91
Expansion: The magpie effect.....	95
Sidestepping ground truths	100
Conclusion.....	104
Chapter 6 Seeing affect	106
Introduction	106
Skill, expertise, and power in data structures.....	109
Contested visions.....	113
Expert coders: The Facial Action Coding System	114
Lay annotation: Dimensional models of affect	119
Automating visions: The NOVA interface	127
“It sounds like magic”: Representations of data work matter	131
Seeing data work, accepting uncertainty	137

Conclusion.....	140
Chapter 7: The promise of Affective Computing.....	142
Introduction	142
Know Thyself	145
The search for objectivity	148
“It ticks all the boxes”	155
Normative visions of health and gender	159
What kind of care?.....	164
Conclusion.....	167
Conclusion.....	169
Seeing affect within knowledge infrastructures.....	169
Contributions to knowledge and policy	173
Limitations, future research, and “good” technology	176
“I prefer not to”: From ignorance to uncertainty	180
Final remarks.....	186
Bibliography	187
Annex A: Consent form	209
Annex B: Participant information sheet	211
Annex C: Interview guide	214

List of Figures

Figure 1. Photographs of the six basic emotional expressions used in Ekman's cross-cultural study (reprinted from Ekman 1971)	43
Figure 2. FAST facial areas for the emotion ‘surprise’ (reprinted from Ekman, Friesen, and Tomkins 1971)	48
Figure 3. Expression of “gustative pain” in a Black person, from Paolo Mantegazza’s Atlas of Pain (reprinted from Moruno 2016).	67
Figure 4. Duchenne de Boulogne measuring expressions of surprise on a Salpêtrière patient (W. Bruce and Delaney H. Lundberg Fund)	69
Figure 5. The VAD model mapped on the six basic emotions (reprinted from Mitruț et al. 2019).....	122
Figure 6. Discrete and continuous annotation schemes supported by the NOVA interface (reprinted from Heimerl et al. 2019).....	128
Figure 7. LIME explanations for four emotion scores (reprinted from Heimerl et al. 2019).	130
Figure 8 Electrical activity of four facial regions detected by EMG sensors, as shown in (Schwartz et al. 1976).....	150

List of Abbreviations

16PF – 16 Personality Factor

ACII – Affective Computing and Intelligent Interaction

ACM – Association for Computing Machinery

ADHD – Attention-Deficit/Hyperactivity Disorder

AI – Artificial Intelligence

AIES – AI Ethics and Society

API – Application Programming Interface

ARPA – Advanced Research Projects Agency

ASC – Autism Spectre Condition

AU – Action Unit

AVEC – Audio/Visual Emotion Challenge

BET – Basic Emotion Theory

BIPOC – Black, Indigenous, and People of Color

CEO – Chief Executive Officer

CHI – Conference on Human Factors in Computing Systems

CIA – Central Intelligence Agency

COO – Chief Operating Officer

CSCW – Computer Supported Cooperative Work

CSO – Chief Scientific Officer

CTO – Chief Technology Officer

DARPA – Defense Advanced Research Projects Agency

DSM – Diagnostic and Statistical Manual

ECG – Electrocardiogram

EDA – Electrodermal Activity

EEG – Electroencephalogram

EMFACS – Emotion Facial Action Coding System

EMG – Electromyography

FAccT – Fairness, Accountability, and Transparency studies

FACS – Facial Action Coding System

FACSAID – Facial Action Coding System Affect Interpretation Dictionary

FAST – Facial Action Scoring Technique

FBI – Federal Bureau of Investigation

FER – Facial Expression Recognition

HCI – Human Computer Interaction

HPD – Histrionic Personality Disorder

ICO – Information Commissioner’s Office

IEEE – Institute of Electrical and Electronics Engineers

LIME – Local Interpretable Model-Agnostic Explanations

MIT – Massachusetts Institute of Technology

NIMH – National Institute of Mental Health

NSF – National Science Foundation

R&D – Research and Development

RU-FACS – Rochester/UCSD Facial Action Coding System Database

SCAN – Systematic Classification and Analysis of Nonverbal Behavior

SSK – Sociology of Scientific Knowledge

STEM – Science, Technology, Engineering and Mathematics

STS – Science and Technology Studies

TAC – Transactions on Affective Computing

UCSD – University of California San Diego

UCSF – University of California San Francisco

VAD – Valence Arousal Dominance

VID-R – Visual Information Display and Retrieval

VR – Virtual Reality

Introduction

“Computation is a process of determining indeterminacy. This indeterminacy, however, is internal to the process itself, and a condition for computation’s ontological production.” (Fazi 2018).

“Determining indeterminacy”

It is the height of the Covid-19 pandemic, and I am battling with an unsteady Internet connection while Pierre, a cognitive neuroscientist based at a British university, sits comfortably on the other side of my grainy screen. He is wearing a headset microphone that makes his voice sound distant and metallic: “I use things like heart rate monitoring and skin conductance, pupil dilation.. I’ve always done a lot of eye tracking... looking at where people’s attention is, and then linking all that information together”.¹ Pierre studies nonverbal behaviors – or, as he calls it, “paralinguistic communication” – and uses the data generated by this observed physiological activity to improve people’s interactions with computer systems and virtual agents. Typically, his experiments involve three or four different biosensors: an ECG, or Electrocardiogram, to monitor changes in the heart rate; an EDA, or Electrodermal activity sensor, which measures variations in the skin’s electrical activity; and a video camera tracking the participant’s gaze as well as their changes in facial expressive behaviors. Pierre cautions me that these sensors should never be used in isolation, and that a robust study of affective behaviors demands a combination of multiple data streams. However, as he admits, all

¹ Interview conducted on November 4th, 2020. All the following quotes are from the same interview.

but the latter require that his subjects be physically present in his lab – a space he has no longer access to:

A lot of these other sensors involve a person sometimes literally being wired up, things attached to them, or devices put on them or in front of them.. I haven't been able to do any of that since February. So everything had to shift online, but not everything can be done online. So, the facial analysis stuff... It essentially just needs a webcam.

Throughout 2020, Pierre and his team have recruited participants online to measure their emotional responses or, as he put it, “how they biologically sort of respond in terms of facial expressions” when presented with a range of stimuli. At the moment, the researchers are testing a navigation system that recognizes users’ facial expressions to assess their levels of confusion and frustration and adapt the routing instructions in real time. To collect this data, Pierre has purchased a license for the use of the iMotions Lab, a behavior analysis platform that integrates multiple sensors from over 20 different independent vendors.² Among these, is a facial expression analysis module developed by Affectiva, a Boston-based AI company born in 2009 as a corporate spin-off of the MIT Media Lab and co-founded by former student Rana el Kaliouby and professor Rosalind Picard, the scientist responsible for inaugurating the field of “affective computing” in the 1990s. Affective computing is an interdisciplinary research area combining computer science, engineering, and psychology and dedicated to the development of computer systems trained to recognize and simulate human affective behaviors or, as declared by Affectiva’s mission statement, “humanize technology to bridge the gap between human and machines.”³ Central to affective computing’s project is in fact the recognition that emotions play a fundamental role in human

² <https://imotions.com/>.

³ <https://www.affectiva.com/>.

perception and, therefore, equipping computers with the ability to understand and simulate emotions would contribute to the development of a more sophisticated form of artificial intelligence.

To recognize expressive behaviors, Affectiva’s algorithm identifies what are called “key landmarks” on the face – such as the tip of the nose, the corners of the lips, or the arches of the eyebrows – and uses these coordinates to classify emotional expressions according to a rather narrow emotional taxonomy: as stated in iMotions’ guide to facial analysis, the data gathered by Affectiva’s algorithm “gives rise to probability values that show the likelihood of one of the 7 basic emotions being exhibited: anger, contempt, disgust, fear, joy, sadness or surprise”.⁴ Like most contemporary producers of “emotional AI” technologies (McStay 2018), Affectiva’s approach to affective behaviors is largely grounded on the “classical view of emotions” (Barrett 2017), which posits the existence of discrete (or “basic”), biologically-determined expressions of emotions that can be measured objectively and classified through computational techniques. As I illustrate in this thesis, this view of affective behaviors was popularized in the twentieth century by psychologist Paul Ekman and, despite the many critiques, has never been fully abandoned. Further, this conceptual model of emotional behaviors has directly shaped early attempts to train computer vision systems to recognize facial expressions. This approach has proven successful not simply because, as scholar Andy McStay noted “this account of emotional life works well with sensing techniques that classify facial and bodily behaviour” (McStay 2018: 5); rather, it promises a standardized method for the quantification of an object that seems to exceed projects of rationalization: affect. As reflected by this introduction’s opening quote, this constant effort to “determine indeterminacy” (Fazi 2018) is typical

⁴ <https://imotions.com/products/imotions-lab/modules/fea-facial-expression-analysis/>.

of algorithmic modes of knowledge which attempt to “bring the formal abstractions of computation into contact with the indeterminacy of material existence” (6).

It is this tension between the indeterminacy of affective experience and the promise of rationalization offered by algorithmic classification that motivates my work in this thesis. Beginning this study, I set out to understand what informs the design of Facial Expression Recognition (FER) systems, that is, computational models trained to recognize facial expressive behaviors and produce some form of valuable knowledge about what makes people frown, grimace, flinch, or smile. By trying to de-code and re-code human affects, what “figures of the human” (L. Suchman 2006), affect, and the mind do practitioners leverage? And, further, what can the search for an objective classification of affective behaviors tell us about the wider field and culture of computer science and machine learning?

Agnostic visions and the politics of affective classification

Critical accounts of affective technologies have placed great emphasis on the lack of scientific consensus around computational approaches that postulate the existence of measurable, discrete, and universal emotional expressions (Stark and Hoey 2020; Stark and Hutson 2021; Crawford et al. 2019). As many have noted, this narrow understanding of affective experience does not capture the variety of ways in which people express emotions across different cultural and social contexts (Barrett et al. 2019; Crivelli et al. 2016). Further, this model relies on a form of inference defined as “reverse appraisal” (Ong, Zaki, and Goodman 2019), that is, the assumption that we can infer a person’s internal states from their observed facial behaviors. Importantly, scholars and advocacy groups have noted how the automation of this inference – the “causal leap” (McStay and Urquhart 2019) from detecting physical behaviors to making claims about emotional states – risks perpetuating pseudoscientific and harmful assumptions about facial features and social outcomes (Stark and Hutson 2021; Article

19 2021; Crawford et al. 2019). These claims, researchers have argued, echo the 19th century practice of physiognomy, a discredited belief that a person’s facial appearance could reveal something about their moral character (Stark and Hutson 2021; Arcas 2017; Birhane 2021a).

However, as discussed throughout this thesis, practitioners have remained largely peripheral to these critiques and focused their attention instead on either advancements of the technical state of the art or the development of applicable FER systems (R. A. Calvo and D’Mello 2010). In this thesis, I argue that this form of agnosticism – which is not unique to affective computing but represents a rather common feature of computer science – risks obscuring accountability in that it conceals the historical and epistemological assumptions on which much of the development of FER systems rests. In this sense, practitioners’ disconnect from these debates can serve as a strategy to avoid responsibility around the ethics and politics of affective classification, that is, the role that the classification of expressive behaviors can play towards the social regulation of affective experience. As I explore in this thesis, scholars have noted how, historically, scientific efforts to measure and display affective behaviors served to reinforce norms of emotional expression that posited white Europeans as superior to those of people of color and the mentally ill. Historians Peter and Carol Stearns have defined the complex of social and cultural rules regulating affective behaviors as “emotionology”, or “the attitudes or standards that a society, or a definable group within a society, maintains toward basic emotions and their appropriate expression; ways that institutions reflect and encourage these attitudes in human conduct” (Stearns and Stearns 1985: 813). In this sense, the attribution of ‘positive’ or ‘negative’ emotions to particular social groups – such as the “innocence” of White women’s tears as opposed to the “anger” of Black women (Bonilla-Silva 2019) – is central to projects of social and racial classification. Building on these considerations, media scholars and critical data studies researchers have pointed to the

prescriptive and normalizing power of affective technologies, that is, the way in which the automated classification of emotional expressions can shape our affective experience: in search for ‘objective’ markers of emotional and mental states, these systems can instead reinforce normative assumptions about expressive behaviors, gender, race, and mental health (Stark 2020b; Hoffmann and Stark 2015). In this thesis, I situate FER systems within these cultural and social dimensions and offer ways to challenge practitioners’ agnosticism. In particular, I apply the concept of *knowledge infrastructures* to foreground the complex of knowledges, norms, and assumptions that inform and support FER projects. Once these discursive and cultural repertoires are made visible, it is perhaps possible to open up space for accountability.

Situating algorithmic knowledge: Research questions and overview

As mentioned above, I took the controversy over the validity of FER projects as an entry point to investigate the tensions related to the classification of affective behaviors. This concern translated into two broad research questions:

- 1) *Where and how does classification take place in FER research and practice?*
- 2) *How do practitioners validate these classification choices?*

Other questions that allow me to explore classification practices in more detail include: What counts as measurable affective behavior and how do practitioners measure it? What ideas about affect, the mind, and behaviors are encoded into and reproduced by these classification choices? Further, to understand the role that disciplinary and professional boundaries play in the development of FER systems, I asked: What forms of knowledge are involved in FER research and practice? And what norms and routines guide the ways practitioners make sense of affective data?

As detailed in Chapter 2 of this thesis, I address these questions by adopting a qualitative research approach that combines historical inquiry with the empirical

investigation of local and contingent practices of practitioners' sense-making. Emerging from this longitudinal study are three broad themes that constitute, respectively, the focus of the three empirical chapters at the core of this thesis. In particular, the investigation of both historical practices of quantification of affect and computational projects of affective classification reveal: 1) how science gets mobilized to construe affective behaviors as objective and scientific facts; 2) that central to affective classification efforts (the ways practitioners 'see' affect) is a work of interpretation and standardization of affective data; and 3) that quantifications and categorizations of emotional expressions have a prescriptive and normative effect that can contribute to the social regulation of expressive behaviors. Each chapter brings one or more of these dimensions to the fore, situating the 'algorithmic objectivity' of FER projects within specific historical and cultural imaginaries – the infrastructures of knowledge – that support and inform their development. Finally, this thesis is not about 'affect' per se. I do not intend to offer a definition of affect nor an alternative understanding of affective experience. Rather, I am concerned with how practitioners 'see' affect and the norms and practices that shape this vision. By looking at these local norms of interpretation, I seek to show the fundamental uncertainty – what Beatrice Fazi calls “contingent computation” (Fazi 2018) – on which FER projects are grounded.

Thesis outline

The next chapter of this thesis, *On theory and methods*, begins by discussing the theoretical and methodological frameworks that informed the research, and introduces the concept of *knowledge infrastructures*. Here, I argue that the domain independence, abstraction, and agnosticism of computer science are reminiscent of what Donna Haraway called the “god trick” (Haraway 1988), a form of knowledge that presumes algorithmic results to be objective, neutral, and universal. Drawing on feminist studies

of science and technology, I claim instead a form of “situated objectivity” (Ibid.; Vertesi 2020), that is, a recognition of the partial and local conditions of techno-scientific knowledge. I offer the concept of knowledge infrastructures as a way to make system design accountable and challenge claims of universality and objectivity by situating the knowledges, norms, and assumptions that inform and support the development of FER projects. The second section of the chapter provides a detailed discussion of my methods (document analysis and qualitative interviews), along with additional information on sampling, data analysis, research ethics, as well as a personal reflection on my role as a STS scholar intervening within the research field.

Chapter 3, *Performing theories*, traces the genealogy of automated systems for the recognition and classification of affective behaviors, starting from Paul Ekman’s 1960s studies on deception and culminating in the Facial Action Coding System (FACS), a coding framework for the analysis of affective behaviors that directly shaped early developments in computer vision. This chapter highlights the role of US intelligence and law enforcement in influencing technical advancements of FER systems including models’ assumptions (the idea that facial expressions can reveal ‘truth’ or ‘deception’) and data collection practices (the elicitation, collection, and annotation of ‘genuine’ and ‘deceptive’ facial expressions). I argue that practitioners’ efforts to re-frame FER systems as benign, and perhaps even humanitarian when applied, for instance, to healthcare contexts, risks obscuring this military legacy.

Chapter 4, *The making of affect*, places Ekman’s work within the longer history of the quantification of the affects, beginning in the late 19th century with the construction of affect as an object of laboratory knowledge. Here, I show how scientific efforts to quantify and display emotions helped reinforce and normalize assumptions and expectations towards affective behaviors aligned with colonial norms that posited emotional expressions of white Europeans at the peak of the civilization process. Importantly, this chapter foregrounds the role of interpretation and standardization

practices necessary to construe affective behaviors as objects of scientific knowledge and how these practices are reflected in computational projects of affective classification.

Chapter 5, *Mobilizing science*, constitutes the first of the three empirical chapters and focuses on participants' construction of objectivity and validity through means of discursive and rhetorical practices. As mentioned in this introduction, despite the public and scholarly controversy over efforts to classify expressive behaviors, many practitioners maintain an agnostic view with regards to the conceptual models of emotional expressions they adopt to develop FER systems. However, to classify affective expressions, and infer meaningful information from this classification, practitioners must make some fundamental assumption about what constitutes a valid and measurable affective expression, and how to extrapolate valuable knowledge from it. The chapter takes this paradox as a departing point to explore the various forms of boundary work through which participants establish and maintain epistemic authority over their claims.

Chapter 6, *Seeing affect*, takes a closer look at the data work that underlies FER projects. In particular, I focus on two annotation schemes used to label (‘see’) training data for FER algorithms: the Facial Action Coding Scheme (FACS) and the Valence, Arousal, Dominance (VAD) model, regarded respectively as forms of “expert” and “lay” or “unskilled” annotation. Here, I center the experience of two participants, Isabella (a FACS coder) and Olivia (a VAD annotator), and show how the definition and attribution of skill and expertise in data structures reflects broader power asymmetries in machine learning design that are further reinforced through practitioners' perception and representation of data work. Importantly, by adopting analytical tools borrowed from anthropological studies of professional vision (Goodwin 1994; Grasseni 2007), I show how both coding schemes involve local and contingent

practices of negotiation and interpretation of the data, thus unearthing the uncertainty that lies at the core of affective classification.

Chapter 7, *The promise of affective computing*, shifts to *healthcare* applications of FER systems, that is, computational models that measure changes in facial affective behaviors to diagnose and manage mental health conditions (such as depression and anxiety) and neurodevelopmental disorders (such as autism or bipolar disorder). Here, I explore the social and cultural imaginaries that inform the development of these tools and show how the knowledge produced through techno-scientific discourse and practice can shape and determine bodies and behaviors. In particular, I show how normative ideas of health and illness that are at play in the development of FER systems can facilitate and reinforce the expression of certain behaviors aligned with socio-cultural expectations around, for instance, gender and health. Taken together, chapter 5 to 7, illustrate the infrastructures of knowledge on which practitioners draw to classify and make sense of affective behaviors. In particular, these chapters show how, despite their professed agnosticism, participants ways of ‘seeing’ affect are shaped by a complex of norms, practices, and normative assumptions about affect, identity, the mind, health, and illness.

In my concluding chapter, I bring together these findings and discuss my contributions to knowledge and policy. Here, I also reflect on the limitations of this work and suggest avenues for future research. Finally, I return to the issues of agnosticism and uncertainty with which this thesis started. I first compare participants’ agnosticism to what Robert Proctor and Londa Schiebinger called *agnogenesis*, that is, the construction of ignorance within techno-scientific contexts to avoid accountability (Proctor and Schiebinger 2008). I then advocate for the recognition of uncertainty as constitutive to the project of automation and the need to create space within design pipelines to collectively contest and challenge algorithmic certainty and objectivity.

Chapter 2 | On theory and methods

Introduction

In this chapter, I discuss the theoretical and methodological frameworks that informed this thesis, from conceiving of the research to the analysis of the data and writing. The first section of the chapter presents conceptual tools drawn from feminist studies of science and technology and histories of objectivity. This section also explores the concept of *knowledge infrastructures* and discusses its conceptual as well as methodological relevance. The second section of the chapter offers a detailed discussion of my methods: document analysis and qualitative interviews. Here, I reflect on the role that different interpretive frameworks played in my analysis of practitioners' own interpretation of affective information. Finally, the chapter concludes with a reflection on my role as a STS scholar intervening within computer science spaces.

Theoretical position: Knowledge, vision, and power

Scholars have studied how computer science can shape politics and society and, particularly, how predictive models can produce and reinforce certain views of the world over others. As Louise Amoore noted, algorithms can be understood as “ethicopolitical arrangement of values, assumptions, and propositions about the world” (Amoore 2020: 6), that demarcate what is to be considered normal from abnormal through practices of classification and categorization. These predictions are often grounded on claims of objectivity, neutrality, and – as an inevitable corollary of those first two properties – accuracy. Cloaked in the persuasive certainty of numbers and rules, statistical decision-making “has at least the appearance of being fair and

impersonal” (Porter 1996: 8). This form of scientific objectivity, argues Theodor Porter, “provides an answer to a moral demand for impartiality and fairness. Quantification is a way of making decisions without seeming to decide” (Ibid.). Objective, fair, and impersonal, numbers aim to set feelings aside: “the more neutral, the more objective; and the more objective, the more true—or so this line of reasoning goes” (D’Ignazio and Klein 2020: 6). Ironically, in the projects and practices discussed in this thesis, it is precisely the expression of feelings that is made object of aseptic quantification, stripped of the contexts that may have elicited it. Aiming for unbiased and dispassionate measurements of affective behaviors, the practitioners interviewed here mobilized a form of objectivity that relies on a claim of conceptual (and epistemological) agnosticism: a detachment from the complex of theories, ideas, and assumptions that make the measurement of affect possible in the first place. This is not unique to the technologies and practices described in this thesis; many have pointed out how abstraction and formalism – as defining features of computer science – can obscure practitioners’ accountability. For example, in discussing the role that abstraction plays in computer science education, James Malazita and Korryn Resetar have argued that, despite practitioners’ efforts to portray abstraction as “a collaborative knowledge framework that affords interdisciplinary activity, abstraction instead serves as an epistemic, cultural, and ideological wall to integrated critical-technical education” (Malazita and Resetar 2019: 301). Similarly, Stephen Slota et al. have noted how the “domain agnosticism” or “domain independence” of data science positions practitioners “in the middle of all things—capable of engaging this, that, or any domain” (Slota et al. 2020: 1).

This agnostic view from everywhere – and, at the same time, from nowhere – is reminiscent of what Donna Haraway called the “god trick”, a form of vision that, disguised as neutral and universal, conceals instead “the unmarked position of Man and White” (Haraway 1988: 581) and, we might add, able-bodied, cisgender, and human.

This becomes particularly relevant in computer vision applications, where this seemingly ‘objective’ gaze is central to automated regimes of recognition and classification. To attend to the ethics and politics of algorithms – to contest the god trick – is to follow and unpack claims of agnosticism and illuminate, layer by layer, the infrastructures of norms and assumptions on which algorithmic systems rely.

A view from “somewhere”

As feminist scholars of science and technology have long noted, a critique of objectivity is not a contestation of science altogether. As Haraway reminds us, a key principle of the ‘strong program’ in the Sociology of Scientific Knowledge (SSK) emerged in the 1970s was that symmetrical and impartial explanations should be applied to all sides of a scientific controversy, in an effort to avoid normative statements about scientific ‘truth’ and ‘falsehood’ (Bloor 1976). Here, Haraway joins other scholars who pointed out how this methodological posture risks lending itself to critiques of relativism:

The strong program in the sociology of knowledge joins with the lovely and nasty tools of semiology and deconstruction to insist on the rhetorical nature of truth, including scientific truth. History is a story Western culture buffs tell each other; science is a contestable text and a power field; the content is the form. Period. (Haraway 1988: 577)

A feminist critique of scientific objectivity, on the other hand, claims a “no-nonsense commitment to faithful accounts of a ‘real’ world” (579) that can enable ethical and political projects of solidarity and sustainability. To do so, argues Haraway, it is necessary to go beyond the traditional binaries of *biased vs unbiased science* or *science vs pseudoscience*. As shown in this thesis (and particularly in chapter 4 and 5), these binaries can be in fact mobilized – meaning that the boundaries of science can be shifted by means of discursive and material strategies – to privilege particular knowledge claims over others. By recognizing the “situated” quality of all knowledge

claims (Ibid.), feminist techno-science can instead offer “a strong tool for deconstructing the truth claims of hostile science by showing the radical historical specificity, and so contestability, of *every layer* of the onion of scientific and technological constructions” (578). Against both the abstract view from nowhere of positivist objectivity, as well as the toothless view from everywhere of relativism, the feminist form of objectivity points to the local, situated, and embodied condition of all forms of vision:

Relativism is the perfect mirror twin of totalization in the ideologies of objectivity; both deny the stakes in location, embodiment, and partial perspective; both make it impossible to see well. Relativism and totalization are both “god tricks” promising vision from everywhere and nowhere equally and fully, common myths in rhetorics surrounding Science. But it is precisely in the politics of epistemology and partial perspectives that the possibility of sustained, rational, objective inquiry rests (584).

Importantly, feminist objectivity recognizes what Haraway calls the “apparatus of bodily production”, that is, the performative power of scientific knowledge to produce and reproduce bodies and “other objects of value” (595). In this sense, bodies are not natural entities waiting to be discovered, but rather objects constructed through techno-scientific knowledge. Examples of this are the construction of sex and gender (Butler 1999; 2011) and mental illness (Michel Foucault 2013) through medical and scientific discourse. I bring these dimensions to the fore in chapter 7, where I explore the prescriptive effects of FER systems developed for diagnostic uses. In search for ‘objective’ markers of mental behaviors, participants reinforced instead normative assumptions about mental illness, behaviors, and gender. Informed by feminist studies, this research aims to recognize and show how techno-scientific knowledge – knowledge produced through technological and scientific discourse and practice – can shape and determine bodies and behaviors.

Knowledge infrastructures

This thesis offers ways to make system design accountable and challenge claims of universality and objectivity by foregrounding – or, as Haraway called it, “positioning” (Haraway 1988)– the knowledges, norms, and assumptions that inform and support FER projects. To do so, I attend to the ways practitioners ‘see’ affective information, that is, I pay attention to the complex of theories and practices that shape practitioners’ vision of affect. The three empirical chapters that constitute the core of this thesis map participants’ efforts to construe affective behaviors as “objective facts”: the various discursive and material strategies by which practitioners formulate and validate their claims of objectivity. However, a seemingly straightforward task – the measurement of facial expressions – requires significant work of human interpretation and negotiation, a “collaborative craftwork of hands, eyes, and signs” (Suchman and Trigg 1993: 173). This becomes more evident in chapter 6, where I discuss first-hand accounts of the data annotation practices necessary to label facial expression datasets and the role that coding frameworks, as well as organizational dynamics, play in shaping annotators’ visions.

I understand these ensembles of theories and practices as *knowledge infrastructures*. Despite the homonymity, the concept is not (or not entirely) inspired by critical infrastructures studies (P. Edwards 2002; P. N. Edwards et al. 2013; S. L. Star 1999; Bowker and Star 2000; Karasti et al. 2016) but, rather, by critical studies of technology at the intersection of STS and HCI.⁵ In this sense, I draw on Lucy Suchman’s notion of “configurations” as a way to pay attention to the “imaginaries and

⁵ Critical infrastructure studies argue that mundane aspects of infrastructures are imbued with politics. Common examples of infrastructures include railroads, plumbing, electricity, as well as information and communication infrastructures – all of these are “something that other things ‘run on’” (Star 2002: 116). Importantly, the political nature of an infrastructure lies on its ability to enable or, conversely, foreclose the circulation of goods, bodies, and information: “For the blind person, the graphics programming and standards for the World Wide Web are not helpful supporters of computer use, but barriers that must be worked around” (Ibid.).

materialities” that inform technology design and, continues Suchman, “an orientation that resonates as well with the term’s common usage to refer to the conjoining of diverse elements in practices of systems design and engineering” (Suchman 2012: 48). Configuration, in this sense, becomes an analytical lens to trace the “heterogeneous relations that technologies fold together” (ibid.). Similarly, I articulate the notion of knowledge infrastructures as a conceptual tool to recover the discursive and material repertoires that underpin practitioners’ claims. In an effort to make knowledge and power visible, I choose the term ‘infrastructure’ for its material quality.⁶ Similarly, throughout the thesis, I often adopt a spatial language, using terms such as *layers*, *structures*, *grids*, and *maps* to locate practices of sense-making and render them visible.

Finally, to foreground these norms and assumptions, I borrow conceptual tools from histories of scientific objectivity (Daston and Galison 2007) as well as the classification of affect (Dror 1999b; 2001; 2011). In so doing, I hope to bring to the fore how the long history of quantification of affective experience informs modern computational efforts to measure facial expressive behaviors. As Suchman argued, when designing socio-technical systems, “the politics of cultural historical imaginaries need to be on the table” (Suchman 2012: 52), and recognizing those politics “may require reanimating the figurations that hold particular relations of persons and things in place” (Ibid.). From a methodological perspective, tracing these imaginaries involves what Suchman, borrowing from John Law, called “method assemblage” (Law 2004), that is, a way to

make relations between what is *present* (including knowledges, representations, subjects and objects) and what is *absent* or part of the latter’s ‘hinterland’ (both *manifestly*, for example in the form of things articulated as ‘context’ for what is present, and *othered*, in

⁶ Susan Leigh Star argued that infrastructures are usually “deeply invisible, as is the work involved in creating and using them” (Star 2002: 113). In this sense, my analysis aims to make these infrastructures, as well as the work that supports them, visible.

the form of an open-ended horizon of the unremarkable and/or repressed) (Suchman 2012: 55).

In this sense, a critical study of facial expression recognition systems requires to trace and connect the historical and cultural imaginaries that support and inform their development – both explicitly and implicitly. From a methodological perspective, this means paying attention to how historical and cultural imaginaries are enacted and shape the local and situated development of FER systems. In the next section, I discuss how these theoretical considerations translated into my research practice.

Methodology and methods

To address my research questions – *how do practitioners make and validate their classification choices?* –, I adopted a qualitative approach that would allow me to locate and investigate the sense-making practices of FER development. Science and Technology Studies (STS) have long highlighted the importance of local contingencies for technology development and its social impact. In this respect, qualitative accounts of science and technology have avoided simplistic and linear portrayals of design practices, treating them instead as heterogeneous negotiation processes where knowledge and expertise are distributed among the networks of people that engage with technology development (Law and Callon 1988; Woolgar 1990; Stewart and Williams 2005). However, as mentioned in the previous chapter, this sensitivity towards the local and contingent must be contextualized within broader socio-political as well as historical dimensions. In this regard, Critical Data Studies have proposed to examine the complex of elements underpinning the production of data-driven technologies, including

systems of thought, forms of knowledge, finance, political economy, governmentalities and legalities, materialities and infrastructures, practices, organizations and institutions, subjectivities and communities, places, and the marketplace where data are constituted (Iliadis and Russo 2016: 3).

This combined approach allowed me to look at the diversity of processes occurring at the different scales at play in the development of FER systems. At a larger scale, I looked at the broader history of efforts to quantify and govern affective behaviors (as shown in chapters 3 and 4), and how these might inform local organizational and design practices. At a smaller scale, I looked at the local and situated negotiation and interpretation practices that underpin technology design and development (as shown in chapters 5 to 7). This methodological approach also allowed me to circumvent the “opacity” of algorithms – the technical as well as commercial inscrutability of proprietary algorithms (Burrell 2016) – and, to borrow from Paul Dourish, investigate instead algorithms’ “others” (Dourish 2016): data structures and practices, coding schemes, as well as representations and disseminations of technological imaginaries. Rather than stable and fixed technical objects, I viewed FER systems as socio-technical practices, emerging from complex configurations of knowledge, power, vision, and labor. As Nick Seaver has argued, an “anthropology of algorithms” can bring to the fore these various dimensions (Seaver 2018) and offer new entry points into analyzing, criticizing, and transforming technology. The next sections discuss my methods of choice – document analysis and semi-structured interviews – in more detail, along with the challenges faced throughout the research.

Document analysis

Documents represented an important source of information for this study. In reviewing technical as well as biographical material, I was able to explore the genealogy of contemporary FER systems, and the epistemological foundations on which they are

grounded. In chapter 3 and 4, which traditionally would constitute a review of the literature, I explore the early developments of FER systems, the inception of the field of affective computing, as well as the logics and imaginaries that drove these projects. In chapter 5 and 7, I also make use of scientific literature on expressive behaviors and machine learning to contextualize participants' responses and the projects discussed. Some of this literature was signposted to me by the participants' themselves as evidence of the scientific rigor of their methods. I analyze this material and expand on its socio-ethical implications. In addition, industry participants provided a number of original documents such as technical guides, white papers, and blog posts, which helped me familiarize myself with their products and approaches and add depth to the interview responses. However, for reasons described below in the research ethics section, I could not cite these documents without providing identifiable information about the organizations that produced them.

Semi-structured interviews

Qualitative interviews occupied a central space in my data collection practice. Between the fall of 2020 and the summer of 2022, I interviewed 22 practitioners working on the development of FER systems, from across both the industry and the academic sectors. The next sections offer a more detailed description of my research participants and my sampling strategy; here, however, I wish to discuss some methodological aspects of the data collection process.

Semi-structured interviewing is a qualitative research practice that involves a set of pre-determined questions (the 'interview guide' described later in this chapter) but gives participants sufficient flexibility to explore new and unexpected topics. This flexibility allows the researcher to gain rich insights into the participants' views of the

world (Bryman 2016). I considered this approach to be promising for obtaining first-hand accounts of design practices – how they classify affective behaviors and how they navigate the tensions associated with this classification. It is worthwhile to stress here the importance of the notion of ‘accounts’ in qualitative interviewing. Produced through these conversations are in fact “social constructs”, generated by “the self-presentation of the respondent and whatever interactional cues have been given off by the interviewer about the acceptability or otherwise of the accounts being presented” (Dingwall 1997: 11). This does not imply that interviews produce false or unreliable accounts, only that these are situated within participants’ experience. While ethnographic observations within one or multiple organisations would have enriched the data collection process (providing more texture to the ‘situatedness’ of participants’ responses), these had to be forfeited due the Covid-19 pandemic. However, as shown in chapters 5 and 6, interviews allowed me to bring to the fore the discursive repertoires adopted by participants to negotiate the tensions arising from affective classifications.

A further point of consideration is the terminological choice with regards to how I refer to the participants I interviewed. Often, qualitative studies of scientific and technical knowledge refer to research participants as “experts” by virtue of their role as informants, differentiating between scientific and technical expertise from other forms of everyday knowledge (Döringer 2021; Bogner, Littig, and Menz 2009). However, literature in the sociology of knowledge and expertise has problematized this definition to include forms of non-institutional or specialized knowledge and account for new forms of knowledge production and foregrounded the political role that attributions of expertise and authority have in decision-making processes (Epstein 1995; S. S. Jasanoff 1987). In addition, over the course of this research, my experience within the ethnographic field has further complicated participants’ attribution of expertise. For instance, those who did not hold qualifications in computer science (but were nonetheless responsible for essential work, such as data annotation), repeatedly

cautioned me that they “were not the experts”. As I discuss in chapter 6, these considerations foregrounded the political role that definitions of expertise play within data structures. Further, as discussed in the final section of this chapter, my own intervention in the field of computer science as a peer reviewer made the designation of ‘expert’ even more discomforting. For these reasons, I use ‘practitioners’ to refer to my participants, in the pragmatic and literal sense of the term as “those who practice a profession or an art”. With this definition, I aim to be inclusive with respect to the various forms of knowledge, roles, and career stages involved in the development of FER systems.

Sampling and recruitment Strategy

Participants were selected based on their direct experience in developing or working with facial expression recognition systems and techniques. Given the relative scarcity of insights into commercial developments of FER systems, I originally planned to sample exclusively industry practitioners and complement interviews with in-situ ethnographic observations within one UK-based commercial organization (for which I had begun negotiating access). However, as mentioned above, the outbreak of Covid-19 early on in my research forced me to re-adjust my methods and to instead study practitioners “at-a-distance” (Forsey et al. 2015). Further, the organization I had originally selected as field-site was not comfortable with having me access their digital spaces (such as online meetings and Slack channels) due to confidentiality reasons. Therefore, I decided to extend my sampling strategy to practitioners from multiple organizations and from across both industry and academic settings. The choice to include participants from both sectors has both pragmatic and conceptual reasons: first, because I primarily relied on “snowballing” – meaning that I asked interviewed participants for recommendations on other potential interlocutors (Bryman 2016) – I

was often re-directed to academic practitioners. Industry practitioners were in fact less likely to recommend their own competitors and, given the underlying theme of my research (questions about the validity of affective classification), I often felt that by recommending an academic colleague, participants were signalling their membership or proximity to a community of 'legitimate' scientists and practitioners. Second, some of the participants worked or had experience of working at the interface between public and private sectors: some of them, for instance, had been encouraged by their own academic institution to spin out their research into commercial areas. This fluidity of the boundaries between industry and academia within my sample reflects the growing influence of market dynamics on university departments and, particularly, computer science and engineering spaces (Whittaker 2021; Meyer 2003; Nyeko and Sing 2015). Further, while I observed some significant differences in how industry and academic participants viewed their practice (which I discuss more in-depth in chapter 5), both groups shared fundamental assumptions about the classification of affective behaviors, as shown throughout the thesis. The limitations of this approach are, however, discussed in the concluding chapter of this thesis.

Participants were recruited from the UK, Europe, and the United States. These are by no means the only geographical contexts where FER systems are designed and developed. For example, scholars and advocacy groups have noted how, in recent years, China has rapidly expanded its market for emotion recognition and behavior analysis technology, and urged critical attention into its development and deployment contexts (Article 19 2021). However, I focused on these three regions due to a) my familiarity with the interviewees' broader socio-cultural contexts, and b) participants' proximity to the debates over the validity of FER systems taking place within English-speaking academic or public spheres. Indeed, participants often made direct references to some of the critiques of emotion recognition mentioned throughout this thesis.

Further, while I originally aimed to represent a variety of diverse voices and experiences, most of the practitioners who accepted to take part in the study were males. The final sample is, after all, a more or less accurate representation of the fields of computer science and engineering in the regions studied, where women account for equal or less than one third of STEM graduates as per result of historical gender inequalities.⁷ It is worthwhile noting that, while information about participants' gender was available either through their public information (i.e., use of pronouns in online profiles) and self-presentation, I did not inquire about aspects of their identity associated with race and ethnicity, as this seemed unnecessarily invasive. I therefore choose to leave this information out of the analysis, not as a form of "color blindness" but rather to avoid issues arising from what Miri Song calls "visible minority status", that is, the attribution of "visibility as a non-White person" (Song 2020: 1).

Finally, to recruit participants, I approached potential interviewees either through "cold emailing" (that is, contacting practitioners without any prior knowledge), or through mutual connections. In this latter case, I was introduced via email or permitted to mention these mutual acquaintances when contacting the potential interviewees. In some cases, these interactions took place on the social media platform LinkedIn's messaging service. It is worthwhile mentioning that, due to the impossibility of attending in-person events (such as academic and industry conferences, workshops, and other social events) for almost the entire duration of the data collection process (which took place between the fall of 2020 and the summer of 2022), obtaining participants' trust has at times proven exceptionally difficult – particularly in the first stages of the pandemic when industry practitioners were forced

⁷ The World Bank, Gender Data Portal, <https://genderdata.worldbank.org/indicators/se-ter-grad-fezs/?fieldOfStudy=Science%2C%20Technology%2C%20Engineering%20and%20Mathematics%20%28STEM%29&view=bar>

to prioritize – as one participant put it – “economic survival”.⁸ Participants were not compensated for their time and took part in the study on a voluntary basis.

Participant characteristics

As mentioned above, throughout the research, I interviewed 22 practitioners from across the industry and academic sector. As shown in Table 1, participants spanned various career stages as well as different roles within the design pipeline, including CEOs, CTOs, machine learning engineers, data annotators, professors, post-doctoral researchers, and PhD students. A first point of consideration is the difference in participants' definitions of their disciplinary identity. While all participants worked on the development of facial expression recognition and analysis models, only a handful of academic practitioners defined their area of specialization as ‘Affective Computing’, and even fewer ever used the term ‘emotion recognition’. As I show throughout the thesis, participants’ definition of professional and disciplinary boundaries constitutes a key strategy for the validation and legitimation of their practice. Given the growing criticism, and in an effort to produce ‘objective’ and ‘accurate’ knowledge about users based on their facial expressive behaviors, participants often preferred avoiding the use of terms like ‘emotions’ or ‘affect’, to focus instead on the – somehow less charged – measurement of facial and expressive behaviors. Despite this caution, however, the terms ‘emotion’ or ‘affect’ (and related concepts) re-emerged in participants’ professional spaces, populating their organizations’ websites, promotional material, journal articles, publication venues, or the conferences they attended.

Further, the use cases for participants’ FER models spanned various research and application areas, including research to improve Human-Computer Interaction (HCI) (that is, interaction between humans and digital agents or robots), research in

⁸ Interview with Marcus, conducted on December 10th, 2020.

behavioral and medical science, as well as applications of FER to market research, advertising, in-car safety, and healthcare. This latter area, as I discuss in chapter 7, was of particular significance, with almost half of participants' efforts focused on the development of FER systems for the detection, assessment, and management of mental health and neurodevelopmental conditions such as depression, anxiety, autism, ADHD, and bipolar disorder. As I show in the chapter, here participants' work is informed by a variety of assumptions about expressive behaviors, health, and illness inherited by experimental psychology and freely encoded in automated systems.

Finally, as it is common in qualitative studies, not all of the interviewed participants are quoted in this thesis. In my analysis, I have relied on a few key voices whose accounts were particularly significant and which best encapsulated participants' overall experience. Although not cited in this work, some of the interviewees were however fundamental for my understanding of the technical aspects of FER systems as well as the political and economic dimensions of both industry and academic spaces.

Table 1 List of Participants

Pseudonym	Sector	Role within Organization	Area of Specialization*	Use-case(s)	Location	Gender	Date
Robert	Academia	PhD student	Computer Science	HCI (Robotics)	UK	M	14/10/2020
Pierre	Academia	Senior Laboratory Manager	Cognitive Science	Behavioral science research	UK	M	4/11/2020
Marcus	Academia /Industry	Co-Founder, CEO, and Professor	Computer Science (Social Signal Processing /Machine Learning)	Healthcare (Mental health, Neurodevelopmental disorders, Pain assessment)	UK	M	10/12/2020
Nathan	Industry	Co-Founder and CTO	Computer Science (Machine Learning)	Healthcare (Mental health, Neurodevelopmental disorders, Pain assessment)	UK	M	21/04/2021
Dennis**	Industry	CEO	Computer Science (Hardware and Software Engineering)	HR, Advertising, Education, Market research, Entertainment	EU	M	29/04/2021
Valerio**	Industry	Co-Founder and CEO	Computer Science (Behavioral Modelling /Machine Learning)	HCI (Digital agents and robotics), Healthcare (Stress and pain)	EU	M	3/05/2021

				management), In-car safety			
Otis	Industry	Senior Customer Manager	Computational Neuroscience	General purpose API (Behavioral science research, Healthcare, Market research, In-car safety, R&D, Workplace training and wellbeing)	EU	M	14/06/2021
Marcelo	Industry (Private Research Institution)	Senior Researcher	Computer Science (AI/HCI)	HCI, Mental health	US	M	17/06/2021
Peter	Industry	Co-Founder and CTO	Computer Science (Machine Learning)	Advertising, Market research	EU	M	8/07/2021
Daniel**	Industry	Co-Founder and COO	Advertising / Certified FACS coder	Advertising, Market Research, HCI (Robotics)	EU	M	12/07/2021
Olivia	Industry	Data Annotator and Annotation Team Lead	Data Segmentation and Annotation	Healthcare (Mental health, Neurodevelopmental disorders, Pain assessment)	UK	F	15/07/2021
Nash	Industry	Machine Learning Engineer	Computer Science (Machine Learning)	Healthcare (Mental health, Neurodevelopmental disorders, Pain assessment)	UK	M	16/07/2021

Julius**	Academia	Professor	Computer Science (Social Signals Processing, Affective Computing)	HCI (Emotion recognition)	EU	M	21/10/2021
Edmund	Academia	Research Fellow and Head of Subject Group	Computer Science (Affective Computing)	Social Agents, HCI, Healthcare (Mental health diagnostics and therapeutics)	EU	M	22/10/2021
Alex	Academia	Post-doctoral Researcher	Computer Science (Pattern Recognition)	HCI (Emotion recognition)	EU	M	10/11/2021
Cedric	Academia	Associate Professor	Computer Science (Machine Learning)	HCI, Human- Centered AI, Affective Computing	US	M	23/11/2021
Frankie	Academia	Professor	Computer Science (HCI, Robotics)	HCI, Affective Computing, Social agents	UK	F	24/11/2021
Marvin	Academia	Professor	Computer Science (Machine Learning)	Human behavior understanding, Mental health (Diagnosis and treatment)	US	M	3/02/2022
Quentin	Industry /NHS	Co-Founder, CSO, and Practicing Surgeon	Medical Science (Reconstructive Surgery)	Behavioral science research, Content creation, Training, Healthcare (Diagnosis and treatment of mental	UK	M	16/02/2022

				health conditions, neurodevelopmental disorders, and facial paralysis)			
Alice**	Academia /Private teaching institution	MSc Student and Educator at private coding school	Computer Science (AI)	Introduction to programming emotion recognition	UK	F	17/02/2022
Isabella	Academia	Professor	Psychology (Certified FACS coder)	FACS coding, Behavioral science research	US	F	8/03/2022
Erik	Academia	PhD Student	Computational Social Cognition	Emotional inference, HCI, Social agents	UK	M	18/08/2022

*Information provided by participants or available in their public profile.

** Interview conducted in Italian.

Interview procedure

Interviews took place online through video-conferencing platforms (Microsoft Teams and Zoom) and lasted between 30 and 90 minutes, with most of them lasting around an hour. As mentioned, semi-structured interviews involve a core set of questions, leaving sufficient flexibility to explore unexpected themes and topics. Building on Alan Bryman's guidance on qualitative interviewing (Bryman 2016), I formulated an interview protocol (see Annex C: *Interview Guide*) structured around three main conversational points:

- 1) *Contextual information*: With this first set of questions, I aimed to gather information about participant's background, as well as their role and responsibilities within their organisation.
- 2) *Understanding design practice*: Here, I inquired about organizational dynamics (such as professional/disciplinary boundaries and task distribution within research teams and organization) and asked a set of more specific questions about data classification, models, and evaluation measures.
- 3) *Understanding issues and challenges*: In the last part of the interview, I asked participants about challenges associated with their practice and connected these challenges with questions about broader debates about the science of emotions, research in facial expression recognition, and automation.

All interviews were conducted in English, except where otherwise indicated.

Data analysis: Interpreting interpretations

All interviews were audio-recorded, transcribed, and coded using the NVivo coding software for qualitative data analysis. To make sense of the data, I adopted an abductive approach, that is, an iterative approach to data analysis that aims to generate novel theoretical insights by combining deductive and inductive reasoning (Timmermans

and Tavory 2012). Abductive analysis requires a degree of what Mats Alvesson and Kaj Sköldbberg called “reflexive interpretation”, an attention to the interpretive and political role of both theoretical frameworks and empirical data in qualitative research (Alvesson and Sköldbberg 2017). In this sense, this work encompasses multiple degrees of vision and interpretation: as argued in the thesis, despite practitioners’ professed agnosticism and quest for objectivity, their ways of ‘seeing’ affective data are shaped by various norms, routines, and epistemological assumptions. This becomes particularly evident in the ways that, for example, coding frameworks (and their underlying assumptions) guide practitioners’ process of sense-making. In interpreting practitioners’ interpretations, I myself make use of interpretive grids that help me categorize and analyze my data. In chapter 5, for instance, I adopt Thomas Gieryn’s concept of boundary work (Gieryn 1983; 1999) to map the multiple ways in which participants’ represent and validate their practice. In chapter 6, I use Cristina Grasseni’s notion of “enskilment” (Grasseni 2004) to describe how data annotators negotiate the interpretation of affective data. To borrow again from Alvesson and Sköldbberg, I am aware that, by prioritizing this “repertoire of interpretations” (Ibid.), I may be foreclosing other possible understandings. However, in line with the theoretical framework described at the beginning of this chapter, my analysis does not intend to provide a singular truth about practitioners’ experience. Both participants’ accounts, as well as my own interpretations of such accounts, are to be intended as partial and situated. Hopefully, the reader will identify a common thread bringing together all the experiences described here. In this sense, I join here Janet Vertesi (who herself draws on feminist studies of science and technology) in claiming a form of “situated objectivity”, a form of reliability that derives from the “accumulation of so many local and different experiences” (Vertesi 2020: 25) across professional and disciplinary boundaries and sectors.

Research ethics and the politics of anonymity

In conformity with the University of Edinburgh's research ethics guidelines, I ensured that participants were giving their informed and voluntary consent by a) providing a clear overview of the nature and scope of my research prior to the interview session, allowing them time to ask for further clarifications or additional details beforehand; b) providing a participant information sheet that summarized the purpose of the research, explained the participant's rights (the right to rectify, restrict, object to data processing, and erase the data), and ensured secure storage of the data; c) asking for written consent through a participant consent form. In accordance with the General Data Protection Regulation, the Data Protection Act 2018 and the University of Edinburgh data protection policy, personal data was stored securely and safely. I submitted my ethics form to the School of Social and Political Science and received approval in February 2020 (level 1). A copy of my consent form and participant information sheet are included in the Annex A and Annex B.

As showed in Table 1, participants were given pseudonyms and granted full anonymity. In describing their experience, I tried to keep concealed as many indirect identifiers as possible.⁹ I made this decision after the first interview with an industry participant who, as the CEO of his company, wondered whether he should appear in my research with his full name to "set the record straight" on questions of scientific validity and legitimacy. I feared that this methodological choice would have set a rather different tone for participants' responses, more akin to a publicity practice than a research interview. Further, two of the participants here interviewed happened to be in a subordinate relationship with some of the other interviewees. Due to these power dynamics, I have obscured some of the details about their roles and projects. While leaving this information out of the narrative might contribute to some tensions, the full

⁹ Non-personal information that can be nonetheless used to identify participants.

anonymity allowed participants to freely express doubt and uncertainty about the field's practices.

My position as a researcher: Doing STS in Computer Science spaces

When I began this research, in 2019, the field of computer science and engineering (and its related industry) was undergoing a considerable transformation. Three years after Cambridge Analytica, and six years after Edward Snowden's revelations about the true extent of the US government's digital surveillance practices, faith in a techno-utopian AI future had somehow started to fade.¹⁰ Crucially, critical scholarship and advocacy in this area had started to expose the structural and historical harm perpetuated by automated systems at the expense of the most vulnerable (O'Neil 2016; Noble 2018; Eubanks 2018; Buolamwini and Gebru 2018; Benjamin 2019; Angwin et al. 2016; Stop LAPD Spying Coalition 2018). Media reports changed tone as well, giving more space to critical voices urging to reconsider the deployment of AI systems in public spheres (Hill 2020). This resulted in a proliferation of ethical guidelines and frameworks that – however abstract and toothless (Green 2021; Hu 2021; Bietti 2020) – signalled corporate interest in promoting responsible innovation. These transformations have shaped academic spheres as well: conference and publication venues that traditionally did not focus on the socio-ethical implications of technology, have instead started to pay more attention to these aspects and promote special conference editions and journal issues dedicated to the ethics of automated systems. In practice, this has meant that, throughout my research, I have been invited to review papers and manuscripts for a range of AI/HCI conferences and journals, which included: the 2021 Affective Computing & Intelligent Interaction (ACII) conference

¹⁰ This is, by no means, the first time that the public confidence in AI technology has plummeted. For a timeline of the so-called "AI winters" see [\(Hendler 2008\)](#).

on ‘Ethical Affective Computing’; the Institute of Electrical and Electronics Engineers’ (IEEE) *Transactions on Affective Computing, Special Issue on the Ethics of Affective Computing*; the 2022 ACM Conference on Human Factors in Computing Systems (CHI), and the *AI & Society, Journal of Knowledge, Culture, and Communication*. In peer-reviewing research written primarily by computer science and engineering practitioners, I was somehow invested of the role of “expert” in matters of technology, ethics, and society. This has forced me to reflect critically on STS as a form of knowledge production, as well as on my impact on those cultures of computer science and engineering that constitute the object of my research – a form of intervention in the field that Emily York called “critical participation” (York 2018). However, inhabiting both of these spaces (the social and, although peripherally, computer sciences) has proven at times discomforting. On one occasion, for instance, I recognized one of my participants (who I had initially contacted based on his “expertise”) as the author of a paper I had previously reviewed. At the time of the interview, this was not yet clear to me, as he had not named the publication venue. However, a few months later, he sent me the article as a follow up to our conversation. I recognized the paper and, in reviewing the interview transcript, I wondered if the considerations he was sharing during our meeting (particularly with regards to the ethics of FER systems) had not, in fact, been shaped by my review. This event unearthed a strange circularity: was he feeding back to me my own feedback? As mentioned in the previous sections, this also complicated my use of the term “expert” in the context of qualitative interviewing – who was to be considered the “expert” in this scenario? For this and the reasons mentioned above, I decided to refer to all participants as “practitioners”.

I have kept that particular portion of the interview out of the data analyzed in here. However, since then, I grappled with questions about what should be the intended audience and impact of this research. Inspired by these considerations, in 2022, I

submitted a partial and preliminary account of my findings to a well-known affective computing journal. Translating my insights into a common research language proved very difficult and, unfortunately, unsuccessful. The style and format guidelines provided by the journal (a rigid structure typical of the natural sciences that, for example, treats the data separately from its discussion) were often at odds with the more fluid and narrative approach of qualitative research (where the data cannot be separated from its interpretation). Further, in providing recommendations for ‘ethical design’ (as encouraged by the editorial guidance), I struggled with the anxiety over the potential instrumentalization of the more normative dimensions of the paper. As others have noted before me, co-participatory forms of STS studies (where social scientists can shape techno-scientific projects) risk being “instrumentalized and performed with a narrow scope. Despite the claims of the social scientists involved, their work may appear as a tool for promoting acceptance of emerging technologies” (Joly 2015: 235). A more concrete example of this is the proposed interventions to increase the “accuracy” of facial recognition technology: as many have noted, most available systems fail to correctly identify individuals with darker skin and, particularly, Black women. As work in this area has suggested, this is largely due to the lack of intersectional representation within face datasets (Buolamwini and Gebru 2018). However, increasing the accuracy of these systems might prove equally problematic when they are used for harmful surveillance practices. What needs reformulating, then, is not a technical understanding of fairness, but rather, its fundamental assumptions and use-cases.

As I argue in this thesis, the development of FER applications rests on a multifaceted infrastructure of equally problematic assumptions; a repertoire of normative ideas (about affect, behavior, and identity) on which practitioners – despite their professed agnosticism – must routinely draw on when making design choices. However, offering this argument within computer science spaces (and particularly

affective computing and related practices) may have looked like an attempt to invalidate the whole field, its ambitions, and concerns. While the feedback I received from the anonymous reviewers of my paper cannot feature in the corpus of data discussed in this thesis, it is in line with the responses I received from my participants.

Chapter 3 | Performing theories

Introduction

In the summer of 2022, a particular genre of social media content suddenly flooded the Internet: a deluge of videos and memes that featured “body language experts” dissecting the nonverbal behaviors of celebrities Johnny Depp and Amber Heard, as observed during the public broadcasting of the Depp vs Heard defamation trial. Actor Johnny Depp had sued his former wife Amber Heard following a 2018 op-ed published in the Washington Post in which she described her lifelong experience of domestic violence and abuse. Although the article did not mention Depp by name, he sued her for defamation, claiming that her testimony had ruined his reputation and future career opportunities.

The broadcasting of the trial became the theatre for a particular kind of gaze: every week self-professed “experts” tuned in to scrutinize the couple’s facial and body movements for signs of truth, deception, and mental illness, using an array of analytical techniques ranging from nonverbal communication, body language, and military interrogation. These diagnostic efforts quickly uncovered a vast industry of lie detection, often promoted by individuals with close ties with US law enforcement and military agencies. One example is The Behavior Panel, a YouTube channel run by four self-described “world's top body language and behavior experts” involved, to different extents, in military and US intelligence interrogation training.¹¹ Prior to the trial, the channel offered detailed analyses of the involuntary tics and twitches of politicians, celebrities, known murderers, and suspect criminals – and even featured a guest appearance of America’s most popular psychologist and tv personality ‘Dr Phil’. In the

¹¹ <https://www.youtube.com/c/TheBehaviorPanel/about>

eight, long episodes dedicated to the Depp vs Heard case – which average an hour and a half and 1.5 million views each – the four analysts describe the inner motives and truths of the two celebrities as “leaking” or “bleeding out” through their body language. A smirk is simultaneously an indication of Depp’s difficult past, as well as a signal of Heard’s moral contempt and deception. Categorical emotions such as ‘anger’ or ‘sadness’ are identified by breaking down facial expressions into discrete smaller movements: a brow furrow, a corner-of-the-lip raise, an eye twitch. The emphasis on the anatomical aspect of deception gives these analyses the semblance of scientific legitimacy, presenting subtle and transitory physical movements as incontrovertible facts of either truth or lie. After a meticulous examination of one of Amber Heard’s fleeting facial expressions, Behavior Panelist Chase Hughes declares:

Body language and behavior profiling is a lot like meteorology. It’s based in science and gives you likelihood. Where we deal in term of likelihood. But this is a rare occasion where I’m gonna say there is a 99, and maybe a 100% chance of precipitation here, by which I mean deception.¹²

This intense scrutiny is not limited to possible signs of deception, as the four panelists often weigh in to comment on Heard’s potential diagnosis of Histrionic Personality Disorder (HPD), periodically comparing her observed mannerisms with those expected from a person with HPD, and as described in the Diagnostic and Statistical Manual of Mental Disorders (DSM). During the trial, another so-called body language expert, Janine Driver, initiated a prolific social media campaign aimed at determining whether Amber Heard was an “innocent victim or an antagonising narcissist”.¹³ Driver, whose curriculum vitae features multiple three-lettered US intelligence and law enforcement agencies, employs the same language of anatomy and categorical

¹² The Behavior Panel, “Her ACTIONS Betrayed Her - Amber Heard Body Language Deception” https://www.youtube.com/watch?v=hXWpK0fguLc&t=1052s&ab_channel=TheBehaviorPanel.

¹³ <https://www.tiktok.com/@bodylanguageinstitute>.

emotions to provide evidence of the actress' deceit. In a live interview broadcasted by the digital network Court TV, Driver pointed to the subtle twitch appearing on the left side of Amber Heard's face as the smoking gun that would prove her guilt:

It's called contempt. It's on the left side of her face, right there. It's moral superiority. And it's out of place, especially for a battered woman. I work with battered women. I myself was dating a hotshot lawyer in my 20s who beat me up and left me for dead in his driveway. And I have to tell you, battered women or women who are domestic violence survivors, they are not leaking contempt.¹⁴

The assumption that there might be 'right' and 'wrong' ways to express emotions is rooted in an essentialist vision of emotional expressions, inherited from what psychologist Lisa Feldman Barrett calls the "classical view of emotions" (Barrett 2017). According to this view, external stimuli activate particular neural circuits in our brain which, in turn, trigger physical reactions: increased heart rate, sweating, tears, a scowl. This view also contends that these physical expressions are a product of evolution (they are necessary for survival), universal (everyone expresses and recognizes emotions in the same way), and involuntary (emotional expressions are difficult to control). This latter aspect of the classical view, argues Barrett, is central to much of Western philosophical thought which, from Plato to Freud, has opposed the lawlessness of passions and emotions to the civilising power of human reason:

Emotions are thus thought to be a kind of brute reflex, very often at odds with our rationality. The primitive part of your brain wants you to tell your boss he's an idiot, but your deliberative side knows that doing so would get you fired, so you restrain yourself. This kind of internal battle between emotion and reason is one of the great narratives of Western civilization. It helps define you as human. Without rationality, you are merely an emotional beast (Barrett 2017: xi).

¹⁴ <https://www.courtstv.com/title/4-25-22-body-language-expert-analyzes-johnny-depp-and-amber-heard/>.

This classical view of emotions is the primary frame of reference for the diagnostic gaze of body language experts like Janine Driver or The Behavior Panel: a way of ‘seeing’ affective behaviors that, since the 1960s, has gained renewed popularity thanks to an American psychologist named Paul Ekman. As illustrated in this chapter, Ekman’s research on the universality of (some) emotional expressions has since dominated Western mainstream understandings of expressive behaviors. As Driver explains to the tv host, the contempt she so clearly sees on Amber Heard’s face is one of the seven universal emotions identified by Ekman:

There’s seven universal emotions, and it sounds ridiculous, I get it. It looks stupid. You are like, ‘Janine, what are you talking about? Maybe she just had a twitch’. Listen there’s seven emotions. This guy named Dr Paul Ekman, he’s on the same list as Freud as the 100 psychologists to ever influence the world, the planet!¹⁵

As I show in this chapter, Ekman’s research, and particularly his work on expressions of deception, has greatly influenced social and technical imaginaries around affective behaviors: not only are many Facial Expression Recognition (FER) systems grounded in Ekman’s understanding of affective behaviors; Ekman has directly shaped computational efforts to automate the recognition and classification of affective behaviors. As scholar Kelly Gates has noted, the technical development of FER systems is “tightly bound to the field of psychology: psychological theories of facial expressions and emotions inform the design of automated facial expression analysis systems, and those systems in turn promise to advance the field’s knowledge of facial expressions and emotions” (Gates 2011: 152). It is, however, a particular brand of psychology that Ekman sought to automate, often in service of law enforcement and intelligence agencies. With lie detection at the core of his scientific project, Paul Ekman has been instrumental to the creation of a techno-scientific gaze that sees facial expressions

¹⁵ Ibid.

primarily as involuntary “leaks”: a code that can be “cracked” (Ekman 2016), regardless of people’s consent or intentions, to reveal the truth behind their behavior.

This chapter traces the genealogy of automated systems for the recognition and classification of affective behaviors, starting from Paul Ekman’s 1960s studies on the universality of expressions of emotion. Here, I discuss the centrality of Ekman’s work on deception, and how this culminated in his Facial Action Coding System (FACS), a coding framework for the analysis of affective behaviors that has directly shaped the early development of FER systems and, to this day, represents the standard method for the analysis of facial expressions. Crucial to Ekman’s project (and to computational efforts to automate it) was the assumption that it is possible to design a scientific and objective method for the classification and analysis of affective behaviors, stripped from their socio-cultural context and turned into calculable and manageable data. As I discuss in the chapter, this assumption has since been largely contested. As a result, practitioners have progressively abandoned claims of emotional inference (the idea that facial expressions can reveal accurate information about emotional states) to focus on the anatomical description of facial behaviors. However, the common genesis of Ekman’s work on deception, FACS, and automated FER systems show that it is not truly possible to divorce computational efforts to measure facial expressions from efforts to make sense of people’s internal states.

Reading faces: A tale as old as (D)ARPA

Paul Ekman is best known for his cross-cultural study of facial expressions and the theorization of a set of discrete, universal (or pan-cultural) expressions of emotions. Between 1966 and 1971, funded by a subdivision of the US Department of Defense, Ekman set out to determine whether facial expressions of emotions are culture-specific or – as originally theorized by Charles Darwin in *The Expression of Emotions in Man*

and Animals (1872) (Darwin 1998) – whether there are elements of universality that can be attributed to at least a particular set of expressions. To find evidence of this universal grammar of emotions, Ekman ran a series of experiments with members of the indigenous Fore community in the South-East highlands of Papua New Guinea. People from this region, reasoned Ekman, had not been exposed to Western mass media, did not speak English, nor had ever “worked for a Caucasian” (Ekman 1970: 155). Removed from any potential representation of Western expressions of emotions, Ekman could test whether humans can recognize and convey emotions regardless of their cultural contexts. The experiment was fairly simple: he would show the Fore a set of pictures (as shown in Figure 1) representing six stereotypical emotional expressions – anger, disgust, happiness, fear, sadness, and surprise – and asked them to describe them. These experiments however, as he recounts in his self-published memoir, did not initially yield particularly strong results. Ekman had to rely on a translator, and he could not successfully explain the task to his research participants (Ekman 2016). For his second trip to Papua New Guinea, in 1968, Ekman changed strategy: he showed participants three photographs representing emotional expressions and told them three stories associated with different emotions; he then asked the Fore to match each story with a corresponding picture. With the exception of expressions of fear – often confused with surprise –, the Fore demonstrated to recognize emotional expressions in ways comparable to Western subjects. This study alone did not lead Ekman to claim the universality of emotional expressions. In an experiment conducted with Japanese and American college students, he showed his participants videos featuring a travel documentary and a “series of stress-inducing films” (Ekman 1971: 242). Unbeknownst to the students, Ekman recorded their facial expressions and later showed this data to a different sample of participants, asking them to guess whether the viewer was watching the neutral or the distressing material. According to Ekman, “about 60% of

the judgments" (243) were accurate, therefore providing "strong evidence" (244) in support of the universality of emotional expressions of emotions.

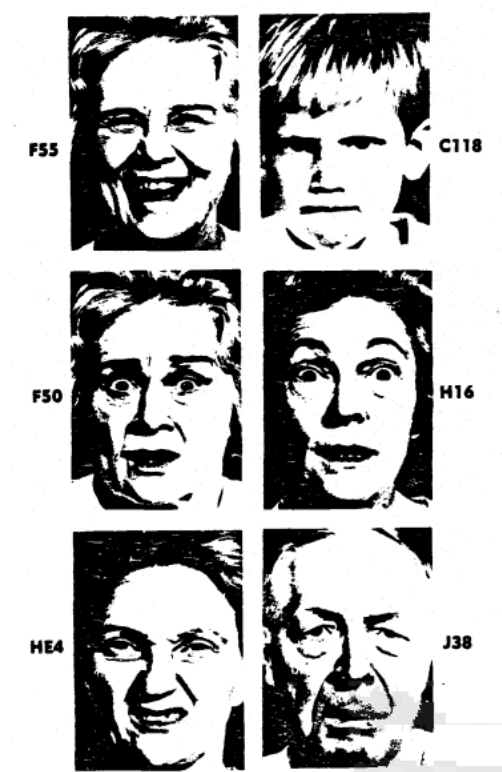


Figure 1. Photographs of the six basic emotional expressions used in Ekman's cross-cultural study (reprinted from Ekman 1971)

Ironically, the study on the universality of emotions was not Ekman's idea. As he acknowledges in his autobiography, the research had been written and assigned to Ekman by Lee Hough, the head of behavioral sciences at the US Department of Defense's Advanced Research Projects Agency (ARPA).¹⁶ Ekman, who had a training in clinical psychology, did not know how to run cross-cultural studies, nor was familiar with the relevant literature. However, despite his initial reticence, he received close to a million dollar – a sum equivalent in 2022 to nine million dollars – to conduct the research. Years later, he discovered that the reason behind ARPA's insistence had to

¹⁶ Today, the agency is known as the Defense Advanced Research Projects Agency (DARPA).

do more with the personal affairs of Hough and less with intellectual curiosity. At the time, Hough had been caught using money allocated for social science research to gather information that could be used to remove from power Chile's left-wing president Salvador Allende. As Ekman recounts,

Lee had money earmarked for overseas research, and he had to spend it quickly, before the end of the fiscal year, or he wouldn't get as much money the next year. And I walked through his door. An innocent who could do overseas research that wouldn't get him into trouble! (Ekman 2016: 46)

At the heart of Ekman's efforts – and crucial to his professional career – was, however, his fascination with lie detection. By studying facial expressions, Ekman had hoped to design a scientific method to spot clues of deception from nonverbal behaviors. His first, although unpublished, study, funded by the National Institute of Mental Health (NIMH) in 1955, sought to show that facial micro-expressions were accurate indicators of people's emotional states and, particularly, of stress. The experiment, termed the "emotional status exam", made use of a photographic camera to record the changes in nonverbal behaviors occurring in psychiatric patients admitted for severe depression. Ekman positioned the camera in front of the patients, gave them a remote control, and asked them to take photographs when they thought their facial expressions were matching their emotional states. He later showed these pictures to "untrained people", asking them to indicate which ones had been taken at admission and which ones had been taken at discharge. As he recounts, the participants "achieved 100% accuracy" (Ekman 2016: 36). These seemingly – by his admission – banal results were the first piece of a larger puzzle that would lead Ekman to argue that, because largely unintentional, facial micro-expressions can "leak" accurate information about a person's emotional state. As illustrated in the next section, the concept of "leakage"

would later become central to Ekman's efforts to design a scientific method for the detection of deception.

Anatomy of a lie: the Facial Action Coding System

In 1967, the director of research in the psychiatry department at the University of California San Francisco (UCSF), Enoch Calloway, asked Ekman if his research on nonverbal behaviors could help address a problem encountered while treating psychiatric patients. As he explained, those hospitalized for suicide attempts would ask for a day or a weekend pass but would sometimes lie about their recovery progress so that they could "complete the act" while outside of clinicians' purview (Ekman 2016: 93). Convinced of the role played by micro-expressions in deception detection, Ekman set out to determine whether his research could help psychiatrists assess patients' behavior and, potentially, save those patients' lives.

Early in his research, Ekman had come to realize that still photographs were insufficient to capture micro-expressions and had turned instead to 16mm film recording. To perform meaningful analysis of patients' expressive behaviors, however, he needed to be able to sort through the footage to compare and contrast different video excerpts. He then received funding to develop a computer interface that could allow him to perform high-speed retrieval of video data. The interface, called Visual Information Display and Retrieval (VID-R), was designed to search and retrieve visual records, view them at different speeds, as well as code video data. This latter function was particularly important for what Ekman called the Systematic Classification and Analysis of Nonverbal Behavior (SCAN), the application of VID-R on patient data "for classifying all observable body and facial movements without reference to an a priori theory about classes or types of movements" (Ekman, Friesen, and Taussig 1969: 3). The coding function would allow an operator to mark the beginning and end of any observable movement and save it for later analysis. This procedure did not aim to

impose meaning to the data, as “SCAN does not itself tell us the meaning of the nonverbal behavior; the SCAN output is the beginning point for our analysis of the data” (10). However, by defining when one expression ended and a new one began, SCAN helped demarcate the single units of expressive movement. Through this coding process, Ekman could extract individual body and facial movements, strip them of their social context, and transform them into manageable *data*. This was the first step in Ekman’s lifelong effort to code affect.

In the eighteen months that it took engineer Thomas Taussig to develop the VID-R system, Ekman published his first article on deception, “Nonverbal leakage and clues to deception”, co-authored with Wallace V. Friesen (who will become Ekman’s primary collaborator for the next twenty-five years). This was the first published occurrence of the term “micro-expressions”, which duration the authors describe as “so short that they are at the threshold of recognition unless slow motion projection is utilized” (Ekman and Friesen 1969: 93). Central to his investigation of deception clues, Ekman theorized micro-expressions to occur at 1/50 of a second, corresponding to a “one motion-picture frame” (94). Serendipitously fitting the affordances of video technology, these micro-movements would be almost imperceptible to the naked eye, Ekman reasoned, due to “their being embedded in other expressions which distract attention, their infrequency, or some learned perceptual habit of ignoring fast facial expressions” (Ibid.). However, with the development of the VID-R system, Ekman was now seemingly able to unlock the truth behind nonverbal behaviors. Missing from his systematic study of deception was a formal method to code and analyze facial expressions.

In 1971, together with psychologist Silvan Tomkins, he published a coding framework called Facial Action Scoring Technique (FAST), a tool for the measurement and analysis of facial behaviors. FAST built on Ekman’s cross-cultural study and Tomkins’ own work on the universality of facial displays. As Ekman recounts in his

memoir, Tomkins had been, together with Charles Darwin, an intellectual father figure: he (like Darwin) had postulated the existence of discrete, biologically determined, affective expressions. Further, Tomkins' fascination with the face was said to manifest in his legendary ability to interpret human expressions. According to a *New Yorker's* article, Tomkins was known to be "the best face reader there ever was" and that he could "walk into a post office, go over to the 'Wanted' posters, and just by looking at mug shots, tell you what crimes the various fugitives had committed. Tomkins felt that emotion was the code to life and that with enough attention to particulars the code could be cracked" (Gladwell 2002).¹⁷ Tomkins was also known for his early work on the use of computers in psychiatric research inspired by cybernetics and information theory.¹⁸

Ekman and Tomkins based FAST on the same six basic emotions the New Guinea study had concluded to be universal: happiness, sadness, anger, fear, surprise, and disgust. Next, they divided the face into three areas: forehead, eyes, and mouth (as shown in Figure 2). The reason behind this tripartition was that it would allow to easily mask the two areas of the face that were not being coded when using photographs or video footage. As the authors described, three scorers received a six-hour training on how to separately code the three facial areas and identify individual movements, called FAST 'items'. The scorers were then asked to match individual FAST items with

¹⁷ Critical readings of Ekman's work often point to the influence that Tomkins had on the theorization of the six basic emotions. Adam Frank and Elizabeth Wilson, however, have argued that Tomkins' role in Ekman's research is often over-stated and that there are significant theoretical and political differences between Tomkins' and Ekman's understanding of the universal nature of emotions. As they describe, "Basic-ness, for Ekman, isn't just a measure of the discrete categorical differences between emotions; it is also an argument that emotions are biological in a fundamental, invariant way. In this sense, basic emotions are a weapon against the cultural relativism, linguisticism, and social constructionism that he feels brought the study of emotion into disrepute". Tomkins, conversely, "uses the universality of facial expression to give affects a psychological distinctiveness that has been eliminated in a discipline that has increasingly come to regard cognition as king" (Frank and Wilson 2020: 37-38).

¹⁸ For a discussion on the influence of cybernetics on Tomkins' thought, see (Sedgwick 2003; Wilson 2011; Leys 2011).

emotion categories. The purpose of this task was to show that individual facial movements could be used to predict emotions, without any information regarding the social or cultural context in which the emotional expression occurred. As the paper describes, scorers did not always agree with each other, and more often than not, combined FAST items (meaning, multiple facial areas) were necessary to successfully predict emotions. Such disagreement was not seen as proof of methodological flaws; it only suggested that “on occasion, any scorer would fail to grasp all the details in the facial area to be scored, and thus would disagree with the other two scorers” (Ekman, Friesen, and Tomkins 1971: 48). The results from the study were deemed significant enough to confirm FAST a reliable tool for the measurement of facial expressions. However, a year after the publications of FAST, anthropologist Wade Seafood showed Ekman a facial movement that was not described by his facial taxonomy. As Ekman reports in his autobiography, Seafood: “pushed up his lower lip activating the *mentalis* muscle in his chin area. I was devastated by Wade’s demonstration, not knowing how much else, how many other facial movements we had missed in creating FAST” (Ekman 2016: 97).



Figure 2. FAST facial areas for the emotion ‘surprise’ (reprinted from Ekman, Friesen, and Tomkins 1971)

Ekman was determined to create a comprehensive tool for the measurement of facial movements that could be used to spot clues of deceit. Therefore, he decided to use the remaining of his NIHM grant, originally allocated to the study of depressive patients' nonverbal behaviors, to develop what would later become the Facial Action Coding System (FACS). For over a year, Ekman and 'Wally' Friesen sat in front of a camera to record as many individual micro-expressions as possible, aiming to create an exhaustive catalogue of all facial muscle movements. Central to this endeavour was the work of Duchenne de Boulogne, a 19th century neurologist working at the Salpêtrière asylum in Paris, where (as described more in depth in chapter 4) he applied electrical shock techniques on patients' face to stimulate and document facial muscle movements. Using Duchenne's muscle descriptions as a starting point, Ekman and Friesen identified a little over than forty movements, which they called Action Units (AUs), as well as three hundred combinations of AUs. These efforts were apparently made easier, as Ekman describes, by his "exceptional control" of his facial muscles (Ekman 2016: 101). Finally, the researchers used the collected data to compile a "self-instructional manual" that included photographic and video examples of each AU and their combinations. An additional grant allowed them to test the reliability of FACS by asking three postdoctoral researchers and two research assistants to self-study the manual and separately code a video excerpt. The results confirmed the reliability of the systems and, in 1978, Ekman and Friesen published the Facial Action Coding System manual.

FACS is described as an objective and comprehensive tool for the measurement of facial behaviors (Rosenberg and Ekman 2020; Ekman 2016), its objectivity deriving from a purely anatomical description of facial expressions. However, to extract meaningful information from such description – to determine whether combinations of AUs can be indicative of emotional or mental states – a "higher level of interpretation" (Gates 2011: 172) is necessary to make sense of the data. For this reason, Ekman and

Friesen developed a “selective system” (Rosenberg and Ekman 2020: 16), called “Emotion FACS” (EMFACS) focused only on the facial movements associated with basic emotions. As the authors describe, when using EMFACS, “the coder describes only those facial events that include AUs or combinations of AUs that are ‘core’ to certain emotion expressions. EMFACS is less time consuming than FACS. It should be understood that EMFACS is FACS selectively applied” (Ibid.). In addition, Ekman and colleagues created a database to support and facilitate EMFACS coding, called Facial Action Coding System Affect Interpretation Dictionary (FACSAID), which contained representations of facial expressions previously coded, and the meaning assigned to them. Scholar Kelly Gates has noted how, in an online description of the database (no longer available at the time of writing), “one of the weaknesses in the database in its current form is that the rules for interpreting behaviors are not explicit, the authority of the experts who interpreted them being the only basis for confidence in their accuracy” (Hager in Gates 2011: 173). This approach, she argued, performed a reverse inference of the affective behavior it aimed to classify. By relying on expert coders, the accuracy of EMFACS is “constructed in part through the claims to objectivity associated with the initial stage of FACS coding” (Ibid.). Despite these claims, however, FACS (like SCAN and FAST) is in itself an effort to create new meaning about facial expressions. As Gates notes, FACS “does not delineate a set of facial expression components placed there by God, evolution, nature, or culture, waiting for psychologists and computer scientists to lift them off the face and place them on a grid so that their true meaning can be established scientifically” (174). With FACS, Ekman established a new grammar of affective expressions.

Automating visions: FACS meets neural networks

Since its publication, FACS has been adopted by large corporations and research institutes (such as Pixar, Apple, and the Salk Institute) as the standard coding framework for the analysis and modelling of facial behaviors. However, FACS coding is a lengthy, painstaking process. Learning how to FACS code requires about a hundred hours of individual study, and coding a minute of video data can take up to a hundred minutes. Ekman hoped that advancements in computing techniques could help automate, at least partially, the coding process. Colleague and psychologist Robert Levenson warned Ekman that the complexity of FACS coding was an indication of its scientific rigor: “only the most serious dedicated scientists use FACS, because it took a lot of work. If it could be done quickly, then hordes of slipshod scientists, who abound in academic psychology, would do slipshod research on the face” (Ekman 2016: 104). Ekman abandoned the idea until, in the early 1980s, he believed to have found a potential solution. As he describes in his autobiography, he was attending a conference in Wales where he heard a researcher describing an automated facial recognition system called “Wizard” that could “spot problematic people who might be approaching a secure facility from fifty feet away” (106). While Ekman does not name the researcher, Kate Crawford identified him as Igor Aleksander, a computer scientist working for the British government who had trained his neural network on a dataset of criminal mugshots, a practice that will later become commonplace in government-led facial recognition advancements (Crawford 2021). Aleksander, as Ekman recalls, was facing a technical challenge: the system could recognize individuals unless they were making a facial expression. This would change their appearance to the point that the computer would no longer be able to recognize their face. After his presentation, Ekman approached Aleksander and told him that “his noise – facial movements – was my data, while his data – different facial appearances – was my noise” (Ekman 2016: 106). The two researchers spent four days at Aleksander’s lab in London, in an effort

to train the neural network to recognize different AUs. If successful, Ekman believed that the system could save FACS coders more than half of their coding time.

In the early 1990s, inspired by these events, Ekman applied to the National Science Foundation (NSF) for funding that would allow him to organize an international conference on automated measurements of facial expressions, bringing together computer scientists, psychologists, behavioral neuroscientists, and physicists. In the final report of the conference, Ekman and neural networks pioneer Terrence Sejnowski described how the automation of facial expressions promised insight into people's affective states, cognitive processes, personality traits, "truthfulness", which included "the leakage of concealed emotions, and clues as to when the information provided in words about plans or actions is false" (Ekman, Sejnowski, and Hager 1993: 4), as well as psychopathology, including signs of mental illness and other neurodevelopmental disorders. The conference working groups outlined the challenges for the automation of facial expression recognition, and made a series of technical recommendations, with a particular emphasis on the need for a shared dataset containing still photographs, video and audio data of spontaneous facial expressions, as well as "psychophysiological correlates of specific facial actions, and interpretations of facial scores in terms of emotional state, cognitive process, and other internal processes" (16). These considerations shaped the development of future research on FER systems and, in particular, the creation of public facial expression datasets.

Among these was the RU-FACS, a collection of video recordings of spontaneous facial expressions "with rigorous FACS coding" (Bartlett et al. 2006), developed at the Machine Perception Lab at the University of California San Diego (UCSD) by Sejnowski's collaborators and former students. As the authors explain, the dataset's subjects were recorded while taking part in a "false opinion paradigm", a psychological experiment replicated from Ekman's own studies on deception shown to elicit a wide variety of facial expressions: participants were asked to fill in a questionnaire about

their opinions on social or political matters. Then, interviewed by a retired police officer or an FBI agent, participants were told that if they said the truth about their opinions and were believed, they would receive \$10, whereas if they lied and were believed they would receive \$50. If they lied but were not believed, they would receive no compensation and would be required to fill another long questionnaire (Bartlett et al. 2006). In the original instantiation of the experiment, it was Ekman himself interviewing his subjects while holding a copy of his book *Telling Lies*, as a proof (and a threat) of his exceptional lie detection abilities. Like in the RU-FACS experiment, successful liars would receive a compensation of \$50. Those who failed to convince Ekman of their opinions, however, would face a punishment “as severe as I could get the Institutional Review Board, the committee protecting human subjects, to approve” (Ekman 2016: 138). In a rather unsettling description of the experiment, Ekman explains that,

If they [the participants] were accused of lying – like in the real world, it didn’t matter whether they were actually lying or not, innocent people get jailed if the judge or jury convicts them – they would get punished. I then gave them a taste of the punishment. They entered a totally dark room the size of a telephone booth, and once they were seated, they heard a sound as loud as a firecracker, just ten decibels below the level at which there is a risk of damaging hearing. When they came out of the room labelled in large letters “punishment chamber” I told them that if they were accused of lying, they would have to be in the punishment chamber for sixty minutes and hear forty blasts of noise. If they didn’t want to take that risk and opted not to participate in the experiment, I offered them \$5 for their time so far. None refused. They were all males; no females responded to our posted invitations to participate (Ibid.).

As one of the first datasets of spontaneous expressions, RU-FACS laid the foundations for the application of FER systems outside of the lab. As the authors describe in their paper’s conclusions, the automated measurement of spontaneous facial expressions had the potential to bring about “paradigmatic shifts” in behavioral science and psychiatric research as much as in law enforcement, counter terrorism, and security

practices (Bartlett et al. 2006: 11-12). In 2012, Marian Stewart Bartlett and her colleague at the Machine Perception Lab, Javier R. Movellan, founded Emotient Inc., an emotion recognition company that would later become famous within Silicon Valley media spheres for its ‘Sentiment Analysis Glassware’, a Google Glass facial expression analysis application (Lunden 2014). Around that same time, Ekman had learned that the CIA was interested in the development of automated facial expression recognition, and had convinced the intelligence agency to arrange a “horse race”, a government-funded technical competition between Emotient and another AI team (formed by researchers Takeo Kanade and Jeffrey Cohn) to advance progress in the development of FER systems (Ekman 2016). The race was soon abandoned with the appointment of a new CIA executive director (a woman who, in his memoir, Paul Ekman calls ‘the Witch’), however the two teams continued their work separately, shaping the future advancements of FER systems within both the industry and academic sector.

At this time, Ekman, who had been sitting on Emotient’s scientific advisory board, began expressing concerns for the privacy implications of the commercialization of his methods. In a 2015 interview with the *Wall Street Journal*, he took distance from misuses of FACS: “I can’t control usage, [...] I can only be certain that what I’m providing is at least an accurate depiction of when someone is concealing emotion” (Dwoskin and Rusli 2015). Despite this recent cautiousness, Ekman has played a central role in shaping social and technical imaginaries around affective behaviors. In the 2000s, the idea that facial expressions could reveal meaningful information about people’s mental and emotional states had become popular outside psychology (and law enforcement) circles thanks to its media representation in *Lie to Me*, a tv show (for which Ekman was a scientific consultant) about a deception expert who could tell lies from truth simply by observing people’s nonverbal behaviors.

Within computer science and engineering domains, as hopefully illustrated so far, Ekman’s research on micro-expression, and particularly on FACS as a method to

de-code facial behaviors, has been key to developing FER systems. Central to efforts to automate the recognition and classification of facial expressions, FACS has also shaped the standardization of methods to generate animations of affective behaviors. Indeed, FACS is at the core of the MPEG-4 Facial Animation Standard, a set of international parameters for the animation of digital characters' facial expressions (Pandzic and Forchheimer 2002; Aleksic, Potamianos, and Katsaggelos 2005).

“We don’t do mind reading”

Over the years, Ekman’s research on the cross-cultural aspects of emotional expressions has received many critiques. In a 1994 review of Ekman’s methods, American psychologist James A. Russell questioned the validity of Ekman’s comparative study of Japanese and American students, arguing that the forced-choice format of the experiment had been shown to influence participants’ responses, leading to consensus rather than disagreement (J. A. Russell 1994). In the 2010s, Russell partnered with a team of psychologists and anthropologists to replicate the New Guinea Study with a fundamental difference from Ekman’s original research: the team spoke tribe’s local language and had “extensive prior fieldwork experience” in the area (Crivelli et al. 2016: 1). They found that, among members of the same population, expressions of fear were interpreted as conveying threat, a result that seemed to challenge the existence of cross-cultural basic emotions and support instead a view of affective behaviors that sees facial expressions as context-dependant “social tools aimed at influencing others in social interactions” (Ibid.) – as opposed to universal and involuntary leaks of internal states. The authors emphasized the significance of the study’s results for the application of Ekman’s theories within social contexts. As they argued, their findings were important not only in “challenging psychology’s approach of allegedly pancultural

‘basic emotions,’ but also in applications such as emotional intelligence tests and border security” (Ibid.).

More recently, in a comprehensive review of the literature on emotional inference and facial expressions, psychologist Lisa Feldman Barrett and colleagues have argued that the available evidence fails to support a cross-cultural view of basic emotional expressions (what they call the “common view” of emotional expressions) and that the ways in which people communicate emotions vary significantly across cultures, and even across the same individual (Barrett et al. 2019). Moreover, the authors noted how most experiments that have tested the universality of emotional expressions have largely ignored how contextual information influences people’s expressions and perception of affective behaviors. As they argued, someone’s expression of anger, for example, might depend on a variety of contextual factors, such as,

the situational context (e.g., whether a person is at work, at school, or at home), social factors (e.g., who else is present in the situation and the relationship between the expresser and the perceiver), a person’s internal physical context (e.g., how much sleep they had, how hungry they are), a person’s internal mental context (e.g., the past experiences that come to mind or the evaluations they make), the temporal context (what occurred just a moment ago), differences between people (e.g., whether someone is male or female, warm or distant), and the cultural context, such as whether the expression is occurring in a culture that values the rights of individuals (compared with group cohesion) and is open and allows for a variety of behaviors in a situation (compared with closed, having more rigid rules of conduct) (11).

According to the review, other factors can influence the reliability of the “common view” on emotional expressions: often experiments ask participants to use a set of pre-determined labels (joy, anger, sadness, etc.) to recognise emotional expressions, influencing participants’ choices and responses. When those labels are removed, studies have shown that agreement rate among participants drops significantly, almost at chance levels. Moreover, research on emotion perception often employs images of

actors performing emotions or uses computer-generated facial expressions, which do not correspond to people's real, existing emotional state (Ibid.).

Barrett has since then shared the wider implications of her research on emotional expressions with the AI industry. In a 2020 public lecture at Microsoft, the psychologist asked the audience in the room:

Can a machine actually read emotion in your face? Well, there are plenty of companies who claim to have done it already, including yours. But what I am going to tell you today is that these claims are based on some fundamental assumptions about emotions which are probably wrong.¹⁹

In describing her review of the research on universal emotions, Barrett showed the room a picture of the six posed expressions of emotions – happiness, sadness, fear, anger, surprise, and disgust – as routinely used in psychology experiments. What is interesting about these expressions, she described, is that “they were not discovered by observing how people actually move their face in everyday life. They were stipulated by a handful of scientists and then adopted as universal truth, and then scientists built a whole science around them.”²⁰ However, as she explained throughout her lecture, AI companies often rely on the assumption that these expressions are a valid representation of people's affective behavior, and ground their technical systems on this very assumption.

Following a mounting body of work presenting evidence of the societal harms, and particularly racial discrimination, propagated by AI systems, (Noble 2018; Eubanks 2018; Benjamin 2019; O'Neil 2016; Browne 2015), scholars have pointed to the disparate impact that automated systems claimed to classify people's emotional states based on their facial expressions can have on marginalized groups. In a comparative analysis of commercial FER systems Microsoft AI and Face++, researcher Lauren Rhue

¹⁹ <https://www.microsoft.com/en-us/research/video/can-machines-perceive-emotion/>.

²⁰ Ibid.

showed that both systems attributed more negative emotions to Black men than their white counterparts. Testing the software on a publicly available image dataset of professional basketball players, Rhue provided evidence that “Face++ consistently interprets black players as angrier than white players, even controlling for their degree of smiling. Microsoft registers contempt instead of anger, and it interprets black players as more contemptuous when their facial expression is ambiguous” (Rhue 2018: 1). This type of analysis is part of a broader scholarship on bias and fairness in automated systems that in recent years, by auditing different commercial software, has brought to the fore the disparate impact that computer vision systems have on Black, Indigenous, and people of color (BIPOC), and gender nonconforming individuals (Buolamwini and Gebru 2018; Raji et al. 2020; Raji and Fried 2021; Scheuerman, Paul, and Brubaker 2019; Scheuerman et al. 2020; Keyes 2018). However, as Rhue notes in the conclusions, while her study has presented evidence of racial disparities in emotional classification, the question of whether AI systems can *actually* detect emotions, and can do so better than humans, remains unanswered (Rhue 2018). Others have taken this question as a point of departure and moved away from technical understandings of algorithmic bias to focus instead on the legitimacy of the conceptual underpinnings of FER systems. Perhaps the most vocal among these is media scholar Luke Stark who has included FER systems in the array of data-driven technologies that he and Jevan Hutson termed “physiognomic AI” to describe

the practice of using computer software and related systems to infer or create hierarchies of an individual’s body composition, protected class status, perceived character, capabilities, and future social outcomes based on their physical or behavioral characteristics (Stark and Hutson 2021: 7-8).

As the authors note, AI systems claiming to produce meaningful knowledge about individuals based on their physical appearance propagate assumptions grounded in

“scientifically baseless, racist, and discredited pseudoscientific fields” that “should be anathema to any researcher or product developer working in computer science today (4-5). Stark and Hutson refer here to the 19th century practice of physiognomy, the widespread belief that a person’s facial appearance could reveal something about their moral character. As they argue, AI systems claiming to recognize and classify emotional states based on facial expressions, like physiognomic practice, assume a correlation between physical expressions of emotion and internal states:

Often grounded in the Basic Emotion Theory (BET) developed by American psychologist Paul Ekman, these systems assume discrete categories of human emotion are universally legible via external signals such as the movement of the face, or tone of voice, and that emotions motivate human behavior in ways that are hard to consciously suppress — claims that have been contested on multiple scientific and social grounds (27).

As the authors note, the resurgence of physiognomic claims in computer science practice requires urgent attention as it poses a threat to human and civil rights. For this reason, they have called for a ban on all developments of physiognomic AI “such that it is as legally and politically unpalatable as it is morally” (30). Others have joined Stark’s and Hutson’s call for prohibitions of the development and deployment of these systems based on their lack of scientific validity, often citing Barrett’s review of Ekman’s methods (Crawford et al. 2019). As a result of these critical efforts, in 2022, Microsoft has limited both commercial and public access to some of the facial analysis features of its facial recognition service (Azure Face), particularly with regard to emotional states, gender identity, and age.²¹

As described in this thesis’ introduction, however, most practitioners have remained largely peripheral to this debate. Research publications as well as tech companies’ marketing material continue to adopt Ekman’s methods (and vocabulary)

²¹ <https://learn.microsoft.com/en-us/legal/cognitive-services/face/transparency-note>.

to describe and classify facial affective behaviors. This is not out of pure stubbornness. As scholar Kelly Gates had already observed in 2011, “there is a concerted effort in the technical literature to distinguish facial expression analysis from the analysis of emotions, an insistence that they represent two distinct problems” (Gates 2011: 155). Indeed, following Paul Ekman’s more recent caution mentioned above, practitioners have progressively abandoned claims of emotional inference and focused instead on the recognition of facial expressive *behaviors* – a rhetorical as well as technical move that I describe more in depth in chapters 5 and 6. As Gates noted,

This move to separate the specific problem of facial expression recognition from the recognition of emotion defines the aims of any particular effort at computerization as narrowly as possible, separating it from the messy and exceedingly more difficult problems associated with the meaning of facial displays and their relationship to what people are thinking and feeling (Ibid.).

In practice, this has resulted in efforts to decouple FACS coding (as an “objective” and descriptive method to code facial expressions) from higher levels of interpretations of affective behaviors, as well as from its association with Ekman’s research on expressions of emotions. As testified by the practitioners interviewed in this study (and as illustrated in chapters 5-7), only a few “uninformed” or “bad” actors would try to infer categorical emotions from image or video data of people’s faces; most well-meaning practitioners are instead aware that emotional inference lacks substantial scientific validity. As participants frequently described, “We don’t do mind reading”; they would limit their practice to the “objective” description of affective behaviors. And yet, to produce any sort of meaningful knowledge about facial expressions – and justify potential applications of FER systems – practitioners must make assumptions about what those expressions might mean. In addition, despite practitioners’ efforts, it is not truly possible to divorce FACS (on which most FER systems rely) from its genesis as a coding framework historically developed to identify liars and criminals, in service of law

enforcement and intelligence agencies. I follow here Deborah Raji who, in mapping the historical developments of facial recognition technology (as in, biometric systems developed for identification purposes), has argued that “current attempts to revisit the narrative and re-frame the purpose of the technology to supposedly benign commercial applications” (Raji et al. 2020: 8) cannot and should not obscure the military legacy of these systems. As illustrated throughout this chapter, Ekman’s legacy, and particularly his work on deception, has greatly shaped technical developments of FER systems, from models’ assumptions (the idea that facial expressions can “leak” some form of knowledge about internal states) to data collection practices (the elicitation of “genuine” or “deceptive” facial expressions, such as in the RU-FACS dataset), and evaluation measures (the evaluation of computers’ accuracy compared to that of human FACS coders).

However, Ekman’s work on deception is just one of the many layers of a multifaceted infrastructure of knowledge on which FER projects rest on. As I illustrate in the next chapter, FER systems are just the latest episode in a broader history of quantification of human experience that, at least since the 19th century, includes the scientific classification of affective behaviors. Further, in an effort to distance their practice from inferential claims about emotions, practitioners are increasingly turning their attention to adjacent theories of the mind and expressive behaviors. As shown throughout chapters 5 to 7, these theories, although not directly or explicitly associated with Ekman’s work, rest on similar assumptions about affective behaviors: namely, that these can be objectively observed, measured, and used to produce inferential knowledge about people’s internal states. Borrowing from scholar Ruth Leys, the reason behind practitioners’ reticence to abandon categorical and deterministic accounts of affective expressions, might lie in the fact that this would “force them to provide thick descriptions of life experiences of the kind that are familiar to

anthropologists and novelists but are widely held to be inimical to science” (Leys 2011: 471).

Conclusion

In this chapter, I traced the genealogy of Ekman’s work on expression of emotions and showed the significance of its legacy for modern FER systems. In particular, I focused on Ekman’s work on nonverbal clues of deception and how this led to the development of the Facial Action Coding System (FACS). Here, I showed how FACS, designed as an objective and scientific framework for the analysis of facial behaviors, has directly shaped early developments of automated FER systems. Despite practitioners’ recent efforts to abandon inferential claims about facial expressions and emotions, the entangled roots of Ekman’s research on deception, FACS, and the development of FER systems indicate that it is not fully possible to divorce the classification of facial expressions from its genesis as a tool to obtain the *truth* – sometimes by means of intimidation, threat, and deceit, as showed by Ekman’s perhaps questionable research methods. Often supported by or in service of US military and intelligence agencies, Ekman designed FACS as a scientific method to spot lies: a type of inquisitive gaze inherited by modern FER models and transposed to various application domains under the guise of statistical objectivity.

The next chapter places Ekman’s work within a larger history of the quantification of the affects contributing to the social regulation of affective behaviors. As I show, efforts to develop a techno-scientific framework for the measurement and classification of affective expressions served a broader project of social classification undertaken by Western societies in the 19th century – efforts that are reflected in modern affective computing practice.

Chapter 4 | The making of affect

Introduction

In the late 1990s, Affective Computing pioneer Rosalind Picard set out to promote visions of a symbiotic fusion between humans and machines – one where machines would adapt to humans and promote human flourishing and well-being. Despite Picard’s efforts, however, today Affective Computing practice resembles earlier historical attempts to quantify and classify universal affective expressions. This chapter traces the discursive and material practices that inform the classification of affect, from late 19th-early 20th century physiological studies of emotions to Affective Computing. Here, I review secondary sources recounting the constitution of affects as objects of laboratory knowledge and then move to Rosalind Picard’s first-hand accounts of the field’s inception. This historical perspective places FER systems into a broader trend seeking to quantify and mechanize affective human experience, starting with the creation of scientific procedures in the late 19th century to visualize, stabilize, and assess human behavior, to late 20th century efforts to quantify emotional expressions through computational means.

The establishment of protocols and standards for the measurement of affective behaviors would eliminate the “suspect mediation” (Daston and Galison 2007) of human intervention and turn emotions (and their representations) into ‘objective’ scientific facts. Tracing these efforts and, particularly, the scientific practices that legitimated and co-constituted them, historicizes the rhetoric and practices underpinning computational forms of classification of emotional expressions, and highlights the role that human intervention and interpretation played in the measurement and classification of affective experience. In addition, the chapter draws

attention to the performative power of this classification – the ways in which the measurement and classification of affective behaviors can shape our emotional experience – and how the establishment of scientific protocols, standards, and techniques can reinforce these performative and normalizing dimensions. Building on chapter 3, this section concludes with a reflection on the role that conceptual models of emotional expressions play in technical developments of affective computing applications and in practitioners’ articulations of social responsibility.

Affects in the laboratory

The quantification of affective behaviors is not new. Scientific knowledge about affective experience has, in one way or another, long legitimated projects of social classification: the identification of the criminal, the insane, the ‘unfit’, the unproductive. According to historian of science Otniel Dror, the numerical representation of affects has its origin in late 19th century physiological studies of affective behavior and the creation of “emotion-as-number” (Dror 2001). In this new form of disembodied representation, emotion was disassociated from feminized portrayals of the affects and from ideas of emotion as “disruption” to industrial labor (Ibid.). New scientific procedures, ‘objective’ techniques, and “inscription devices” (Latour 2012) allowed men in the laboratory to gauge intimate emotional knowledge from their research subjects and elevate it to the status of scientific object that could be recorded, quantified, manipulated, standardized, and reproduced.²² A central figure in this affective turn was Italian physiologist Angelo Mosso who, experimenting on

²² By translating emotions into “the scientific language of numbers and graphs”, argued Dror, scientists could “explicitly discourse about the emotions of their laboratory animals without reverting to what they defined and perceived as a feminized and oppositional type of emotion talk. The number was an important technology for the reframing of ‘emotion’ and its integration into the discourse of the laboratory” (Dror 2001: 371).

animals to study blood circulation, observed that expressions of pain could disrupt and obfuscate the study of physiological processes (Dror 2001). Expressions of affect captured the interest of other physiologists and turned into a site of investigation in itself for the study of the relationship between internal states and physiological responses. New instruments and protocols proliferated. Italian criminologist Cesare Lombroso – founding father of phrenology as well as Mosso’s mentor – applied these new techniques to “the psychological study of the delinquent – in order ‘to penetrate, as with an instrument of precision, into their inner psychology’” (Dror 1999a: 213).

Implicit in these investigations was the claim of a stable and objective relationship between affective experience and its scientific representation, the latter ultimately replacing and standing for the emotion it meant to represent. However, the interpretation of these representations was not straightforward; borrowing from Gaston and Dalison, it required the exercise of “trained judgements” (2007): the ability of a trained professional, or “expert”, to discern meaningful patterns in the data. This new form of objectivity – as opposed to truth-to-nature and mechanical objectivity²³ – required a “*physiognomic sight*”, the ability to “synthesize, highlight, and grasp relationships in ways that were not reducible to mechanical procedure, as in the recognition of family resemblance” (324), such as the group stereotyping invoked in racial classifications of faces. As it will become apparent in chapter 6, these subjective – albeit shaped by social and cultural factors – dimensions of late 19th-early 20th century science of emotions are still central to computational models of facial expression recognition and largely overlooked in current critical analyses of FER systems.

²³ Lorraine Daston and Peter Galison have shown how scientific objectivity has emerged and changed meaning throughout history. In particular, they noted how, in the 19th century, the “mechanical objectivity” of photography (which aimed to eliminate all human intervention by means of mechanical reproduction) replaced the “truth-to-nature” of scientific illustrations (which aimed to portray an ideal, abstract depiction of natural phenomena) (Daston and Galison 2007).

Classification breeds performance

This transformation of affective experience into an object of scientific knowledge, along with the rise of numerical and mechanized form of representation, resonates with the broader epistemological and cultural transformation of Western sciences in the late 19th century, seeking to anchor human experience into biological, quantifiable truths. Social Darwinism, combined with new technologies and experimental procedures, crept into different scientific disciplines, both enabling and harnessing positivist and eugenicist projects of social progress. According to its proponents, this new physiology of emotions would shed light on the underlying causes of social ills – fatigue, insanity, or criminality – and encourage the reproduction of productive, sane, and ‘fit’ individuals. One example of this Paolo Mantegazza (1831-1910), an Italian physician and anthropologist with a profound admiration for Darwin’s ideas on evolution, who set out to identify the hereditary laws of emotional expressions unique to each culture. His *Atlas of Pain*, which collated original photographs along with artistic representations of emotional expressions taken from Ancient Greece, Christian and Renaissance art, aimed to shape the recently created Italian identity (Chiarelli 2020). To further serve this purpose, he “consciously made a comparison with other physiognomies from distant cultures, such as Polynesians, Jews, Japanese, and what he called ‘the Negroes’” (Martin Moruno 2016: 153).

However, as historian Dolores Martin Moruno noted, Mantegazza was conscious of the unscientific value of artistic representations of pain. To transform his photographic evidence into “scientific facts”, he developed a scientific protocol that included standardized head orientation (a full frontal and a profile picture), cranial measurements for each racial group, and the use of the same oval frame and neutral background “as a rhetorical strategy to homogenize pictures and present them as data that could provide scientific evidence for the law that he had previously enunciated”

(Ibid.). To record genuine expressions of pain, Mantegazza would inflict painful or unpleasant experiences to his subjects: exposure to bright lights or loud noises, ingestion of sour food, and even physical harm (see Figure 3).

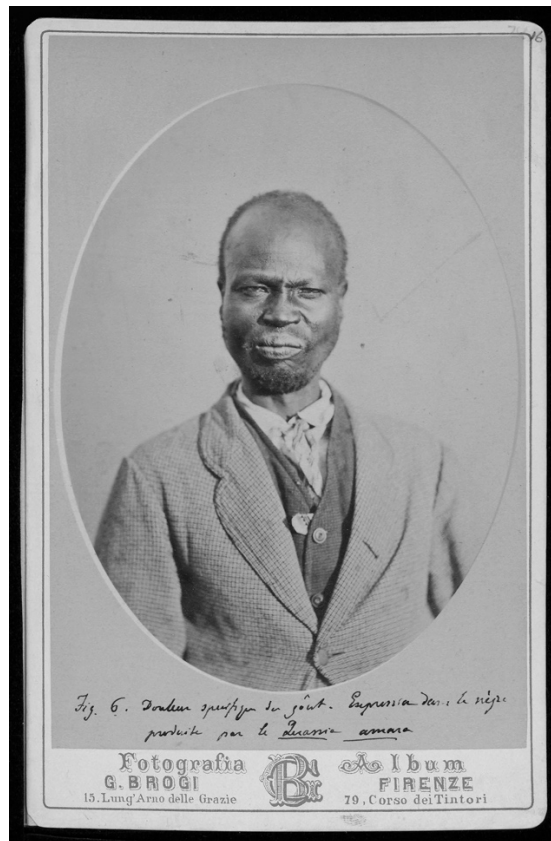


Figure 3. Expression of “gustative pain” in a Black person, from Paolo Mantegazza’s *Atlas of Pain* (reprinted from Moruno 2016).

In illustrating representations of pain, Mantegazza reinforced colonial ideas of Western superiority: Italian men’s sophisticated expressions of pain conformed to social and cultural norms of emotional expressions, as opposed to those of the “savage” and the “insane”. Moreover, notes Moruno, Mantegazza’s careful selection of photographs and use of captions aimed at persuading the reader into this way of seeing. “All these decisions show that Mantegazza had consciously defined what pain should look like throughout the photographs that he included in the *Atlante*” (157).

Similarly to Mantegazza, and perhaps more famously, French neurologist Duchenne de Boulogne (1806-1875) employed the new photographic technique to develop a scientific study of the face, *Mécanisme de la physionomie humaine* (Mechanism of human physiognomy) (Duchenne 1876). Duchenne's work aimed at mapping all muscles of the human face through the combination of photography and electricity. As shown in Figure 4, he applied electrodes to the faces of his subjects – patients of the Salpêtrière psychiatric centre – to stimulate involuntary movements of facial muscles and capture photographic evidence of their expressions, often staging dramatic poses that suggested sexual objectification or class prejudice (Mauro 2000). As Kate Crawford notes, Duchenne's photographs were not representations of genuine expressions but simulations. Nonetheless, he believed that by adopting photography and other technical and scientific procedures he could “transform the squishy business of representation into something objective and evidentiary, more suitable for scientific study” (Crawford 2021: 163). In his study on the universality of emotional expression, *The Expression of Emotions in Man and Animals*, Darwin praised Duchenne's scientific efforts to codify human expressions and used his photographs as evidence of his theory (Darwin 1998). As mentioned in the previous chapter, Ekman himself built his work on Duchenne's study on facial muscle movements (Ekman and Rosenberg 2005; Ekman 2016). What is now known as a *Duchenne smile*, a “configuration of the enjoyment smile identified by Duchenne: the orbicularis oculi with zygomatic major”,²⁴ has now become the benchmark in psychology and Affective Computing to describe genuine expressions of happiness (Campos, Keltner, and Tapias 2004).

²⁴ <https://www.paulekman.com/blog/fake-smile-or-genuine-smile/>.



Figure 4. Duchenne de Boulogne measuring expressions of surprise on a Salpêtrière patient (W. Bruce and Delaney H. Lundberg Fund)

Building on Foucault's study on power and governmentality, Nikolas Rose noted how the creation of these new forms of documentation and recording of affective information in the late 19th century had enabled the psychological sciences to visualize, stabilize, and assess human behavior. "The formation of a plane of sight and a mean of codeability establishes a grid of perception for registering the details of individual conduct" (Rose 1988: 187). Human behavior had become both visible and intelligible, "no longer lost in the fleeting passage of space, time, movement, and voice but identifiable and notable in so far as [it] conform[s] to or deviate from the network of norms which begins to spread out over the space of personal existence" (Ibid.). These new techno-mediated forms of emotional representation

were not predicated on a theorized sympathy, empathy, or emotional “contamination,” but on a new mode of retrieving knowledge of the intimate that did not depend on the traditional values and time-consuming practices that heretofore had enabled intimate emotion-talk. They created new possibilities by mechanically transgressing the emotional divide between private and public (Dror 1999b: 392-393).

Through quantification, mechanical representation, and exteriorization, the scientific study of human expressions became a site for the redefinition of emotions themselves. Representations of emotions would serve as blueprints of what emotional expressions should look like, often shaped by class, racial and gendered assumptions underpinning the physiognomic project of social classification. Enabled by new and mobile “emotion-gauging technologies” and widespread circulation of emotion images, physiologists, psychologists, and clinicians “actualized and painted the world, its objects, and its inhabitants in emotions” (Dror, 1999b: 401). This new entangling of emotions, machines, and science facilitated the construction of cultural norms of emotional expression and a hierarchy of feeling that privileged certain expressions – and certain emotions – over others. As it will become more apparent in chapter 7, contemporary affective technologies – combined with the predictive power of machine learning – can shape our affective experience in a similar manner.

Cybernetic feelings: between feedback and control

In the early 2010s, Professors Rafael Calvo, Sidney D’Mello, Jonathan Gratch, and Arvid Kappas sent a proposal to *Oxford University Press* for the publication of *The Oxford Handbook of Affective Computing*, a project aiming to be “the definite reference for research in the burgeoning field of affective computing” (Calvo et al. 2015:1), a comprehensive anthology of almost two decades of research, theory, methods, and state-of-the-art applications in Affective Computing. To open the 41 chapters, authored both by “world leaders” (Ibid.) in the area and emerging

researchers, they invited pioneer of Affective Computing Rosalind Picard to contribute with an adaptation of a paper originally published in 2010 for the first issue of the IEEE's *Transactions on Affective Computing* (TAC). Here, Picard traced her personal history and experience in launching the field of Affective Computing – a discipline bridging computer science, engineering, psychology, cognitive and neuroscience – starting in the early 1990s. At the time, she recounted, “emotion was associated with being irrational and unreasonable” (Picard in Ibid. 13). Being a woman in engineering and cognizant of the derogatory meaning assigned to the word “emotional”, Picard did not want to risk her reputation by working on the affects.

Computer scientists, indeed, did not take Picard seriously. Ingrained in the field was the assumption that computer science – and, by association, intelligence – had to do with logic and abstractions, not that memory, attention, and decision-making might be influenced by emotions. However, with the support of the MIT Media Lab, the dissemination of her ideas through the publication of her foundational book *Affective Computing* in 1997, Picard started witnessing a growing interest in the topic and its centrality to the development of artificial intelligence and, in particular, computer vision. In trying to give computers more sophisticated forms of vision, Picard had realized that emotions played a fundamental role in human perception: “Building a vision system is not just about detecting high-contrast oriented lines or telling a dog from a cat. Vision is affected by attention, and attention is affected by what matters to you. Vision—real seeing—is guided by feelings of importance” (13). Developing emotion recognition and computer vision were, to Picard, linked endeavors: you could not train machines how to ‘see’, without training them how to understand emotions.

Affective technologies – of which wearables constituted an essential component – were intended to *adapt* to humans and promote human well-being. Among the possible applications, Picard had imagined a “computer-interviewing agent” capable of coaching its users for a date or a job interview. In addition, she reasoned, affective

technologies would help autistic children developing their emotional skills or reduce anxiety in mental health-care patients. Finally, the collection of emotional data through wearables could help broaden scientific knowledge about how we learn, communicate, and feel (Picard 2000; 2010). Picard's idea of a symbiotic fusion between humans and machines echoed that of cybernetic and cyborg visions. Indeed, Picard drew on cybernetic principles of "affective regulation", "feedback", and "control" (Robinson and Smith-Lovin 2006) to develop computer systems for the detection of affective behavior. In the above-mentioned IEEE inaugural paper, Picard describes the influence that Manfred Clynes – the scientist who coined the term "cyborg" with co-author Nathan Kline (Clynes and Kline 1960) – had on her work:

One morning over breakfast cereal and the Wall Street Journal (the only nontechnical journal I regularly read) a front-page article about Manfred Clynes caught my eye. He was described as a brilliant inventor who, among better-known inventions that became commercially and scientifically successful, also invented a machine for measuring emotion. His "sentograph" (*sentire* is Latin for "to feel") measured slight changes in directional pressure applied to an immovable button that a person pushed on. The finger pushes showed characteristic patterns related to joy, sadness, anger, sex, reverence, and more. This is not a list approved by mainstream emotion theorists — they don't include sex or reverence — and Manfred is far from main-stream, a child prodigy with a fan letter from Einstein for his piano playing and coauthor on the 1960 paper that coined the word "cyborg." But he *measured* emotion, and later, the measures were replicated by others. I was amused by this crazy fact, although not enough to do anything more than file the article. The article mentioned my friend, Marvin Minsky, who many years later introduced me to Manfred, and we then became friends (Picard 2010: 11-12).

In the late 1990s, together with her colleague at the MIT Media Lab, Jennifer Haley, Picard developed a prototype: the StartleCam, a "cybernetic wearable camera" (Healey and Picard 1998) connected to a physiological sensor supposed to detect the body response to emotional and psychological stimuli. The increase in arousal (determined through changes in "skin conductance") would trigger the recording of events

supposed to be of interest to the user. The two inventors used the expression “cybernetic control” to describe

the control loops formed by the nervous system, the brain, and the wearable mechanical-electrical communication system. [...] This control loop might also be considered ‘cyborgian’ as the term cyborg denotes man in quasi-symbiotic union with electro-mechanical homeostatic control systems (1).

The human-machine entanglement that she envisioned was seemingly far from that of predictive computing. To achieve affective perception, machines required “information that we value, information that is relevant to personal needs, goals, and preferences” (Picard 2000: 240). Technology, Picard reasoned, needs information that is only important to us. Indeed, she opposed Orwellian visions of social control:

The goal should not be to control people or to tell them what to do; this is both unethical and likely to lead to data that indicates primarily stress and anger. Instead, the idea is that such a system [wearable technologies] would only be worn in a voluntary mode, by somebody who is interested in either learning about their affective patterns, or in communicating them to a computer, friend, or trusted physician, for reasons of benefit to the wearer (237).

Scholars have pointed to the lack of critique in Picard’s project and her seeming inability to foresee the power dynamics that would have turned her cyborgian visions into a site of profitable extraction (Angerer and Bösel 2016; Zuboff 2019). In the early 2000s, Picard met Rana El Kaliouby, a PhD student at Cambridge University who had developed a facial-expression reading software – *Mind Reader* (R. A. E. Kaliouby 2005) – supposed to assist autistic people in social interactions.²⁵ The two moved their research to the MIT Media Lab, where they were quickly inundated with requests from corporate sponsors (Microsoft, Honda, NASA, Nokia, among others) to use their system to monitor consumers behavior. Here, the Media Lab encouraged Picard and El Kaliouby to “spin off” their research into a start-up. In 2009, they co-founded

²⁵ Following disability scholars and advocates, I use here and throughout the thesis an identity-first approach.

Affectiva, an emotion analytics company that quickly shifted its focus from assistive technologies to market research and, among the affective recognition techniques, favored facial analysis. According to a New York Time piece, Picard was later forced out of the company due to conflicting views over the company's direction (Khatchadourian 2015).

Data-driven emotions?

It is worth noting that, in Picard's original vision, "emotion recognition" hinged on the supposed ability of computers to *measure* emotions, intended as "observations of motor system behavior that correspond with high probability to an underlying emotion or combination of emotions" (Picard 1995: 6). She was not, however, concerned with the definition of emotion per se (e.g. as a psychological or neurological category). Indeed, she often employed emotion, affect, and feelings interchangeably. For Picard, the lack of consensus on what conceptual models best measure affective behaviors – whether emotional expressions are universal and involuntary or culture-dependent and intentional – would not interfere with the field's ultimate objective: "If the problem can be solved in a speaker-dependent way, so that your computer can understand you, then your computer can translate to the rest of the world" (4). With a speaker-dependent approach, speech-recognition systems are trained only on the user's data and cannot respond accurately to other users. In this sense, whether emotions are universal or culture-dependent is not necessarily relevant for the development of affective technologies. The task of identifying affective states "*only need[s] to demonstrate consistent patterning for an individual in a given perceivable context*" (Ibid. emphasis added). As mentioned earlier in this thesis, this conceptual agnosticism lies at the core of Affective Computing and, perhaps more broadly, of data-driven science. In this sense, Affective Computing resonates with what commentators would

have later termed the “end of theory” (Anderson 2008). Famously, Chris Anderson – former editor-in-chief of *Wired* – proclaimed the beginning of a new epistemological paradigm enabled by the sudden deluge of Big Data and the spread of computational statistical models, which had transformed the scientific production of knowledge from *theory*-driven to *data*-driven: “Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough” (Ibid.).

Rob Kitchin however, among others (boyd and Crawford 2012; Gitelman 2013), pointed out how this narrative conveniently served the interests of data analytics companies and similar “knowledge-oriented businesses” who could now promise “the possibility of insightful, objective and profitable knowledge without science or scientists, and their associated overheads of cost, contingencies, and search for explanation and truth” (Kitchin 2014: 5). Data-driven science, on the other hand,

is guided in the sense that existing theory is used to direct the process of knowledge discovery, rather than simply hoping to identify all relationships within a dataset and assuming they are meaningful in some way. As such, how data are generated or repurposed is directed by certain assumptions, underpinned by theoretical and practical knowledge and experience as to whether technologies and their configurations will capture or produce appropriate and useful research material (Ibid.).

Such is the case of Affective Computing and, particularly, the development of models for the classification of affective facial behaviors. While theoretical commitment to a certain conceptual model of emotions and emotional expression is not deemed necessary for the task of facial expression recognition, mainstream affective models – such as FACS/EMFACS – guide the development and the application of machine learning systems. At the same time however, and as shown previously in chapter 3, the affordances of computer vision played a central role for the development of emotion science itself. As Anna Tuschling eloquently puts it:

On the one hand, computers are the technological basis on which certain psychologies of emotion and affect (Tomkins', Ekman's and Friesen's psychologies, most notably) were developed; through affective computing, on the other hand, computers now 'receive' the scientific knowledge on emotion they made possible in the first place (Tuschling 2014: 189).

Borrowing from Donald MacKenzie and Yuval Millo's study on the impact of economic theories on financial markets, "It can reasonably be said of this technosystem that it performs theory" (MacKenzie and Millo 2003: 128) and, at the same time, that theory shapes the development of these systems. The epistemic affordances of computer vision and machine learning facilitate the quantification of facial expressions as conceived by theories of expressive behaviors. These theories then guide practitioners' ways of *seeing* affective behavior. As a result (and as shown in chapter 6), significant human mediation is required to fit the model's assumptions and expectations. This interplay of theory and technical development has another important effect: in navigating the debate concerning the scientific foundations of their systems, participants often described themselves as merely performing 'technical work', thus building practical and ethical boundaries between technical developments and emotion science and deferring social responsibility only or mostly to the latter. This instrumentalization of science – how practitioners perceive, use, and *mobilize* theory – is investigated in more depth in Chapter 5.

Conclusion

The quantification of affective behavior is not a novel phenomenon. Historians of science have located the origin of this trend in 19th century physiology when the creation of new instruments, new scientific protocols, and new ways of seeing made it possible to elevate emotional expressions to objects of scientific knowledge and use this knowledge to legitimize projects of social classification. This scientific effort to measure, quantify, and display emotions helped corroborate and normalize implicit or explicit assumptions or expectations towards affective behavior dis/aligned with social and cultural norms. Despite Picard's seemingly techno-utopian visions of feeling machines, affective computing practice still reflects earlier historical efforts to classify human behavior.

19th century physiologists' standardization practices echo today's *data wrangling*: the process of cleaning, standardizing, and annotating the many thousands of images that constitute computer vision training datasets. As for Mantegazza's choices about head orientation, background color, and captioning, the process of collecting, cleaning, and annotating training images for machine learning datasets "is itself a form of politics, filled with questions about who gets to decide what images mean and what kinds of social and political work those representations perform" (Crawford and Paglen 2021: 33). These computational *atlases* of emotional expressions contribute to what Anna Tuschling has called the "globalization of the affect [...]" intended to be valid irrespective of space and time and thus also attain global reach in the service of a better understanding of facial expressions" (Tuschling in Angerer, Bösel, and Ott 2014: 188). Indeed, media scholars have noted how the design choices about the affordances of affective technologies can facilitate the stabilization of particular logics of feeling (Stark 2020b). As I will discuss in chapter 7, the use of affective technologies can encourage users to conform to hierarchies of expressive behaviors that conform with dominant socio-cultural norms. Gabriele Gramelsberger

has called this phenomenon “externalized introspection”, a “new” form of representation of individuality that “transgresses the traditional form of statistics, because it is linked to predictivity for individuals and it directly maps the artifactual information back on us” (Gramelsberger in Bösel and Wiemer 2020: 46). Similarly, Marie-Luise Angerer and Bernd Bösel have pointed to the “normalizing dimensions” of this “new affective programming [which] promises to work far more subtly than anything from the age of mass media ever could have done” (Angerer and Bösel 2016: 49). However, by drawing attention to the continuity between 19th century physiologists and contemporary Affective Computing practitioners, I argue that this performative element was already prominent in historical efforts to quantify human emotional experience. In both cases, scientific practice is mobilized to legitimize hierarchies of affective behaviors. Similarly to the protocols, standardization efforts, and visualization techniques that informed the creation of 19th century *Atlases* and affective records, the ‘objective’ classification of automated facial expression recognition systems is made possible by local and situated practices enacted throughout the design pipeline.

Altogether, this brief historical account of techno-scientific efforts to capture affective behavior can help persuade against framings of affective computing developments that focus on the field’s “novel” element. The *making of affect* forms the historical backdrop for the next three empirical chapters. Interview data will illuminate: the role that mobilizations of science and practical knowledge play in efforts to legitimize or discount projects of affective classification (Chapter 5); the role that human mediation and interpretation plays in the classification and automation of affective experience (Chapter 6); and, finally, the performative and normalizing dimensions of such classification practices, with a focus on healthcare applications of FER systems (Chapter 7).

Chapter 5 | Mobilizing science

Introduction

As mentioned in this thesis' introduction, , despite the public and scholarly controversy over efforts to classify affective behaviors, many practitioners maintain an agnostic view with regards to the conceptual models of emotional expressions they adopt to develop FER systems. This chapter explores the multiple ways participants working on FER systems mobilize science to establish the credibility of their claims. Drawing on interview data, the chapter analyzes the rhetorical practices shaping participants' articulations of validity and how these articulations provide strategies to navigate social and ethical responsibility. First, I discuss various forms of "boundary work" (Gieryn 1983) as the ways participants construct professional and ethical boundaries to establish the legitimacy of their practice. Here, the analysis is modelled around STS scholar Thomas F. Gieryn's categories of "expulsion", "protection of autonomy", and "expansion" (Gieryn 1995). This tripartite analysis is not meant to represent a systematic catalogue of all forms of boundary work in affective computing and/or machine learning, as Gieryn's categories do not always perfectly overlap with this study's findings. Indeed, all of the episodes of boundary work here recounted can be read equally as expressions of either expulsion, protection of autonomy, or expansion. The analysis offered here, however, intends to be a useful guide to map practitioners' effort to establish credibility in the face of criticism. Moving through the rhetorical practices of expulsion, autonomy, and expansion, the chapter surveys the various forms of boundary work necessary to participants to ascertain authority over the "cartographic space of science" (Gieryn 1995): the moral demarcation between *good* and *bad* actors, the profession of *agnosticism* over conceptual models of expressions of

emotions, the characterization of engineering as *ambiguous* practice, the demarcation of *pure science* from its applications, the absorption of other knowledge domains' epistemological assumptions into FER practice, and, simultaneously, the entrenchment of affective computing inferences within complex socio-technical contexts. Then, I consider the material practices and tactics that allow participants to validate their system's predictions: since performance metrics cannot establish the 'truth' of emotional experience, participants describe how they focus instead on the ability to track and predict changes in expressive behavior. This chapter's responses are situated within the broader literature around algorithmic harm and predictive technologies.

Boundary work in affective computing practice

The contextual backdrop of this chapter is the ongoing public and scholarly debate over the legitimacy of FER systems claimed to infer inner truths – from affective states to moral values – from the analysis of facial expressions. As discussed in chapter 3, these criticisms are primarily directed to computational approaches relying on psychological theories that postulate the existence of discrete and universal facial emotional expressions. Scholars have drawn attention both to the lack of scientific consensus around such theories (Barrett et al. 2019) and to these systems' potential to perpetuate pseudo-scientific claims, hidden behind the rhetoric of scientific and mathematical objectivity (Stark and Hutson 2021; Sloane, Moss, and Chowdhury 2022; Birhane 2021b). Despite this debate, the field of affective computing continues to be “primarily driven by computer scientists and AI researchers who have remained agnostic to the controversies inherent in the underlying psychological theory. Instead, they have focused their efforts on the technical challenges of developing affect-sensitive computer interfaces” (Calvo and D'Mello 2010: 19).

However, to develop FER systems, practitioners must make assumptions about what counts as measurable affective behavior and what knowledge can be inferred from these measurements. This fundamental assumption shapes practitioners' knowledge claims (what are the intended uses of the system), data practices (what data types are needed and how they will be collected, cleaned, and annotated), model development (what key features will be selected), and accuracy metrics (how success will be defined). Common approaches to the measurement of affective behavior include categorical models based on FACS (which breaks down the emotional expression into discrete facial movements) and EMFACS (which assign discrete emotional labels to a limited set of facial movements), dimensional models (which classify emotional expressions based on their perceived positive or negative quality), as well as combinations of these methods.

In light of this apparent dissonance, this chapter maps participants' struggle to establish and maintain epistemic authority over their claims when the scientific validity of their methods (and the underlying assumptions) is challenged. As Gieryn noted:

When credibility is publicly contested, putatively factual explanations or predictions about nature do not move naked from lab or scientific journal into courtrooms, boardrooms, newsrooms, or living rooms. Rather, they are clothed in sometimes elaborate *representations* of science – compelling arguments for why science is uniquely best as a provider of trustworthy knowledge, and compelling narrations of why my science (but not theirs) is bona fide (Gieryn 1999: 4).

In surveying practitioners' representations and interpretations of science, the chapter maps the rhetorical practices of “boundary work” through which participants demarcate legitimate from illegitimate scientific knowledge and practice (Gieryn 1983b). Among earlier episodes of boundary work, historians of science have documented 19th century physiognomists and phrenologists' efforts to defend their practice as legitimate against accusations of ‘pseudoscience’ (Gieryn 1999; Shapin

1979). The rhetorical tactics adopted by this study's participants can be somehow read against these historical examples: in both cases, practitioners aspire to establish their practice as part of a broader project of social progress. Proponents of physiognomy and phrenology believed their knowledge could be used to improve society (encouraging the reproduction of those deemed 'fit', according to eugenic principles) and provide evidence for reforms in social policy (Pearl 2010; Thompson 2021). In a similar vein, participants interviewed here often viewed their work as part of a project of democratization of psychological and medical knowledge around affective behaviors and, more broadly, saw AI as a positive force for social change. However, whether FER systems (and affective computing at large) could provide genuine – let alone useful – insights into the affective life of users is precisely what is contested within the above-mentioned scholarly and public debates. This chapter's findings foreground the boundary work required for participants to validate the legitimacy of their claims and practice: how they establish their epistemic authority, how they define the ethical and professional limits of their practice, and what rhetorical strategies they devise to maintain credibility. Representations of legitimate scientific practice take here a fractal form: the act of demarcation is repeated at progressively smaller scales, carving out additional spaces of cognitive authority inside previously demarcated ones. In its final dimension, the space of legitimate affective computing knowledge and practice appears distant and removed from the space of ethics and social responsibility.

Expulsion: Truthful science and ill intents

The establishment of legitimate scientific knowledge involves the demarcation of *science* from *non-science*, as the institution of boundaries between the *inside* and the *outside* of what counts as science. This entails a process of *expulsion* of what is

perceived as illegitimate scientific practice and can take various forms. “When the goal is monopolization of professional authority and resources, boundary-work excludes rivals from within by defining them as outsiders with labels such as ‘pseudo’, ‘deviant’, or ‘amateur’” (Gieryn 1983: 792). To establish credibility, and distance themselves from accusations of pseudoscience, participants reported different rhetorical strategies of expulsion.

Marcus, a CEO and computer scientist with a background in electrical engineering, described to me his disappointment in how the affective computing community had recently been portrayed by a widely-cited report on the social implications of AI, which called for a ban on the use of affect recognition in high-stake decision-making processes and a moratorium on its development on the grounds of its contested scientific foundations. For Marcus, it is important to demarcate “good science” from improper scientific practices:

I get angry when I read reports saying that the affective computing community in particular is not doing proper science. I think that is entirely wrong. Yeah, and I think that betrays a lack of understanding of the field, to be honest. [...] there are lots of researchers who do really good research actually trying to solve the problem of, let's say, emotion recognition in itself and how to best interpret social signals in context, getting the right answers to be used in a real-world problem that involves emotion. And most of that is done properly. I'm sure there are some pieces of research out there that are not proper. But but that is the minority. And it is definitely not something that we approve of or indeed try to encourage.²⁶

Perhaps due to his dual role within both the industry and the academic sector, Marcus' response reflects a view of the field of affective computing as a cohesive community,

²⁶ Interview conducted on December 12th, 2020.

where those doing “improper science” are cast out as a rare “minority”. For Frankie, a professor in computer science, the outer limits of scientific knowledge are instead defined precisely by such professional identity. In particular, she dismissed commercial applications of FER as “snake oil”, unsupported by scientific evidence. By virtue of being profit-oriented, tech companies cannot be trusted:

Their assertions are unproven, and if you would ask to provide evidence for them, I don't think they could do so. Since it's commercial companies and it's unregulated, they can assert whatever they like. That sells stuff. Ok, so this is marketing, it's not science. There is science in there, but they're not being truthful about the science, I'm afraid [...] Some of them don't know that they're mistaken. While some of them know very well they're lying. Now, I put it as strongly as that, there are organizations selling this stuff and know that they're lying. We know that what they're saying and doing is not what they are actually doing, the marketing people will make the lies for you. The scientists who work for these companies will just keep their heads down and wish that they didn't do that. Because the scientific people know that these are lies.²⁷

Frankie's response expresses multiple forms of boundary work: first, it strongly demarcates “the marketing people” from scientists. Second, it attributes malicious intent to those excluded from the realm of science (“there are organizations selling this stuff and know that they're lying”, “the marketing people will make the lies for you”, “the scientific people know that these are lies”). This attribution of moral qualities further establishes the limits of science and is reflected in other participants' responses. Another example of this is Otis, a computational neuroscientist working as a customer manager for a major FER vendor. He helps clients who have purchased the licensed software to make sense of the results of its facial expression analysis. The majority of

²⁷ Interview conducted on November 24th, 2021

his clients are academic researchers, often using the FER module in combination with other physiological sensors (such as eye-tracking, ECG, and EEG) for medical or psychological research. Many of his clients however, he tells me, are marketing agencies who often use the FER module alone to gauge their customers' emotional responses to products and services without relying on multiple data channels. For Otis, the issue lies with the misuse of FER systems for unscientific purposes:

I think some are really aware of it and try to design studies that.. Whether study design is so good that the conclusion is quite easy.. But others just greatly ignore it, and I think especially in the commercial world where we get a lot of these requests, like 'I really want to measure the emotions of my consumers'. Well, that's not what you're measuring at all. You're measuring a facial expression. Uhm, so I think some researchers know what they're doing and put a lot of effort into this and others have just, you know, heard about it as a fancy thing. And they want this view into the mind of the respondent. [...] I think maybe that's also why neuromarketing has a bit of a reputation. Because you will see these gung-ho agencies promoting, you know, 'we can give you the emotional insights' and just bypassing this whole discussion.²⁸

Like in the previous response, the profit-oriented logics that govern the marketing sector are in fundamental opposition with what Otis considers a rigorous use of the API supported by scientific evidence. As he explains to me, "it all comes down to a really sort of nice and tidy study design".²⁹ Quentin, a CEO with both a medical background and a computer science training, reported similar sentiments, drawing a comparison between inappropriate uses of technology and pseudo-therapies that promise to treat diseases without legitimate scientific evidence:

²⁸ Interview conducted on June 14th, 2021

²⁹ Ibid.

I am worried as well, because people with the hype that goes around technology could end up damaging the reputation of using technologies by not doing things the right way. So, we have a good example in healthcare. We have people with these crazy cures for conditions and some of these.. I'm not saying all kinds of medications are bad. But there are some things that just don't work. There's no evidence behind them. And I think super important that we have an evidence-based approach to this technology because it has great power potentially. And if we don't have that approach of saying, okay, well, does it work? And how then there is a risk that either it won't be adopted to be in the health service where it could do tremendous good, or there's a risk that people will inadvertently cause great harm.³⁰

Emerging through these responses is the role that the scientific method, as a “rhetorical resource”, plays in boundary work (Derkatch 2008: 374). In asserting credibility over their claims, both industry and academic participants often invoked scientific “rigor” and “evidence-based” methods as rhetorical strategies to defend the boundaries of legitimate knowledge. This resonates with the “legitimation narratives” invoked by medical professionals to either validate or discount emergent and novel treatments (Perrotta and Geampana 2020). Finally, when accusations of pseudoscience are followed by calls for bans or moratoria on the development of FER systems (and affective computing technologies at large), the attribution of malignant intent moves beyond individual and professional responsibility, to social and political institutions. Cedric, an associate professor at a US-based computer science and engineering department, described:

I'm generally of the opinion that it's really difficult to contain technology. I don't think it's a path that will be productive, and I think it will only penalize the good people who listen if you try to do that. Whereas there are rogue governments.. bad governments or

³⁰ Interview conducted on February 16th, 2022

dictatorships, or bad people, or like people in in their basements, or just antisocial, and there are very angry people who will start to abuse things and so the only people you'll penalize are the good people who are gonna listen to you. I think instead that you have to create the possibility to use that technology for good so that the good outweighs the negatives, and that way you make that change that you want instead of just banning the entire technology.³¹

In the effort to “expel non-real members from their midst” (Gieryn 1995: 432), participants expressed the necessity to draw a moral boundary between “good” and “bad” actors, casting out from the domain of their practice bad scientists, greedy salesmen, and evil governments. The act of expulsion coincides here with the affirmation of technology as a positive social force and reflects the recent effort to rebrand AI as a tool for social good – the so-called field of “AI for Good” (Floridi et al. 2018). Cedric’s response, in particular, is exemplary of the libertarian techno-optimism that surrounds much of AI development. As many have noted, narratives centered on the supposed ability of AI to solve societal challenges, and on practitioners’ individual responsibility to make ethical choices, can often obfuscate existing power structures and perpetuate social harm (Madianou 2021; Cave et al. 2018; Tutton 2021). Finally, Cedric’s response serves a further purpose: by excluding malignant actors from the domain of legitimate scientific practice, it delineates a space of professional autonomy protected from external intervention. Overall, these responses seem to demarcate good and responsible knowledge production – backed by scientific evidence and robust methods – from abuses of science, across both academic and industry sectors. In the next section, I explore the differences (and similarities) between these two contexts and how participants mobilize professional boundaries to establish the validity of their claims.

³¹ Interview conducted on November 23rd, 2021

Autonomy: Agnosticism and ambiguity

To defend the scientific integrity of their claims against accusations of ‘pseudo-science’, participants offered competing characterizations of their practice. This type of boundary work – the protection of autonomy – involves creating boundaries to protect professional autonomy and authority over scientific claims from other interest groups (such as other professional domains, government bodies, civil society, and political groups) (Jasanoff 1987). This section surveys the tensions emerging from developments of FER systems in both the academic and private sectors and maps differences and similarities between the two groups’ responses.

Applied science

Faced with the ongoing criticism surrounding the validity of adopted models of emotions, some industry participants characterized their practice as largely peripheral to theoretical debates and held instead a pragmatist view of their practice as mostly concerned with development of applicable and useful technology. By removing themselves from the dispute over the validity of the theories and methods they adopted, participants aimed at protecting “their right to determine and perform the work as they see fit” (Wulff and Finnestrand 2022: 711). This lateral move serves two distinct purposes: first, it provides immunity from the particular interests involved in the dispute over the definition of emotions, seen as a conflictual ground for the establishment of a research agenda that sees emotional expressions as social and cultural constructs or (*à la* Ekman) as universal and biological traits. Second, it posits

participants as temporarily *external* to science, only tasked with the advancement of practical knowledge. This re-shaping of participants' practice into technical knowledge (Porter 2009) forecloses critiques of FER technology grounded on its contested scientific foundations. This is illustrated in Marcus' response:

We don't do emotions. Well, that's not true. We do a bit of emotion. But emotion is an incredibly ill-defined concept and I, for one, am not going to take any stances in the *emotion theory war* [emphasis added] that's currently ongoing because I think taking a stance is inherently unhelpful. We need science to answer these questions, but what we need is great is good questions and good science to answer those questions.³²

As mentioned above, this agnosticism is mirrored in participants' characterizations of their practice as purely technical. I asked Dennis – a CEO with an electrical engineering background – to describe the forms of expertise and competencies involved in the design and development of his FER system: “It’s all technical skills. We did not get into the psychological or sociological side of ‘what to do with these signals.’ I don’t think it concerns us. We offer a tool that could be used to gauge psychological data, but I didn’t want to be specific about what it should be used for.”³³ This response differs from the ones provided in the previous section in that it does not seek validation from science. Here, instead, it is precisely autonomy from science (here intended as psychology or emotion science) that allows participants to focus on the development of technical tools that can be applied in multiple domains. STS scholarship has pointed to this continuous crossing of boundaries between the scientific and the technical as characteristic of engineering practice (as the denomination *applied science* seems to suggest), where

³² Interview conducted on December 12th, 2020.

³³ Interview conducted on April 29th, 2021. Translated from Italian.

the production of knowledge “appear[s] to occupy a double location inside and outside of science” (Downey and Lucena 2001: 167). As Dennis continued,

I do not take sides. Yes, there is a debate: some people support one thesis, others support the opposite one. I don't think it concerns us. My goal is to entertain, or to build something useful.. If our partners find [our technology] useful... That's why I did not look into psychological evaluations or anything similar. [...] I use AI because otherwise I could not obtain these results. And if this is scientifically rigorous or not it's not up to me to determine, because I am not a scientist, I am a technologist. A product developer. I am not a scientist.³⁴

Moreover, as Downey and Lucena have noted, engineers often have to navigate the ambiguities associated with being an expression of capital and corporate power (Ibid.). Theoretical agnosticism offers a strategy to navigate such ambiguities. Marcelo, an affective computing researcher working at a large tech corporation, mentioned the evolving nature of scientific theories of emotion as a reason for his company's noncommittal to any conceptual model:

Some of the steps that we have taken, for example, is reframing how the API is being put on the website. So, you will see that there is a lot of emphasis on saying that this is perceived emotional states, that it doesn't follow any specific emotional theory because they keep evolving and it's an evolving field. And we don't take sides. We just say what we do with it.³⁵

³⁴ Ibid.

³⁵ Interview conducted on June 17th, 2021.

This form of agnosticism seems to allow industry participants to evade conversations around social responsibility and, as reflected in this last response, to defer accountability to their technology's consumers.

Basic research

Within the academic sector, participants navigated the ambiguities associated with engineering work by demarcating their practice – the creation of novel machine learning techniques or datasets – from the commercial development of affective computing applications. For Julius, a professor at a European engineering and computer science department, engineering is about “creating problems” that others – industry practitioners – can act on to develop real-world applications: “As engineers, we create problems. We don’t necessarily have an application in mind. Although, many have used and implemented our algorithms for their own needs”.³⁶ Autonomy from specific application contexts is necessary for academic engineers to advance state-of-the-art algorithmic techniques. “Academic research”, Julius reasons, “is much more advanced than current commercial products, especially when it comes to computer vision and machine learning.”³⁷ While tech companies have perfected the production of cheap and reliable sensors (such as Bluetooth, microphones, and cameras), according to Julius, academic research remains the driving force behind the development of machine learning techniques. This perspective conflicts, however, with scholarly work pointing to the growing concentration of power (financial and computational resources) in the hands of a few large tech firms. Over the past decade, “the rhetoric and capital flowing from these firms [has] served to redefine the AI research field, flooding it with funding and focusing the field’s attention on data and compute-intensive techniques and research questions” (Whittaker 2021: 52).

³⁶ Interview conducted on October 21st, 2021. Translated from Italian.

³⁷ Ibid.

Moreover, “dual-affiliations” of AI scholars across the industry and academic sector and PhD programs sponsored by tech corporations further blur the boundaries between the two sectors. Perhaps contributing to Julius’ perception of academic vanguard is the race to publish that governs university spaces. As Cedric describes to me:

In research, what we look at is the sophistication of the approach. If that approaches is novel and it's working well in comparison to some other published work. If you can outperform them, it's good. If not, you can't be much lower than them because that means that the model isn't working well. So, it's kind of always like a race between different research groups trying to get results that are slightly better than the other. But then, while you're trying to improve the result, you also want to showcase a new method that you've come up with and innovated on.³⁸

The field of computer science, and machine learning in particular, currently prioritizes a conference publication model with shorter rounds of reviews and lesser chances for authors to improve their research before acceptance or rejection (Bengio 2020). This ever-growing pressure to publish (and publish quickly) shapes participants’ practice and how they conceptualize notions of accuracy and validity in affective computing. Reflecting on his students’ experience, Julius later describes:

If my students want to publish, they don’t need to work with a psychologist or a sociologist. They only need to be able to say that their facial expression recognition is 0.5% more accurate than others. That’s sufficient to publish a paper. If that’s useful... Well, that’s a different question.³⁹

³⁸ Interview conducted on November 23rd, 2021.

³⁹ Interview conducted on October 21st, 2021. Translated from Italian.

The usefulness of a model – its *utility* – emerged here as a demarcating principle to position real-world applications of FER models within or beyond the limits of legitimate knowledge. This rhetorical move rests on the old adagio that “all models are wrong, but some are useful”. As Erik – a PhD student in social artificial intelligence – described, while *applied science* aims to develop systems that might not be entirely accurate but are nonetheless useful, *pure science* aspires instead to produce *true* knowledge; therefore, its models will be likely useless:

It comes down to this for me at least, it's thinking about that the truth is somewhere here in the middle and we can approach the truth from below by fundamental research which we know for a fact that things are true. So, we know they're true, but it's not the entire picture but this tiny, tiny little bit is true and in itself it is not useful. Or we can do the engineering problem which is what's happening a lot of the time is we could make very broad assumptions and we could get models that work and then we can sort of approach what the truth is from above where we're saying we're making these very broad claims and then we can maybe refine them a little bit and the truth in the middle and we need people working from both sides to reach the truth. [...] And I think the aspect here is saying we might make very useful models when approaching things from above, but we might not make very true models.⁴⁰

Erik's lab has made important progress in demonstrating that facial emotional expressions are culturally specific and challenged mainstream theories of basic and universal emotions. The team is now working on a novel framework for the identification and generation of facial affective expressions which they hope will help model emotions in robots and digital agents. This work, however, is based on an experiment that does not replicate the real-world conditions of human emotion

⁴⁰ Interview conducted on August 18th, 2022.

recognition, as participants were asked to assign emotion categories to computer-generated images of facial expressions. As he explained to me, these images are “floating heads on a blue background. They have no hair, they have no beard, they have no piercings, they have no makeup. There is no information about the situation they are in. They’re all lit the same way. That’s not how we do emotion inference in the real world”.⁴¹ In this sense, the results of this study are “true”, but they are not generalizable to the real world. Here, Erik further demarcates fundamental research (“approaching the truth from *below*”) from engineering (“*approaching the truth from above*”), as the latter tolerates assumptions about the real world in order to produce useful models. This distinction has implications for practitioners’ social and ethical responsibility:

The scientific process is very rigorous. And because of the type of work we're doing, we're not really claiming that we're making any massive leaps with regards to application. So, it's not that ethics are not at the heart of it or not at least thought about, discussed, but because it is so far away from application, the immediate... The challenges that someone delivering a product has are vastly different from someone doing fundamental research. Right? People who are looking at how two electrons interact have a very different ethical responsibility than someone actively building a nuclear bomb. I'm not explicitly making a link between nuclear bomb and emotional recognition. So, I would say the work that we are doing is much closer to what those looking at electrons are doing. So, it's not to say that there are no ethical implications of it, but it is not as essential as something a little bit more applied.

Approaches to affective computing framed as fundamental or foundational mirror broader claims of neutrality and objectivity that are typical of the machine learning field. When Jeff Brantingham (the Pentagon-funded anthropology professor who

⁴¹ Ibid.

patented PredPol and co-founded the for-profit company named after it) presented a prediction model for gang-related crime (Seo et al. 2018) at the 2018 AIES conference, he was prompted by the audience to consider the potential ethical and societal implications of his work – to which he replied, “It’s basic research” (Hutson 2018). As some have noted, this profession of agnosticism reflects the emphasis that computer science education puts on the principle of abstraction, which “construct programming, and ‘algorithmic’ or ‘computational’ thinking more generally, as epistemically and politically neutral tools that only take political meaning once they are contextualized through the application” (Malazita and Resetar 2019: 301). In a similar vein, the rigor of the scientific method and the perceived distance from real world applications shield academic participants from social and ethical responsibility.

Expansion: The magpie effect

A final form of boundary work takes place when “insiders-scientists” (Gieryn 1995: 394) seek to expand the borders of their “cultural authority” (429) into other knowledge domains. One historical example of this, according to Gieryn, is the expansion of philosophy in the 18th century into domains traditionally within the purview of theology: namely, ethics and spirituality. Here, this type of boundary work is, for some participants, a distinctive tract of practitioners working in AI. As Frankie described:

I’ve often characterized our relationship as being a bit like magpies to other disciplines. We ‘flap, flap, flap, flap [*makes a wing-flapping gesture with her hand*]. We need a

theory, we need a theory, we need a theory. Oh, here's a shiny theory. This one's easy to implement. We'll have that one.' Yeah. Now, that's not a very responsible attitude.⁴²

More concretely, this is reflected in how Julius described the application of Ekman's theory in his own work. Julius has been part of the affective computing community for over two decades and, among his most cited work, is a system for the classification of facial expressions in video sequences that he developed in the early 2000s. Inspired by the Facial Action Coding System (FACS), Julius reduced Ekman's 46 action units to only 12. This method aimed to expedite the annotation process by reducing the number of landmark facial features that needed manual labelling. Talking about this work, Julius described:

Of course, theory matters. It can give us inspiration. Then, from a practical point of view, we see what we can do with it. A major problem with emotion theory is that there is no agreement within the emotion science community. There's not even agreement on whether one theory is better than the other! [...] But something I've already said before is that our work, as technologists, is independent from that.⁴³

As previously suggested by Frankie's response, practitioners working in affective computing treat psychological theories as props for the development of technical systems. A similar example is the application of affective computing methods to other psychological theories. Daniel, a CEO whose primary clients are marketing agencies, described his plans to develop a FER system to predict personality types. This model, he reasoned, would combine action unit detection (the detection of individual facial muscles based on Ekman's taxonomy) with the Sixteen Personality Factor model, also

⁴² Interview conducted on November 24th, 2021.

⁴³ Ibid.

known in the psychological literature as 16PF. Developed as a “comprehensive measure of normal-range personality” (Cattell and Mead 2008), the 16PF questionnaire was created by Heather Cattell, daughter of British psychologist and statistician Raymond Cattell. Raymond is infamously known for his views on intelligence, eugenics and race, and for adhering to fascist and antisemitic causes – to the extent that, in 1997, the American Psychological Association revoked his Gold Medal Award (Tucker 2009). Cattell’s 16PF, a variant of the Big Five types model, can be seen as the personality equivalent of Ekman’s Six Basic Emotions (or Basic Emotion Theory): they are both expression of a longstanding effort to anchor psychological and subjective experience to quantifiable truths (Gould 1996).

Daniel is familiar with the 16PF, as it is often used to test a person’s cognitive abilities in clinical, educational, and workplace settings. The relationship between 16PF and FACS, however, is a concept explored in a 2017 machine learning paper, which Daniel promptly sent me after our interview as proof of his project’s scientific integrity. The authors of the paper, “*Predicting the Sixteen Personality Factors (16PF) of an individual by analyzing facial features*”, propose a deep learning model that “computes the personality traits in less than 1 minute and can be used to monitor the personality traits of an individual in real time” (Gavrilescu and Vizireanu 2017: 2). To support such claims, the authors must make a few fundamental assumptions: That emotions are universally and reliably expressed through facial patterns; that FACS, among other methods, is best at predicting genuine emotions; that 16PF is a reliable test for the prediction of personality types (and that such types are a stable and consistent category); and, finally, that “there is a close relationship between personality traits and how emotions are expressed” (Ibid.). This paper is not an isolated exception but, indeed, one of the many publications in machine learning-related journals attempting to compute personality or character traits based on physiological attributes (Wache et al. 2015; Subramanian et al. 2018; Setyadi, Harsono, and Wasista 2015;

Gavrilescu and Vizireanu 2018; Zhang et al. 2017). As explored in the next section, the validity of these claims is typically assessed through statistical measures aimed at determining whether the systems can perform reliable predictions over time (that is, assessing whether systems work as intended) (Jacobs and Wallach 2021). However, many have pointed to the field's lack of reflexivity and seeming inability to question the epistemological and theoretical assumptions that these systems are grounded on (such as the existence of universal personality types or basic emotions) (Sloane, Moss, and Chowdhury 2022; Birhane and Guest 2020). As Frankie reminded me,

The unethical use of technology by making unsubstantiated claims is a general problem of AI, and it's an even bigger problem in affective computing because the claims are bigger, and they really cannot be sustained in most situations.⁴⁴

Perhaps contributing to this problem is the “looping effect” (Vertesi 2020) that the above-mentioned publication practices can have on the field. As some have noted, hype around commercial applications of AI “can spill over into machine learning-based science, leading to overoptimism about their performance. Non-replicable findings are cited more than replicable ones, which can result in feedback loops of overoptimism in machine learning-based science” (Kapoor and Narayanan 2022: 9). Finally, compounding this, is the entrenchment of these epistemological assumptions within larger socio-technical systems. An example of this is the incorporation of FER systems into other disciplines' research process and, in particular, into medical research and healthcare provision. When I first spoke to Marcus in 2020, he described a vision of his technology – a proprietary FER system for the detection of mental health conditions and neurodevelopmental disorders – not only as vehicle for collaboration with other forms of knowledge, but also for their own development:

⁴⁴ Interview conducted on November 24th, 2021.

I think that what we need at the moment is collaboration between neuroscientists, technologists like machine learners, computer vision researchers, humanists... Yeah, and philosophers of science as well. Building and studying together technology that's actually used and looking at how users use it. [...] There needs to be a community that keeps talking to each other regularly. We're working on that. This is a tough sell. We as [company's name], definitely want to play our part in that. So, we definitely want to contribute. For example, we want our technology to be used by other to test things and run such studies.⁴⁵

Two years after this interview, Marcus' company's website lists – among other solutions – the possibility to use his technology to evaluate mental health treatments in clinical trials, providing “reliable and objective measures” of mental health conditions “in a cost-effective way”. This form of “entrenchment” (Collingridge 1979) further complicates the opacity of these systems. What happens when the assumptions underlying FER models ‘disappear’ within larger socio-technical contexts? These concerns, which will be explored more in-depth in chapter 7, resonate with the vast scholarship on algorithmic injustice that in recent years has pointed to the political role that AI and machine learning practitioners play in altering the fabric of society (Buolamwini and Gebru 2018; Raji et al. 2022; Hutchinson et al. 2021; Mitchell et al. 2019; Costanza-Chock 2018; Birhane 2021b; Whittaker 2021) and automating bias and harm (Abdurahman 2022; Benjamin 2019; Eubanks 2018).

⁴⁵ Interview conducted on December 12th, 2020.

Sidestepping ground truths

Allow me to return to Frankie's last response:

The unethical use of technology by making unsubstantiated claims is a general problem of AI, and it's an even bigger problem in affective computing because the claims are bigger, and they really cannot be sustained in most situations.⁴⁶

As some have noted, the “unverifiability, ambiguity, and indeterminacy of possibly incorrect results” is often used to legitimize opaque, uncertain, or contested applications of machine learning (Grill 2022: 15). Cloaked in the objective confidence of numbers and percentages, accuracy metrics are announced and circulated to ensure trust in algorithmic results (D'Ignazio and Klein 2020; Beer 2016). In machine learning, accuracy usually describes how well (how often) the prediction algorithm produces results that match the test data (Kelleher 2019). Louise Amoore, however, has noted how evaluations of performance often boil down to the question: “Is it good enough?”, where a good enough prediction is “one that achieves some level of optimization in the relationship between a given target and the actual output of a model” (Amoore 2020: 67). This became evident in participants' articulations of accuracy in their own work. When asked what counts as a good classification (what is a “good enough” classification), Julius explained to me that accurate and “fine-grained” predictions of emotional states are not necessary if he can successfully predict indicators of people's preferences of commercial products:

It is not important whether I can accurately differentiate between angry and disgust as long as I can differentiate between overall negative and positive expressions. Especially

⁴⁶ Interview conducted on November 24th, 2021.

if you need these predictions to determine people's preferences. For example, you are choosing between products and you say 'I like this, I don't like this.' I don't need a fine-grained emotion recognition system for that.⁴⁷

Erik described to me a similar scenario, where a FER system could be used to track changes in emotional behavior to detect the onset of depression:

If I then speak to this person after my tool has spit out a number, saying they are 80% likely to have depression, I can take this person to see a doctor and if this doctor says, 'you know what? They're depressed.' Then I have one for accuracy, right? That's easy. You can imagine that in many different ways. It doesn't necessarily have to be emotion, that's just accuracy. In general. If we have very clear goals of what we need to define as a count for a hit and a count for a miss, then that's easy. General purpose emotion recognition, I think doesn't have a good sense of accuracy, right? And this comes down to the idea that we don't really know what emotions are.⁴⁸

This understanding of accuracy ("A count for a hit, a count for a miss") circumvents many of the issues associated with creating ground truths in affective computing. Indeed, practitioners often rely on emotion data collected in laboratory settings – where emotional expressions are posed and exaggerated – as well as on observers' interpretation of emotional expression. As Frankie explains, a focus on behavioral changes and their relevance to the context of application sidesteps the need for ground truths:

You may not need to create ground truth. it depends what you're doing. So, if we recognize enough of the negative behavior or negative feelings about students to improve

⁴⁷ Interview conducted on October 21st, 2021. Translated from Italian.

⁴⁸ Interview conducted on August 18th, 2022.

the interaction, that was good enough even if we weren't always right. [...] I think establishing ground truth is probably a bit of the will of the wisp because, you know, do people know how they feel? and I think emotion is a moving target.⁴⁹

According to this reframing, “it is not the accuracy of the algorithm that matters so much as sufficient proximity to a target. Put another way, the algorithm is good enough when it generates an output that makes an optimal decision possible” (Amoore 2020: 67). With these considerations, the ambiguous nature of participants’ practice comes full circle, as they provide justifications for their work that are seemingly capable of evading questions of scientific validity. This is exemplified in both Marcus’ and Dennis’ responses:

We normally look at downstream use. So, how is the app actually used and what does that, what changes does that bring? If you think about behavior change, for example, getting to a healthier lifestyle... *I couldn't care less how accurate my action unit detection is, as long as it creates the desired behavior* [emphasis added]. Behavior change is easier to collect and is an objective measure. And so that's a downstream accuracy that we are really interested in.⁵⁰

Things can be useful even if they are not based on a rigorous scientific model. A system can be useful, especially if it's used in an entertainment context, not a medical one. *It can be useful even if it's not based on scientific, provable, and objective principles* [emphasis added]. If I'm browsing an e-commerce and I like a product more than others, if the system's prediction about my preferences is 70% correct, that's sufficient to make a useful and valuable system. Obviously, if the outcome of a surgical procedure depends

⁴⁹ Interview conducted on November 24th, 2021.

⁵⁰ Interview conducted on December 10th, 2020.

on that 70%, three people out of every ten patients will die. It depends on what the system's use is.⁵¹

Here, usefulness replaces the need to ground participants' practice in scientific knowledge. Borrowing from Crawford and Campolo, the system's predictions are valid as long as they are *consistent* with the adopted emotion theories: "the consistency of uninterpretable correlations replaces a causal epistemology or theoretical explanation [...] and is deployed retroactively to justify the methodological choices" (Campolo and Crawford 2020: 12). Finally, these responses can help re-orient and align critical analyses of affective computing systems with the vast scholarship on algorithmic harm associated with predictive technologies. Indeed, the discursive and practical shift towards behaviors situates FER systems within the larger political economy of AI systems and platform technologies. In defining the value of affective computing applications, Alex described:

What is valuable is the behavior, human behavior. And you see that a lot on social media platforms and different socializing platform, professional platforms as well. So, what people write there, what people do, where they click. So, big companies like Facebook. Obviously, Facebook has monetized this in a huge way. So, I would say behavioral understanding is more important than actual emotions, because that's what is monetizable for big companies rather than actual real emotions, because they're not interested in knowing your real emotions. The people who are interested in knowing your emotions are people who are close to you. Other people are just interested in how you behave. So, I think those kinds of applications will not have a value over a long time.

⁵¹ Interview conducted on April 29th, 2021.

Emerging from this response is a view of affect that is less concerned with the emotional and mental content of human experience than with the ability to extract value from behavioral changes.

Conclusion

This chapter has drawn attention to the role that articulations and mobilizations of science play in asserting participant's credibility. In particular, I have explored the discursive practices through which participants strive to establish credibility when the validity of their claims is publicly contested. Rhetorical practices include forms of boundary work aimed at legitimizing affective computing knowledge and affective computing practice simultaneously inside and outside the space of *science*. Participants did not hold stable conceptions of their practice, but instead offered articulations of their work as ambiguous and, at time, fluid, crossing the boundaries between *science* and *non-science*, and *pure science* and *applied science*. While this is in part due to the heterogeneous composition of the sample here interviewed, these responses can also be seen as a reflection of what Downey and Lucena called "code-switching", a "situationally-specific movement" between different "codes of meaning" that is typical of engineers' knowledge and professional identity (Downey and Lucena 2004). Then, I explored practical ways in which participants validate their systems' predictions. Since algorithmic performance cannot be evaluated in relation to inner subjectivity (the 'truth' of emotional experience), what emerges from these responses is a shift towards behavioral understanding that situates FER systems within broader critical analyses of predictive technologies and algorithmic harm. By unearthing the discursive and material strategies that participants adopt to evade criticism and establish credibility, this chapter has contributed to the "uninventing of accuracy" (MacKenzie 1993), as in, contesting the certainty (Grill 2022) and credibility of their claims.

While here I have focused on participant's discursive practice, the next chapter (chapter 6) takes a deeper look at the material structures and practices of FER design. In particular, I discuss the role that data annotation (and different annotation schemes) play in the design and development of FER systems. Common approaches to emotion AI design involve the annotation of Facial Action Units (or FACS coding), and Valence, Arousal and Dominance (VAD) annotation. While FACS coding is performed by "expert" coders, trained in the "objective and reliable" annotation of facial micro-expressions, VAD is performed by "lay annotators", who assign a set of "more subjective" metrics (perceived positivity/negativity of feelings, intensity of emotional expressions, and control over emotional experience) to the images and videos they annotate. Here, I discuss how coding schemes and organizational dynamics shape data annotators' ways of 'seeing' affective data and bring to the fore the interpretive element of data annotation.

Chapter 6 | Seeing affect

Introduction

“We never look at just one thing; we are always looking at the relation between things and ourselves. Our vision is continually active, continually moving, continually holding things in a circle around itself, constituting what is present to us as we are.”
(Berger 1990: 9)

“The ‘eyes’ made available in modern technological sciences shatter any idea of passive vision; these prosthetic devices show us that all eyes, including our own organic ones, are active perceptual systems, building on translations and specific *ways of seeing*, that is, ways of life.”
(Haraway 1988: 583)

The previous chapter investigated practitioners’ representations of affective computing practice as efforts to establish and maintain epistemic authority. Here, I look at the data work that informs FER design and, in particular, the role that different data structures and practices play in constructing claims of objectivity and accuracy. This chapter focuses on two common annotation schemes used to label training data for FER algorithms: the Facial Action Coding Scheme (FACS) and the Valence, Arousal, Dominance (VAD) model. FACS is considered to be an anatomically based, “objective” measure of facial expressions, as it breaks down facial muscle activity into discrete micro-movements called “Action Units” (AUs) such as the movement of the corners of the mouth, the corners of the eyebrows, or the outline of the nose. As discussed in the chapter, coders trained in FACS are considered “expert coders” by virtue of their training and the perceived complexity of the annotation task. VAD coders, on the other

hand, are often referred to as “lay annotators”, as this type of coding requires minimal, informal training. Valence, arousal, and dominance are defined as continuous dimensions of facial expressions, indicating their perceived pleasantness, intensity, and degree of control of the emotion. As the chapter will show, the definition of these terms is in itself subject to interpretation and negotiation among annotators. Because it relies on judgments that often escape quantification efforts, VAD is considered to be a more “subjective” form of annotation that aims to incorporate human perception of emotional expressions into FER systems.

In investigating annotators’ experiences of FACS and VAD coding – the ways they ‘see’ and interpret data – this chapter focuses on the interplay of vision, expertise, and power in machine learning applications of affective computing. First, I discuss how the definition and attribution of skill and expertise in data structures reflects broader power asymmetries in machine learning design, maintained and reinforced through uneven labor dynamics. Despite being central to the maintenance and development of machine learning models, the lay annotation of training data is often considered to be low-level and unskilled, as opposed to the high-level work of domain experts. As I explain in the chapter, the annotation of large datasets is often outsourced; annotators’ feedback and concerns are rarely taken into consideration in the model design process. These dynamics are, in some ways, reflected in affective data work consisting of, on one hand, “expert” and objective FACS coding and, on the other hand, “lay” and subjective VAD annotation. First, I show how the practice of FACS coding, despite its purported objectivity and scientific rigor, is itself subject to a process of negotiation and interpretation of the data. I borrow from anthropological studies of vision and professional practice (Grasseni 2004; Goodwin 1994; Ueno 2000) to shed light on the process of “enskilment” (Grasseni 2004) that both FACS and VAD coders must undergo to learn how to ‘see’ data in a certain way. This analysis has two objectives: first, by looking at both FACS and VAD as “ecologies of vision and imagination” (Ibid.)

with their respective norms, grids, and local negotiation practices, I show the subjective and interpretive element that underpin FER projects. Second, contrary to practitioners' perception of VAD annotation as low-level and unskilled, I bring to the fore the experience and expertise of VAD data workers that is essential to affective computing projects. Through this lens, VAD annotators hold forms of specialized knowledge – "skilled vision" (Ibid.) – similar to those of FACS coders. Here, I also discuss the role that material factors, such as annotation interfaces and automation tools that aim to tame the subjective element of VAD annotation, play in guiding annotators' ways of seeing and interpreting the data. These techniques, coupled with practitioners' perception and representation of annotation practices, contribute to obscuring the work of interpretation required to make sense of affective data. The meaning of vision becomes here twofold: vision refers both to the way annotators 'see' the data and to how practitioners 'see' (perceive and represent) data work. In particular, I describe how practitioners perpetuate perceptions of data work and annotation as either objective and scientific or subjective and intuitive, thus overlooking (and at times concealing) the human work of negotiation and interpretation that underpin both coding practices. This is not, however, a unique feature of affective data. As scholarship in this area has shown, the perceived intuitive nature of annotation tasks often results in labor structures that posit annotators as interchangeable and disenfranchised from the overall system design. For instance, Emily Denton et al. have traced the norms and assumptions that underpin ImageNet, a large computer vision dataset that include 1,281,167 training images spanning 1000 object categories,⁵² labelled by crowdsourced data workers 'Amazon Mechanical Turkers'. As the authors have noted,

Framing the label verification as an act that requires little reflective judgement not only suggests that anyone can participate, but that annotators are interchangeable because

⁵² <https://www.image-net.org/about.php>

they share the same innate faculty of seeing objects and because they exercise vision in the same way (Denton et al. 2021: 10).

Therefore, notions of expertise and skilled work become here a crucial site for the re-definition, re-distribution, or conversely concentration of power in machine learning design. Finally, I conclude with a reflection on the role that the situated data work that contributes to the creation of machine learning ground truths plays within broader data structures, and how experiences of data work and annotation can be leveraged to challenge claims of objectivity on which machine learning models rest. Building on critical scholarship on data work, I argue that to make data work visible, it is necessary to accept the fundamental uncertainty that permeates the creation of ground truth datasets.

Skill, expertise, and power in data structures

In 2016, at the Machine Learning and Market for Intelligence Conference in Toronto, the host asked deep learning “godfather”, Geoffrey Hinton, what new developments in the field the audience should prepare for. Famously, Hinton replied:

Let me start by just saying a few things that seem obvious. I think that if you work as a radiologist, you are like the coyote that’s already over the edge of the cliff but hasn’t yet looked down, so he doesn’t realise there’s no ground underneath him. *People should stop training radiologists now*. It’s just completely obvious that, in five years, deep learning is going to do better than radiologists. [...] I said this at a hospital, and it didn’t go down very well [*audience laughs*]. [...] What I think of that now is *business as usual*: take any old problem where you have to predict something and you have a lot of data, and deep learning is probably going to make it work better than existing techniques.⁵³

⁵³ Creative Destruction Lab, 2016, “Geoff Hinton: On Radiology”, 24th November 2016, Available at: https://www.youtube.com/watch?v=2HMPRXstSvQ&ab_channel=CreativeDestructionLab [Accessed 15th September 2022].

As many have noted, Hinton’s predictions did not stand the test of time. Less than five years later, researchers reported “critical failures” in machine learning-powered medical imaging (Oakden-Rayner et al. 2019), and quality assessments of diagnostic accuracy revealed that AI systems fail to replace radiologists’ precision and, particularly, radiologists’ “double reading” in screening programs. (Freeman et al. 2021). Hinton’s overconfidence relied on what researchers have called the “central dogma” of deep learning: the idea that every problem can be solved by collecting large quantities of data, and that the same neural networks can be applied to different problems (Narayanan and Kapoor 2022). Slota et al. have called this tendency of data scientists to make disparate data and knowledge amenable to machine learning techniques “prospecting” (Slota et al. 2020). Prospecting requires that machine learning models be flexible and generalizable, irrespective of domains of application. Here, the “‘domain-agnostic’ positionality of data science” (2) recalls the *magpie effect* discussed in the previous chapter. Domain experts – such as radiologists – are seen as superfluous data workers that can be disposed of once they have completed the annotation of ground-truth datasets necessary to train deep learning algorithms.

This disregard of domain expertise in machine learning is considered to be central to many of the socio-technical failures of automated systems (Sambasivan and Veeraraghavan 2022; Narayanan and Kapoor 2022; Ribes et al. 2019). One reason for this, researchers in computer science have observed, is the systematic “deskilling” of domain experts: rather than essential, expert stakeholders, radiologists are reduced to data workers, that is, to data collectors, cleaners, and annotators (Sambasivan and Veeraraghavan 2022). Implicit in these concerns is the idea that “data wrangling” (the process of collecting, standardizing, and annotating the data) is a menial and low-level task. Indeed, this aspect of the development and maintenance of machine learning is often overlooked by practitioners, as they consider this type of work to be tedious, time-consuming, uncreative, and, importantly, *unskilled* (Sambasivan and Veeraraghavan

2022; Kandel et al. 2012; Irani 2015b). Data workers are here seen as mere computational resources, without expertise of their own. However, data annotation plays a central role in data-driven systems, with many machine learning models requiring manual labelling of training datasets. As work in “critical dataset studies” (Thylstrup 2022) has shown, datasets “form a foundational element of machine learning cultures” (656): training data not only shapes model performance, it also reproduces epistemological assumptions about what does and does not count as valid knowledge. For example, in computer vision, datasets govern the way machine learning models ‘see’ the world. As we increasingly embed these systems into our social life, the project of interpreting and labelling images is political, rather than merely technical (Crawford and Paglen 2021). One example of this is the categorization of race and gender in face datasets. Research in this area has found that assumptions about gender and race are rarely questioned and labels are assigned simply on the basis of visible and physical appearance. As many have noted, however, identity is a socio-historical concept, rather than a merely physical one (Scheuerman et al. 2020).

Therefore, the interpretation and categorization of data yields the power to make decisions about the system’s outputs and, ultimately, to shape societal outcomes. Literature that explores data labelling methods, however, often focuses on the annotators’ role and responsibility in reproducing bias, and points to the issues of data quality and reliability that can arise from inconsistencies in annotation work or annotators’ biased perceptions of the data (Miceli, Posada, and Yang 2021). On the other hand, a growing body of work has emphasized how the power structures that govern data practices within machine learning design pipelines can influence and shape data annotation (Miceli, Schuessler, and Yang 2020; Miceli et al. 2021; Miceli, Posada, and Yang 2021). Notably, ethnographic research on data work has pointed to the global labor dynamics that underpin many commercial machine learning models: workers, often outsourced from low-income regions or vulnerable populations, clean and label

the image, video, text, and sound data used to train and maintain learning algorithms (Irani 2015; Irani and Silberman 2013; Gray and Suri 2019). The “ghost work” behind much of today’s automation (Gray and Suri 2019) is often recruited through crowdsourcing platforms (such as Amazon Mechanical Turk) or impact-sourcing annotation companies that employ workers from the Global South to offer data annotation services at competitive prices to practitioners and researchers largely based in the Global North. Labour dynamics and structures within these platforms and organizations can reinforce the power asymmetries present in the design pipeline: the social and material conditions of data work – the vulnerable status of workers and the complete alienation from their employer (and the purpose of their tasks) – might further infuse the annotation process. Miceli and her colleagues have noted how annotation standards (the requirement and expectations of clients or managers), multiple layers of power (client, team leaders, reviewers, and annotators), and the naturalization of annotation practices (the idea that labels are “self-evident”), can all shape the interpretation of data. This “imposition of meaning” (Miceli, Schuessler, and Yang 2020) has the potential to reinforce normative assumptions around the data and, at the same time, forestall opportunities to challenge the overall system design. Mirroring the concerns mentioned at the beginning of this section, data workers (collectors, cleaners, and annotators) are not seen as relevant stakeholders in the design pipeline.

These knowledge infrastructures – with high-level domain experts and “lay”, low-level data workers – are reflected at a smaller scale in my participants’ experience of data work. However, as the next sections will show, both groups face challenges arising from the ambiguous and highly contextual nature of affective data. To make sense of this data, both FACS and VAD annotators must negotiate and interpret the meaning of what they are observing. Shaping their visions are, however, the norms and practices associated with their respective coding schemes, the material tools and

interfaces that mediate the annotation, organizational structures and workflows, as well as practitioners' perceptions and representations of data work. By unearthing, on one hand, the tensions and disagreements that underpin the annotation of affective data and, on the other hand, the norms and structures that guide annotators' practice, I hope to show the fundamental ambiguity on which FER projects rest. While the inaccurate – and perhaps impossible – annotation of facial expressions might seem innocuous when compared to the annotation of identity attributes such as race and gender, I argue that annotators' disagreements in affective computing have indeed a political role. Brought to the fore, annotators' disparate ways of 'seeing' could counteract machine learning's broader claims of objectivity and universality.

Contested visions

As mentioned, these configurations of knowledge, expertise, and power are reflected in affective data work and, particularly, in the two annotation schemes that inform the development of FER systems: the Facial Action Coding Scheme (FACS) and the Valence, Arousal, and Dominance (VAD) model. As detailed below, while FACS maps combinations of micro-movements supposed to make up basic emotional expressions, VAD focuses on the dimensional qualities of affective expressions, such as perceived intensity and pleasantness of emotional expressions. These two annotator groups are usually considered to be respectively expert and skilled (FACS) and "lay" and unskilled (VAD). The next two sections focus on the experiences of two coders: Isabella, a psychologist certified in FACS coding, and Olivia, a VAD annotator. As a data lead, Olivia is also in charge of managing her company's remote data annotation team. The chapter sheds lights on both differences and similarities of the two coding schemes: while FACS is a highly formalized form of skilled vision, VAD requires little training and is assumed to rely on annotators' untethered intuition with regards to the data they

annotate. However, both FACS and VAD provide coding grids that not only describe but also enact reality (Grasseni 2004; Goodwin 1994): in fact, the adoption of one over the other annotation practice can change the ways practitioners think about the accuracy of their systems' predictions and the validity of their approach. Further, both practices require various discursive practices to negotiate the meaning of data and adapt it to the relevant coding scheme. Both annotation practices constitute forms of what Charles Goodwin has called "professional vision", that is "socially organized ways of seeing and understanding events that are answerable to the distinctive interests of a particular social group." (Goodwin, 1994: 606).

Expert coders: The Facial Action Coding System

FACS coders are usually referred to as "expert" coders, as they are trained and certified through the FACS Final Test, "the only standard for proficiency in FACS coding that is available"⁵⁴, designed by Paul Ekman himself and on sale for \$50 on the Paul Ekman Group's website. Prospective coders can prepare for the test by studying on Paul Ekman's FACS Manual, a \$350 and 527 page-long textbook which, according to the Paul Ekman Group, takes 50 to 100 hours of independent study to complete. Alternatively, students can attend a five-day workshop, the only training course approved and endorsed by Paul Ekman, designed and run by Ekman's protégé Erika Rosenberg and available to students for \$990.⁵⁵ As Ekman advertises on his website, researchers and practitioners who wish to include FACS coding in their projects should verify that coders are certified through his proprietary certification scheme, thus creating demand for his own services. When I spoke to Isabella, who attended

⁵⁴ <https://www.paulekman.com/facial-action-coding-system/>

⁵⁵ <https://www.erikarosenberg.com/facs-training>

Rosenberg's training workshop, she described the intense experience of learning how to FACS-code, and how the certification led to work opportunities:

I learned it through an intensive workshop [...], and it was led by Erika Rosenberg, who's one of the original creators of the system. So, it was very intensive. And I think we spent seven-eight hours a day learning all of the different combinations and then we'd have homework each night and eventually you take a test to see whether you pass or not. And since then, I've just been coding kind of, I don't know, for random people, as opportunities come up, people tend to find the very few certified coders that there are one way or another. And if they're interested in using it in their research, they reach out.⁵⁶

As mentioned above, practitioners consider FACS coding to be an objective, anatomically-based method to map facial muscle activity in behavioral research. Isabella reiterated the scientific validity of the coding scheme, described as an “inherently objective” way to code facial movements. However, she expressed concerns with the use of FACS coding for emotion inference:

What I would say about FACS is, FACS is inherently objective. It's muscle based. I feel like the science is valid. We know that people move their faces in particular ways. Where I often deviate from the system or give people caution is to say, using those combinations that have been set by people like Paul Ekman and Dr. Keltner,⁵⁷ to use it as evidence of a discrete emotion. People's faces are always moving. And it doesn't seem that a particular

⁵⁶ Interview conducted on March 8th, 2022. All the following quotes from Isabella are from the same interview.

⁵⁷ Professor of psychology (and Paul Ekman's collaborator) Dacher Keltner provided scientific advice for the 2015 Pixar's movie *Inside Out*, which portrays five personified basic emotions (Joy, Sadness, Fear, Anger, and Disgust) shaping the actions and thoughts of a young girl (Keltner and Ekman 2015). Anna Lauren Hoffman and Luke Stark have argued how the popularization of Ekman's Basic Emotion theory through media and digital technologies not only reinforces stereotypical and gendered assumptions about affective behaviors; it also has a prescriptive effect in that it can re-shape the way we understand our own emotional and mental experience (Hoffmann and Stark 2015).

combination like, say, an action unit 12 and 6 is automatically indicative of somebody feeling happiness. I think the system is really great in classifying anatomical movements, but it's that leap then to the affective world where I think the system isn't 100% predictive.

Certainty over the coding scheme's objectivity, however, vacillated when Isabella described the process of negotiation between coders' interpretation of the data required to ensure consistency across annotations. Typically, to achieve reliability, two or more coders annotate the same data, but – as she described – they do not necessarily 'see' the same action units:

It's quite surprising. So even though it's an objective system, anatomically based, what one person sees, another person might not see, and a lot has to do with just having a baseline condition. So especially if you're quoting stills and you don't know what a person's natural resting face looks like, it's really easy to mistake a line, a wrinkle for a movement of a contraction of a muscle, and so people will oftentimes disagree. I think the average reliability, a good reliability is like 65%, which obviously is not particularly high. I always shoot for at least 80%, but that often involves multiple samples of coding with another coder establishing that reliability over time, working out your disagreements, trying to figure out why so and so might say that's a 14 versus 12, which are easily... There are certain combinations that are easily confused and that can take a lot of time just to get people's reliability even to 70%, let alone 80%. Yeah, it's not a perfect system by any means.

In addition to the lack of a baseline reference for emotional expressions (what a person's "average" facial expression looks like), other factors can shape FACS coding, such as data quality (image resolution, head pose, and light conditions) and a lack of contextual information. As she described, all these factors can significantly influence coding reliability:

Sometimes people just throw me one picture of an individual and ask me to code it. And there's no context, there's no baseline. And so that can be tricky. And as I discovered... Actually we're working on a project with another collaborator. Even though we had very posed caricature faces, our reliability across the affective domain was still only about 60%. So, without any context and even with the best of photos, it can be low. And the other thing you run into, too, is that sometimes people give you a very grainy photo or half the face is cut off or they're turned, you can't necessarily see what's happening. So that can be tricky, too. Sometimes it's just the quality of the image. Really it is difficult to see some of these movements.

In addition, adherence to the coding scheme does not guarantee agreement between coders. As she explained to me, when coders are asked to annotate only the combinations of action units that map on the six basic emotions (such as happiness or sadness), they can observe different combinations of facial movements for the same overall emotional expression. This leads to disagreements over the definition of inter-reliability itself:

So, in the FACS manual, you can code fully for FACS with all 40 some action units, and then there's a bunch of head movements as well. And some people will not want the full FACS system. They will only want what's called EMFACS, which is "emotion FACS". And so, those are just the combinations that researchers like Ekman have reported map on to different emotional states. That said, though, even if you look at what are they considered the EMFACS code or something, as you might think, is as universal as happiness, there's probably six or seven different possibilities, and so any of them are considered evidence of happiness, which then leads you to the question, are there really a set of individual action units? Because the action units involved can range for anywhere from two to four different combinations, and sometimes they don't even include the same action units. So, there's a lot of variability about what people want and then how people

use the system as well. And then, of course, how you indicate inter-rater reliability. Some people will just say, oh, you said [action unit] 1, I said 1. That's an agreement. But if the full combination is supposed to be [action unit] 1 plus 2, but I've got the 2 and you got the 1, maybe that's considered 50% agreement, or maybe it's considered no agreement because you were looking for the '1 plus 2' combination. So, that's also how you can get differences in inter-rater reliability. Are you coding each individual unit 'present/absent', or are you trying to code for a bigger combination of purported codes that map onto an emotion?

Borrowing from Cristina Grasseni's ethnographic study of professional visions among Northern Italian cattle breeders,⁵⁸ the disagreements and negotiations reported here by Isabella show that even in a standardized practice such as FACS, coders must undergo a "process of 'enskilment'" that involves "disciplining, selecting, re-interpreting, and distancing oneself from one's naïve and undiscerning vision" (Grasseni 2004: 43). Further, the interpretation work foregrounded by these responses seems to challenge the purported objectivity and accuracy of FACS coding. Practitioners, however, are mostly concerned with the material and financial challenges of FACS coding, including the costs of hiring or training FACS coders and the length of the annotation task, which is often in conflict with the faster pace of product development or paper submission.⁵⁹ In this context, disagreements among coders' ways of 'seeing' the data is interpreted as a problem of *capacity* rather than *perspective*. As mentioned in chapter 3, facial micro-expressions – as theorized by Ekman – take place almost too quickly for the human eye to accurately capture them. As Isabella described, machines can help detect those micro-movements: "That's a really valid attempt. My hesitation would be then saying that [a] combination of action units is indicative of

⁵⁸ By studying how breeding experts learn to recognize "animal beauty", Cristina Grasseni argued that vision is trained through apprentice practices that are socially and culturally shaped (Grasseni 2004).

⁵⁹ As mentioned in the previous chapter, the field of computer science currently prioritizes a conference publication model, with a faster submission, revision, and publication process.

fear, but I think it's really a great effort for these systems to be able to pick up these different movements that the human eye doesn't always see". For these reasons, practitioners are increasingly turning to ways to scale and automate the annotation of facial expressions, either by adopting dimensional models of emotional expressions, which rely on larger number of less expensive, "lay" annotators, or by implementing machine learning techniques for the automatic annotation of facial expressions – or combinations of these two approaches. The next two sections focus on these practices.

Lay annotation: Dimensional models of affect

Dimensional models of emotions offer an alternative to Ekman's categorical view of emotions, suggesting the existence of continuous affective dimensions representing the perceived pleasantness/unpleasantness, degree of intensity, and dominance/submissiveness of emotions. This model is generally attributed to James Russell and Albert Mehrabian, who in the late 1970s proposed a 'three-factor theory' of emotions that, as opposed to Ekman's rigid and categorical taxonomy, mapped affective experience along three axes: arousal, pleasure (or valence), and dominance (James A Russell and Mehrabian 1977; James A. Russell 1980). In affective computing, the adoption of the Valence, Arousal, Dominance (VAD) model allows practitioners to "sidestep issues of theoretical validity to instead focus directly on the relevant non-basic affective states" (Rouast, Adam, and Chiong 2021: 527). However, affective computing practitioners often combine dimensional and categorical models by plotting VAD dimensions on to Ekman's basic emotion (Figure 5). As participants explained to me, this method guarantees the 'objectivity' and scientific rigor provided by FACS but offers several significant advantages with regards to coding: first, practitioners do not need to hire or train expensive "expert" coders, as the majority of the annotation can be done by "lay" data workers with minimal and informal training; second, accuracy

can be determined by averaging annotation scores of larger numbers of annotators, rather than having to establish and assess inter-coder reliability; third, VAD annotation is supposedly faster, compared to the lengthy process of FACS coding, which can often take up to 100 minutes to code 1 minute of video data.⁶⁰ Finally, since VAD explicitly relies on human judgements, many participants reported that VAD dimensions allowed them to make their FER systems “more human”. As software engineer and CTO Peter described, while FACS coding is a rigorous and valid method for the detection of facial movements, it does not reflect the ways human themselves perform facial expression recognition:

So, this [FACS coding] is very strict in theory. I think, in theory, it may look like a good framework to use, but in practice, I think when people are expressing facially, I don't think it's a really rigid and fixed set of the same action units that happen. So, people are more fluent in the way that they're expressing with their face. And, therefore, we prefer an approach that we humans ourselves would follow. So, if I, as a human, were to watch somebody and try to identify a facial expression of smile or of raising eyebrows in in awe or in surprise, what would I be watching for? So, without any knowledge of action units and specific muscle movements., me, as a human... And I might be not perfect, and I might make some mistakes, but you know, through living in human society, I've learned how to read facial expressions. You know, some people do that better or some people do that worse. This is a very intuitive way of identifying facial expressions for us humans. It doesn't require any major expensive training like FACS coding, and the benefit of it is that you can, rather than relying on the opinion of one or two FACS coders, you can actually have, you know, hundreds of people conducting the annotations. And you can then again use statistical methods in trying to detect outliers, detect when there's a lot of agreement between many people and then use that data to kind of train an AI algorithm

⁶⁰ <https://www.erikarosenberg.com/more-about-facs>.

to behave the same as an average human being as measured by the accuracy of an average human.⁶¹

Historians of science Lorraine Daston and Peter Galison have noted how, in the twentieth century, the unmediated objectivity provided by scientific instruments (e.g. the “mechanical objectivity” of scientific photographs) began to be accompanied by what they termed “trained judgements”, ways to interpret scientific data that could be learned and taught, irrespective of scientific expertise or skills. In this sense, the interpretative element was not at odds with the search for scientific truths, as subjectivity had become

an important feature of classification because the objects did not demonstrate universal essential properties *and* because in the mid-twentieth century a growing number of scientists across many fields began to take it as a good thing that people could be trained to classify objects univalently even in the absence of strict protocols (Daston and Galison 2007: 335).

Similarly, the interpretative and subjective element of VAD annotation is not in conflict with the purported objectivity of FACS coding. As I will show in this chapter, the outsourcing of classification tasks to unskilled or minimally trained workers allows practitioners to scale up their FER projects and, at the same time, guarantees a form of accuracy that relies on annotators’ supposed ability to ‘see’ affective data in the same way.

⁶¹ Interview conducted on July 8th, 2021.

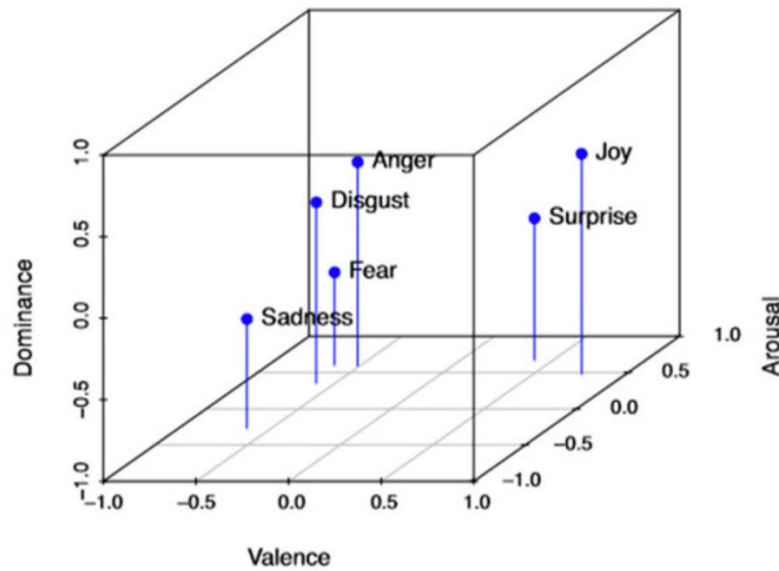


Figure 5. The VAD model mapped on the six basic emotions (reprinted from Mitruț et al. 2019).

When I spoke to Olivia, a data lead and VAD annotator at a FER company, she described VAD as a subjective inference process, in contrast to the perceived objective and binary nature of FACS coding, which instead describes the presence or absence of action units:

Valence, dominance and arousal are quite subjective because, for instance, if I was annotating you now on the screen, I might say that you look quite like interested and relatively happy. These are all inferred. You know, we can't actually say what someone's feeling. It's all inferred from what we believe it to be. But someone else might view that quite differently. So, there is quite a difference between the annotations between the different annotators. *There can be quite a difference within the valence dominance arousal, whereas the AU is quite.. You know, it's either there or it's not. [VAD] it's kind of like a.. It's a scale, you know, it's a sliding scale rather than it being like, it's present or it's not present* [emphasis added].⁶²

⁶² Interview conducted on July 15th, 2021. All the following quotes from Olivia are from the same interview.

Ahead of our meeting, Olivia had prepared a set of definitions of valence, arousal, and dominance, expecting I would ask her about it. As she went through the different meanings of the three dimensions, originally provided by her company when she first started as an annotator, she often dwelled on the difficulty of providing straightforward explanations:

So, valence is ‘the individual’s overall sense of will’, or, ‘does he/she on balance feel positive or negative about the things, people or situations at the focus of his/her emotional state?’ So, overall, I tend to think of this as how happy or sad they are, *but it’s not as obvious as the, you know... it’s an overall state and as always, it’s inferred* [emphasis added]. We can’t actually tell if someone’s happy or you know. But yeah, as I’m annotating that’s sort of my main, thought, it’s how positive or negative someone is. The arousal is the individual’s overall inclination to be active or inactive. Like valence, it is a cover term. Arousal may include mental activity, preparedness to act, overt activity and alertness [...]. And this, what I’m thinking about when I’m annotating is how engaged or bored they are. [...] *But it’s not as straightforward as if they’re bored or... And it is quite difficult to explain it. I think that’s kind of the issue with the VAD annotating that it is so subjective, but that’s also just part of it, I guess* [emphasis added]. And the dominance, dominance rate and scale deliberately mention two related concepts, power and control. Emotion is about people’s sense of their own power, and that seems to be relative to what they are facing. Dominance is ‘the composite sense of being well placed to directly control events rather than being at the mercy of the events.’ So, yeah, this is like how in control of the situation, the person is. A lot of the time in the video and because of we are annotating just a video of someone speaking, I would consider this more like how confident they are about what they’re saying. [...] But yeah, again, it completely depends on what you are annotating. But considering what I’m annotating at the moment and a lot of these, a lot of them, like I said, they do stay quite neutral throughout all of it.

As she recounted, the meaning of valence, arousal, and dominance can be subject to interpretation and negotiation among the annotators themselves. As the lead of her annotation team, Olivia helped her annotators make sense of the data with respect to these definitions. One noteworthy example is the definition of dominance in the context of expressions of anger:

So, for instance, one of my annotators was unsure whether anger would be if someone appeared angry in the video. Would that be the high dominance or low dominance because she was questioning, well, if someone's angry, then is it like a loss of control, whereas I would have annotated it as anger is being more in control than if someone were sad, so then a higher dominance than it would be if it was someone showing sadness. So, I think we had like a bit of a discussion about that, but I think it's just that we've not really, not really discussed the definitions as such. But how to annotate different emotions has come up in the past.

Olivia, who is hired in-house by her company, leads a relatively small team of freelance annotators. Among her responsibilities is making sure that annotations are sufficiently consistent with each other. To keep track of annotators' work and give tips on how to annotate the data, she has set multiple WhatsApp group chats to communicate with them. Due to the subjective nature of the annotation, however, she must often negotiate among annotators' judgements, mirroring the mediation process described in the previous section by the FACS coder Isabella. As Olivia described:

I look at the annotations. If I think something's wildly different, doesn't make sense, then I will go back to the annotator and be like, 'you know, what has made you annotate this in this way? Can you explain it to me?' *We'll have a bit of a chat about it because it might just be that we've viewed something different. We might have a bit of a chat about it and*

then maybe tweak things. A lot of the time it's not that it's wrong, you know, it's just someone else's view on things [emphasis added]. And in that case, because we've got a few different people doing the annotations of each video, I believe. I'm not a 100% sure because it is sort of down the line for me, but I think that what's taken is a bit of an average of those annotations.

In discussing her role within her team and the rest of the company, she often cautioned me that she was “not an expert”. Moreover, despite the small size of the company, her tasks were limited to hers and her team’s annotations, with little or no insight on other segments of the pipeline. Perceptions and attributions of expertise seemed to act as a *barrier* to a more even distribution of knowledge within the company:

As sort of a proper employee, I think that I assumed that I would sort of learn more about that as I went along. But I've not been sort of, not trained, that's not really the right word, but I've kind of been like told, like, this is my part. And then, you know, I just assumed I'll pick stuff as I go along in the role of how it works further down the line. But yeah, I think it would be nice to have a little bit more insight into a little bit further how it goes just beyond me, that would be quite nice. I think, also perhaps I'm like, because I'm not a computer scientist, I'm probably a little bit less likely to ask because in case I don't understand.

Contrary to the perception of this type of annotation work as low-level and unskilled, however, Olivia’s experience reveals the complexity and length of the annotation task which requires multiple breaks and quality checks:

The videos are quite difficult. I tend to do like one video and then I'll do something else, even if it's like answering an email or, you know, do something and not look at it for a little bit, but depends on the video. Some things take me a lot longer than others. The

longer it is, generally, the longer it takes to annotate, but it also depends on what is happening [in the data]. Like I said, they're quite neutral, so there's not much happening throughout it. But if it's someone who's like, really animated, it takes a lot longer.

As suggested throughout Olivia's responses, this type of annotation is further from menial and straightforward, requiring workers to continuously negotiate and fine-tune their interpretation of the data. As this last quote describes, annotators must reconcile the various ways of 'seeing' the data in order to produce consistent labels:

Because obviously it is subjective, so you can't just say, well, I did it different to that, so you must be wrong because how do you know? Yeah, it's probably not, it's just how someone's interpreted differently, but that sometimes just sparks a conversation between us. Like, Well, why? Why did you think, let's have a chat about it? And then we tend to come to an agreement on what we think. *And I think it's definitely kind of difficult. It's difficult when there's lots of different people annotating it because it can be quite different and it's quite interesting looking at the difference between them* [emphasis added]. You know, and you've got two people annotating the same video. And some people annotate slightly differently, just for instance, they might... So, with the with the VAD is sort of like a line. And like you either... if you annotate it positively or negatively. So it's kind of like, modular, like you are kind of, like, more aroused less aroused. And then some people annotate a little bit more smoothly, I think that's just literally just using the tools like how we annotate, but that doesn't really make a difference. I don't think overall because it's sort of an overall interpretation of it. And yeah, it's just a difference between people.

Despite the perceived subjectivity of the VAD coding scheme, annotators must conform to a common frame of interpretation in order to reach a consensus over the meaning of data. However, because of the interpretative and subjective element of the VAD annotation, practitioners must adopt strategies to limit or reduce disagreements

among annotators. The next section investigates the role that automated labelling plays in guiding the labelling of affective training datasets.

Automating visions: The NOVA interface

A major impediment to deep learning advances in affective computing is the relatively small size of available labelled datasets. As shown in the previous sections, the annotation of facial expressions can be lengthy and burdensome, even when employing VAD annotation, making large-scale labelled facial expression datasets hard to build and maintain. Automated labelling promises developments in the area (Rouast, Adam, and Chiong 2021); however, as researchers have noted, the “inherent ambiguity of affective displays” challenges efforts to replace annotators (538). A partial response to this problem is the adoption of machine learning-based labelling techniques that aim to guide annotators’ work and speed the annotation process. At the company where Olivia works, annotators use NOVA, an open-source annotation tool that provides semi-automated labelling for discrete and continuous scores of affective behavior, along with visual feedback that help annotators inspect and correct machine-generated labels.⁶³ NOVA – presented in 2018 in the form of a conference paper at the International Conference on Affective Computing and Intelligent Interaction (ACII) – employs explainable AI techniques so that, according to its creators, “even non-machine learning experts get an idea about the strengths and weaknesses of the underlying classification model and can immediately decide which parts of a prediction are worth keeping” (Heimerl et al. 2019: 2). Ironically, this tool was created with the recognition that, in affective computing, annotators “typically disagree on the labels” (3). However, rather than a threat to the overall methodological approach, disagreement is seen as technical problem that can be solved with automation. The semi-

⁶³ <https://github.com/hcmlab/nova>

automated labelling and the visualization techniques aim to “keep human effort as low as possible” (1) and establish trust in the annotation process.

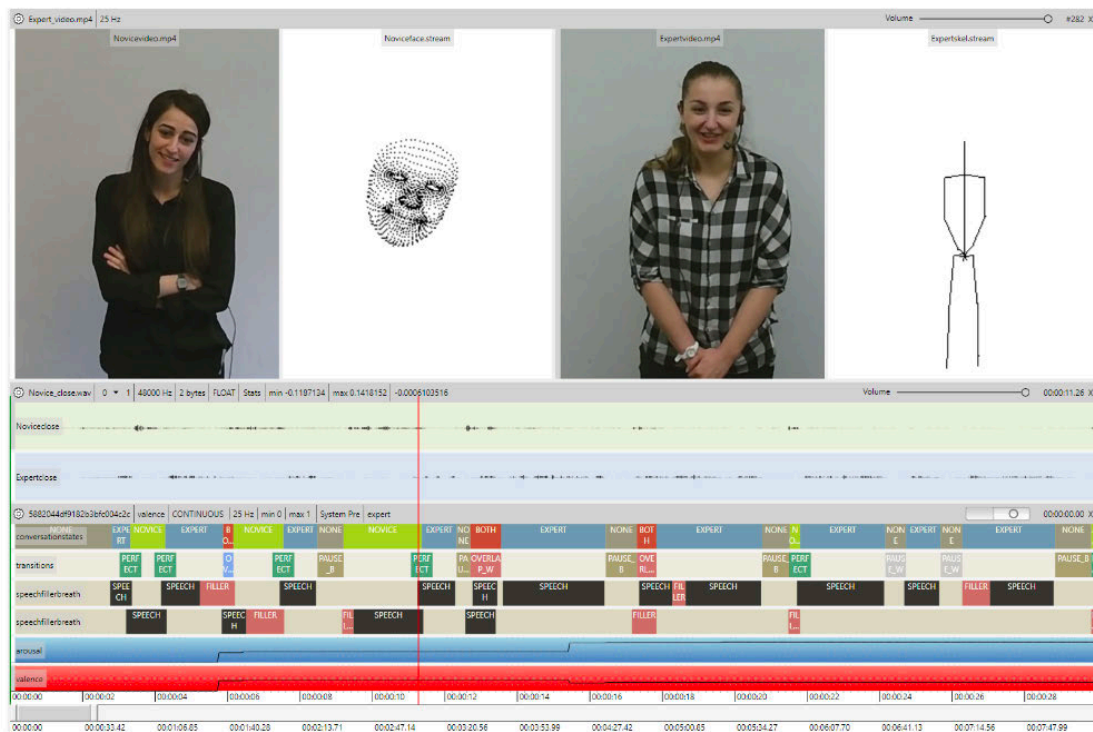


Figure 6. Discrete and continuous annotation schemes supported by the NOVA interface (reprinted from Heimerl et al. 2019).

As shown in Figure 6, NOVA supports the annotation of multiple scores, including continuous dimensions such as valence and arousal. The collaborative element consists of a back-end database that allows annotators to load and save annotations from and to a cloud database that can be accessed by both human annotators and what the authors call “machine users”. Just like human annotators, these algorithmic agents “can create and access annotations” (4). This way, the model is trained on data labelled by human annotators and then used to predict unseen data. During this process, an “active learning” module flags parts of the prediction that require manual revision. The model is then retrained, and the procedure repeated until all data is annotated.

To help annotators inspect and re-label data, NOVA employs explainable AI visualization techniques. Figure 7 shows a visualization representation generated using

the Local Interpretable Model-Agnostic Explanations (LIME) technique, which highlights the features of the image that contributed to each prediction score. For instance, the top-left square shows that the features located around the mouth and chin region were the most salient for the generation of the prediction score associated with the ‘happy’ emotion category. Explanations techniques like LIME are commonly used in areas where machine learning algorithms are part of decision-making processes, such as healthcare. In these contexts, saliency mapping is supposed to help radiologists and clinicians interpret the algorithmic results and improve human-machine collaboration. Scholars, however, have pointed to the potential risks of explainable AI approaches. In addition to the well-documented risk of automation bias – the overreliance on algorithmic results due to the system’s purported objectivity (Logg, Minson, and Moore 2019; Sundar and Kim 2019; Araujo et al. 2020) – scholarship in this area has noted how explanations models are rarely tested, therefore it is difficult to assess their performance. In the case of radiology, for instance,

the clinician cannot know if the model appropriately established that the presence of an airspace opacity was important in the decision, if the shapes of the heart border or left pulmonary artery were the deciding factor, or if the model had relied on an inhuman feature, such as a particular pixel value or texture that might have more to do with the image acquisition process than the underlying disease (Ghassemi, Oakden-Rayner, and Beam 2021: 746).

These visualization techniques require that humans interpret each given explanation, potentially reinforcing selection and confirmation bias, that is, the tendency to select and attribute positive meaning to the explanations that confirm the human operator’s beliefs (Ibid; Lipton 2017). Further, by sidestepping annotators’ disagreement, tools like NOVA can reinforce and normalize assumptions about methodological approaches to annotation, such as the idea that expressive behaviors can be accurately measured through a finite set of numerical scores (either discrete or continuous).

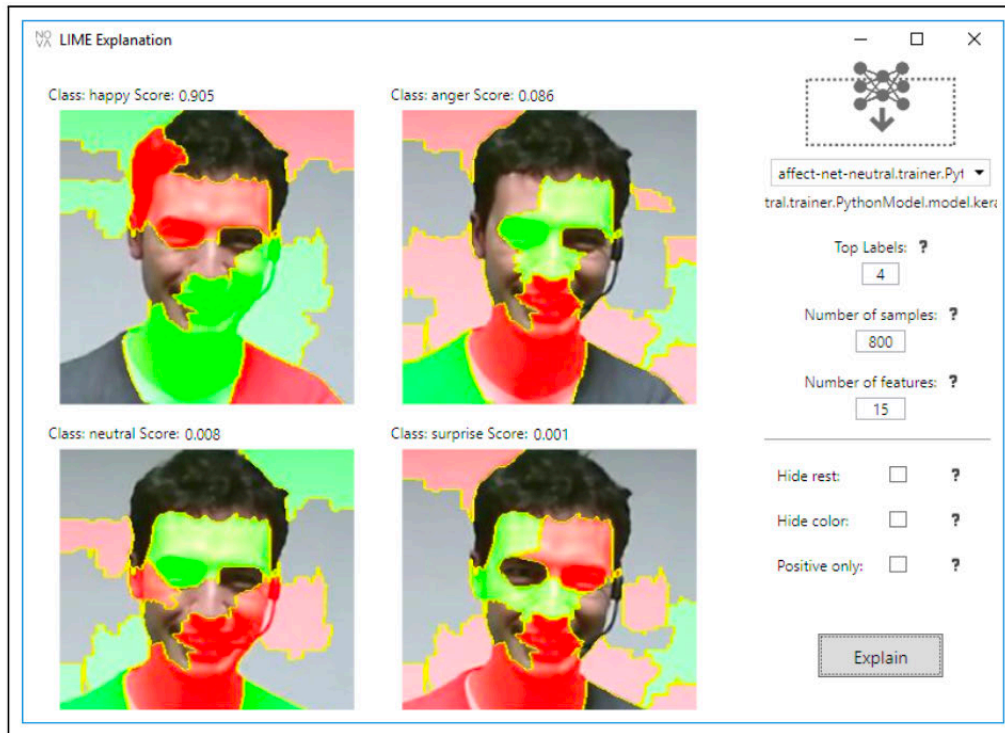


Figure 7. LIME explanations for four emotion scores (reprinted from Heimerl et al. 2019).

Building on Foucault’s investigation of the relationship between power and knowledge, Milagros Miceli and Julian Posada have defined the material tools and interfaces that shape and constrain data work as part of a *data production dispositif*, “an ensemble of discourses, actions, and objects strategically disposed to (re)produce power/knowledge relations in data and labor” (Miceli and Posada 2022: 1). Borrowing from Daston’s and Galison’s study of scientific objectivity, tools and interfaces like NOVA become then the infrastructure through which subjective judgments can be formulated (Daston and Galison 2007). In this regard, it is noteworthy to mention that, while sophisticated labelling tools such as NOVA aim to tame the inherent subjectivity of VAD annotation, FACS coding is usually performed using much more “low-tech” tools, perhaps due to

the perceived objectivity of their coding practice. As Isabella explained to me, FACS coders

Manually take notes and usually just submit [the] codes on a worksheet or Excel worksheet of some sort. The researcher will look for reliability and then say, your reliability is X or Y with this person. Here's a particular action unit that you aren't agreeing on or particular set of action units. And then they have us either revisit those individually or set up a meeting and discuss the differences.⁶⁴

In addition to the automated techniques described here, the next section illustrate how practitioners' perceptions of the complexity or simplicity of annotation tasks can further obscure the work of interpretation and negotiation that underpins both coding practices and reinforce power structures that posit annotators as interchangeable or replaceable.

“It sounds like magic”: Representations of data work matter

In chapter 5, I showed how participants' perception and representation of their own practice contributed to the maintenance of their professional authority. Borrowing from Wulff and Finnestrand, I described this type of boundary work as “the right to determine and perform the work as they see fit” (Wulff and Finnestrand 2022: 711). This section explores participants' representations of data annotation and data workers (FACS coders and VAD annotators), and how these representations might affect experiences of annotations, labor practices, data infrastructures, and annotators' agency within machine learning pipelines. As Lucy Suchman pointed out, representations of work matter, as “work has a tendency to disappear at a distance, such

⁶⁴ Interview conducted on March 8th, 2022.

that the further removed we are from the work of others, the more simplified, often stereotyped, our view of their work becomes” (Suchman 1995: 19). In this sense, representations of other people’s work and their level of expertise can shape and reinforce power dynamics within data structures. Making data work visible, then, is not simply the foregrounding of labour practices but, rather, the annotators’ ability to shape how their work appears to and is represented by others.⁶⁵

All participants in this study described FACS coding as an objective, scientific, and consistent approach to annotation, as opposed to the subjective, lay nature of VAD. As Nathan, a CTO and electrical engineer described, FACS coding is governed by a strict set of rules, with a “right” and “wrong” way to perform the annotation:

So, if you're doing FACS annotation, which is facial muscle action unit annotation, there is a right and wrong way. And there are lots of complicated rules around if this is happening. So if this particular muscle is active and that particular muscle is active and you see this, then it means that in terms of... There are so many... I can't remember, 43 muscles in the face that they monitor and the activation level is either neutral, inactive, or active. There's no subjectiveness.. In some ways, there might be some subjectiveness.. There's some subjectiveness, particularly around the boundary condition, so the change from one [action unit] to another. But there's very clear definitions around what's the difference between a B activation and a C activation, but whether it's a high B, low C... [...] So that's very like... I'm trying to think if there's a good analogy, but there's a set of rules as a handbook [...] and you can flip through it. I can look at your face and kind of say ‘yeah this isn’t a 6 and a 12’, and it's very scientific in the way it's done, whereas arousal, valence, dominance is more, a lot more subjective, but then there's a lot of

⁶⁵ I originally developed this argument in a short position paper co-authored with SJ Bennett and presented at the workshop on ‘Investigating Data Work Across Domains: New Perspectives on the Work of Creating Data’, held in April 2022 at the ACM CHI Conference on Human Factors in Computing Systems. The paper is available as a pre-print at <https://www.ai-ethics.org/hidden-humans-in-the-loop>.

agreement between people without much training. It seems to be something the people can do.⁶⁶

As mentioned in the previous sections, various factors, including financial and temporal resources, influenced the choice of one over the other annotation scheme. While certified FACS coders are expensive and highly sought after, VAD annotators do not require any formal training and can be sourced in larger numbers. As seen in the previous sections, for some participants these reasons were explicit and at the heart of their approach to annotation. For others, the perceived properties of the two annotation schemes guided the overall system design. Marvin, a professor in computer science with experience in consulting, described the trade-off between the reliability of FACS coding and the lay ability to discern facial expressions:

If your goal is to know how people perceive expressions specifically, then you may want untrained annotators because your goal is in itself to know like judgment of people with limited instructions. As you know really well, expressions had over the years nice coding schemes. So FACS coding is one of them. Very popular. So, in this case what you want is reliability. I mean, you want annotation, and you want reliability in your annotators. So, I think there's at least two families of annotations, one where reliability is key and another one which is more about like judgment, and you embrace the subjective aspect of the problem. I think there's at least these two families and their methodology will be different. I mean in the case of subjective... If it's a well-established coding scheme, you should not go too far in the annotation until you get good reliability between your annotators and you train and if needed, you reiterate on that. In the case of where subjectivity is known and embraced, in this case the training of the annotators is less... But at least you should well define your research questions.⁶⁷

⁶⁶ Interview conducted on April 21st, 2021.

⁶⁷ Interview conducted on February 2nd, 2022.

Because of the perceived reliability and consistency of FACS coding, Marvin shared his expectation that, one day, this type of annotation will be fully automated: “I think because FACS coding is something that there is a reliable way of annotating it that is, humans can annotate it very consistently... So, then you will expect with enough work that an AI could do the same job.”⁶⁸ In this case, the perceived objectivity of the annotation task shapes the participant’s belief that coders could be effectively replaced by algorithms, despite the struggles reflected above in Isabella’s experience of FACS coding.

In describing the freelance annotators hired by his company to perform VAD annotation, Nathan mentioned that these “are usually stay-at-home moms.”⁶⁹ This response reflects the overall gender imbalance within machine learning pipelines, where the majority of high-level, engineering roles are occupied by men. This gender inequality is also mirrored in studies that have investigated the demographics of platform workers (such as Amazon Mechanical Turkers), and found that the majority of them are indeed women (Ross et al. 2009). Historians of science Lorraine Daston and Peter Galison have argued that, in the 19th century, the practice of employing women to perform menial and “unskilled” work in science (such as performing calculations or labelling and classifying astronomical images) would make scientific practice more objective, as it

served as a tacit guarantee that data thus gathered were not the figment of a scientist’s imagination or preexisting philosophical commitment [...]. In this respect, the workers were identified with the machines, and, like the machines, in their ‘emptiness’ they offered a transparency through which nature could speak. Second, beyond their supposed ‘lack of skill’, women workers were presumed to offer a ‘natural’ predilection away from the grand speculative tradition (Daston and Galison 2007: 341).

⁶⁸ Ibid.

⁶⁹ Interview conducted on April 21st, 2021.

In a similar way, Nathan described the ability of VAD annotators to recognize facial expressions consistently as a kind of intuition, untarnished by any form of specialized knowledge. As he explained, VAD annotation is more akin to a gut feeling than to adherence to a set of rules (as it is the case for FACS coding). To him, consistency across annotation resembles the kind of serendipitous events that can only be described as *magic*. However, when high consistency among annotators is not met, they will look for discrepancies and exclude inconsistent annotations from the dataset:

There is a [VAD] manual but it's literally like four paragraphs instead of kind of 400 pages [laughs] and it's kind of a bit more *go with your gut* [emphasis added]. But what we tend to find is that, sometimes what you find is that magnitude doesn't necessarily agree, there's a high correlation between annotators. So, yeah, just seems to be something that humans can read about another human. If it sounds... *It sounds a bit like magic* [emphasis added]. And there is an element of it [that] is subjective. And you can't say annotator A is right, annotator B is wrong because that's their opinion of the thing. So, this is why we do it for a number of people, because we expect a high correlation and because we see generally a high correlation, there's somewhere where suddenly there is not a correlation. Then we will manually look at that section, because you know, did they slip or were they just not paying attention? And generally, normally, we exclude it if that annotator has a huge variance from the from the group.⁷⁰

Scholarship that has investigated the use of magical discourse in relation to technology has pointed to the role magic plays in reinforcing techno-determinism and obfuscate accountability. Kate Crawford and Alexander Campolo have defined this discursive strategy as *enchanted determinism*, “a discourse that presents deep learning techniques as magical, outside the scope of present scientific knowledge, yet also deterministic, in

⁷⁰ Ibid.

that deep learning systems can nonetheless detect patterns that give unprecedented access to people's identities, emotions and social character" (Campolo and Crawford 2020: 3). Similarly, Lilly Irani has pointed to the magical discourse invoked to hide the human labor that goes into the process of data cleaning, standardization, and annotation that powers many machine learning systems. As she noted, "that magic always relies on invisible labors" (Irani 2015a). Here, Nathan's invocation of magic seems to obscure the work of interpretation and negotiation that, as shown in the previous sections, forms the "skilled vision" of annotators. Borrowing from anthropology studies on the relationship between magic and technology, Madeleine Clare Elish and danah boyd suggested that "to evoke magic is not only to provide an alternative regime of causal relations, but also to minimize the attention to the methods and resources required to carry out a particular effect" (Elish and boyd 2018: 63), including the struggles, disagreements, and negotiations process that make up the annotation process.

In surveying participants' representations of data work, this section has showed how practitioners' perceptions do not necessarily reflect the reality of affective annotation, nor how annotators conceptualize their practice. This is consistent with other studies of data work that have highlighted how practitioners often consider annotators as low-level, unskilled data workers, as opposed to high-level and qualified stakeholders such as engineers and domain experts. Annotators, on the other hand, see themselves as creative workers, contributing to the overall outputs of the machine learning system but are often excluded from the decision-making process (Gray and Suri 2019; Irani 2015b). Representations of annotation work – namely, the intuitive nature of VAD annotation, or the objectivity of FACS coding – might contribute to the creation of work structures that forestall or exclude annotators' agency within the overall system design. To reiterate this section's main argument, making data work visible must involve annotators' ability to shape how their work appears to others within

the design pipeline. As Miceli and Posada have noted, “breaking with the alienation of data workers means much more than rendering them visible. It rather requires making the rest of the machine learning supply chain visible to them” (Miceli and Posada 2022: 30). The final section of this chapter reflects on the implications that visibility has for annotators’ agency within data structures.

Seeing data work, accepting uncertainty

How does one make data work visible? And what does ‘visibility’ mean in this context? Reflecting on the relationship between researchers and practitioners located in the Global North and the data workers and annotators recruited from the Global South, Noopur Raval has noted how focusing on the invisible or hidden nature of data work does not disrupt the power differentials (and the political economy that sustains them) between Global North and Global South. As she notes, “simply or vehemently emphasizing the role of knowing and seeing (*if only they could see the hidden workers!*) may not be adequate to automatically offer pathways for alternate or emancipatory design in a world that we share unequally” (Raval 2021: 30). Similarly, Miceli and Posada have proposed to shift the focus from workers’ in/visibility to the material factors that preclude annotators from participating in the decision-making process that underpin the design of data-driven systems. As mentioned in the previous sections, these factors include labor dynamics, clients’ and managers’ expectations, and tools and interfaces that can reinforce and normalize annotation schemes and labels. For the authors, interrupting workers’ alienation,

means providing information and education on technical and language matter that could help workers understand how their valuable labor fuels a multi-billion dollar industry. This also concerns questions of labor organization and unionizing: For instance, the recently-created Alphabet Workers Union has taken steps in this direction by including contractors — many of them outsourced data workers. To help counter their alienation,

researchers and industry practitioners need to regard data workers as tech workers as much as we do when we think of engineers (Miceli and Posada 2022: 30).

According to this framing, to address the power inequalities within machine learning design, researchers and practitioners must recognize annotators' contribution to the overall model development, both conceptually and economically. In the context of affective computing, this might mean recognizing that disagreements between annotators are not statistical glitches but rather an inevitable part of the affective annotation process. Even though in the projects described here, annotators were not outsourced, their contribution was still treated as a "menial task", and annotators' disagreement were seen as statistical errors and outliers. In this sense, accepting that there might not be straightforward answers to the problem of ambiguous labels could help avoid inaccurate – or, in the case of affective labelling, perhaps impossible – classification. As Marvin described, this uncertainty should be an integral part of algorithmic results:

I believe that classifiers should always be allowed to say "I don't know". A human is allowed to do it, but classifiers are not allowed to do it, or many of them are not allowed. And the reason for that is that I want AI to fail. AI will fail.. as long as it knows that it failed. So, if it doesn't know that it failed, that's the worst case. And you see often these days with deep learning pressures, overconfident algorithms. So, I think it's not just inaccuracy but it's like, if you're wrong in your prediction and you were so certain that it was the right one, you should be penalized a lot more than if you're wrong but you knew you were probably wrong, and you should not be penalized.⁷¹

For Louise Amore, the recognition of this uncertainty is necessary to counteract machine learning's "vision-dominated objectivity that claims to have a truth beyond

⁷¹ Interview conducted on February 2nd, 2022.

doubt” (Amoore 2020:137). By bringing to the fore the subjective and partial knowledge that constitutes machine learning ground truths, we can then challenge the claims of objectivity and universality brought about by algorithmic results. In this sense, she argues, we must recognize the intrinsic “doubtfulness” of algorithms: “To be doubtful could mean to be full of doubt, in the sense of a fullness and a plenitude of other possible incalculable paths” (142). To be doubtful, however, requires the ability to speak of the limitations of algorithmic systems; a kind of discursive practice that, borrowing again from Foucault, Amoore defines as “risky speech” or “parrhesia” (Ibid.). In ancient Greece, parrhesia referred to a form of political and ethical discourse in which the speaker would speak frankly in front of those in power (Michel Foucault 2011). To speak bravely of the uncertainties that permeate algorithmic results, one must not only point to the contingencies on which ground truths are created but incorporate this doubtfulness into the design process itself. Here, doubtfulness becomes a political project in that, by foregrounding the uncertain process of meaning-making that underpins data annotation, it can re-arrange those configurations of knowledge, expertise, and power explored at the beginning of this chapter.

While uncertainty in machine learning is often treated as a technical problem that can be solved through statistical means (Abdar et al. 2021), this chapter has shown how uncertainty is instead constitutive of the project of meaning-making that makes up the annotation process and the creation of ground truth datasets. The experiences of annotation illustrated so far have hopefully shed light on the existing tensions between instructions, expectations, and the reality of data work. However, as critical scholarship has shown, data workers’ feedback and concerns are rarely taken into consideration, and annotators are seen as providing a discrete service rather than meaningful collaboration. Several scholars have attempted to translate these concerns into organizational practices: Miceli et al. have proposed the documentation of data production practices, including annotators’ needs and concerns (Miceli et al. 2022),

while HCI researchers Lilly Irani and Michael Six Silberman have designed an “activist systems” that allows crowdsourced workers to rate and contest their employers’ requests (Irani and Silberman 2013). While power differentials between data workers and practitioners remain a major challenge, these are some of the ways uncertainty can be brought to the fore within data structures to challenge the epistemological assumptions on which machine learning models rest and help re-evaluate annotators’ specialized forms of knowledge and expertise.

Conclusion

In this chapter, I have focused on two common coding schemes for the annotation of affective data: the Facial Action Coding System (FACS) and the Valence, Arousal, Dominance (VAD) model. Due to different training requirements and to the perception of FACS’ objectivity and scientific rigor, data workers who perform these types of annotation are often referred to as either “expert coders” (in the case of FACS) or “lay annotators” (in the case of VAD). However, despite practitioners’ perceptions of these data practices, both FACS and VAD coders must undergo a process of interpretation and negotiation of affective data. I have borrowed from anthropological studies of professional vision (Goodwin 1994; Grasseni 2004) to describe the “skilled visions” of both FACS and VAD coders (Grasseni 2004). Then, I have discussed the role that material factors such as tools and interfaces play in shaping annotators’ ‘ways of seeing’ the data. In particular, I have discussed the role that automated labelling and explainable AI techniques play in the effort to reduce ambiguity and uncertainty in affective labelling. I then argued that practitioners’ perceptions and representations of data work can further obscure the struggles and disagreements among annotators. I concluded with a reflection on the political role that uncertainty plays in the creation of ground truth datasets. Following critical scholarship on power and labor in data

structures (Miceli and Posada 2022; Raval 2021; Irani 2015a; Gray and Suri 2019), I have argued that by foregrounding annotators' disagreement and by making it an integral part of the machine learning pipeline, we can challenge the claims of objectivity and universality on which machine learning models rest.

The next and final empirical chapter explores three mental health applications of FER technology: mood tracking, depression detection, and autism diagnosis and treatment. Here, I bring to the fore the various social and cultural representations of affect, mental health, illness, disability, and care that inform these projects, as well as the prescriptive and normalizing dimensions of these tools.

Chapter 7: The promise of Affective Computing

Introduction

In the inaugural chapter of the *Oxford Handbook of Affective Computing* mentioned in chapter 4, Rosalind Picard narrates her encounter with Jodie, a young woman who she had met at a retreat organized by the autistic community. Picard describes Jodie, who had a diagnosis of autism spectre disorder, as struggling with stress and anxiety “when unpredictable things happen” (Picard 2015: 11). To help her better understand her own affective ebbs and flows, Picard offered to monitor Jodie’s physiological signals through a wristband that measured skin conductance, motion, and body temperature. “Many people”, Picard reasoned, “don’t have an accurate read on what they are feeling (this is part of a condition known as alexithymia)” (Ibid.). Affective computing could change this. Tracking her physiological data made Jodie realise what strategies comforted her most when facing a stressful event. For example, the wristband’s reading of her signals informed her that pacing was more soothing than “stimming” (the repetitive movement of rocking back and forth). Affective computing promised to advance our understanding of our own inner life, as well as improve our understanding of others’.

Jodie’s appearance in affective computing’s origin story is not an ornamental anecdote nor a mere coincidence. As discussed in this chapter, practitioners have long regarded autism as a prolific area for computational intervention. Over the years, these efforts have been extended to include mental health conditions (depression and anxiety) and other neurodevelopmental disorders (ADHD and bipolar disorder). Today, both industry and academic practitioners see (mental) healthcare as a promising area for the application of FER technologies. Here, computer vision is

harnessed to detect and monitor unusual changes in facial expressive behaviors that, according to a range of psychological theories discussed in the chapter, are assumed to correlate with mental states (such as depressive episodes) or neurodevelopmental conditions (such as autism), with use cases spanning from unregulated “health and well-being” mood-tracking applications to clinically trialled “diagnostic aids”. Accompanying these tools are “promissory claims” (Pickersgill 2019) of “state-of-the-art” technology that can provide “objective and accurate diagnoses” and “improve patients’ health” and “reduce clinicians’ costs”.⁷² These efforts can be seen as part of a broader trend, called “digital phenotyping”, that aims to use digital technologies – and particularly smartphone sensors – to provide objective measures of health conditions (Insel 2017; Jain et al. 2015; Baumeister and Montag 2023). The increasing prevalence of mental health conditions caused by the Covid-19 pandemic, coupled with the growing demand for remote diagnoses and treatments, facilitated the rapid development of a digital phenotyping market for mental health. Here, these “promissory devices” (Birk et al. 2021) are claimed to address multiple (and at times competing) problems at once: offer more personalized healthcare, improve accessibility to treatments, cut costs, reduce clinicians’ workload, and create profit.

These are the logics that drive the projects discussed in this chapter: FER systems designed to track and predict mood changes (as described by Marcus), depression and anxiety (as described by Marcus, Nathan, Marvin, and Nash), and autism and bipolar disorder (as described by Marcelo and Otis), based on facial expressive behavior. In recounting their motivations and aspirations, participants mobilized a range of claims regarding the potential therapeutic benefits and societal impact of their projects. These are what Martyn Pickersgill called “biomedical virtues” that is, “legitimising tropes for promises made” (Pickersgill 2019: 18). Underlying

⁷² These claims appear on the websites and marketing material of the companies I have interviewed, left anonymous for confidentiality reasons.

these claims are, however, social and cultural representations of affect, (mental) health, illness, disability, and, ultimately, care. As Lucy Suchman has argued with regards to AI systems, “Positioned as exemplary of leading-edge thinking and technical practice, these initiatives in new technology materialize the cultural imaginaries that inspire them and which they work in turn to enact” (Suchman 2006: 226). By trying to de-code and re-code human experience, what ideas of the “affects” and “health” do practitioners leverage? As I have argued in chapter 4, affective computing – and particularly FER technology – can facilitate the stabilization of particular affective behaviors (and related logics of feelings) that conform to dominant socio-cultural norms. This becomes particularly evident in mental healthcare, where FER systems are used to track, predict, and differentiate between healthy and ill behaviors, thus normalizing particular ideas of health and care, as well as illness and disability. Writer Xiaowei Wang argued that embedded in these projects is a notion of *cure* rather than *care*. As they have noted, the “logic of cure has a deep-seated finality — we find solutions and find cures, unlike the ongoing process of care” (Wang 2021). In this sense, technologies of cure rely on the “reiteration of what is ‘normal’ within medicine — normal bodies, normal symptoms, normal treatments” (Ibid). These representations of care-as-cure, however, cannot be seen as separate from the social, political, and economic forces that underpin them. To “move away from the logic of cure and towards a logic of actually caring for patient needs, we need to recognize the economic impetus behind machine learning and AI in medicine, and how the political economy of medical AI systems often rely on health inequities to exist” (Wang 2021). The optimization of healthcare, and its promise of cost-efficiency and scalability, is often seen by practitioners as a solution to wider socio-economic inequities in healthcare access and provision (Ibid).

This chapter brings all these dimensions to fore. I focus on three broader application areas in mental healthcare that FER systems promise to revolutionize: mood tracking, depression detection, and autism diagnosis and treatment. Here, I trace

the undercurrent of discursive and cultural repertoires that sustain these projects: neoliberal imperatives of happiness, normative assumptions around disability and gender, as well as Western theories of the affects that are drawn from “experimental psychology and freely mixed with Silicon Valley notions of human optimisation and venture capital logics” (Hagerty and Garofalo 2021). To do so, I move back and forth between the interview data and the recent history of affective computing, showing how these dynamics were present since the inception of the field (and its market). Finally, the chapter ends with a discussion of the notion of care encoded in the projects described throughout. In tracing these dimensions, I hope once again to contest the notions of objectivity and universality on which these technologies (and their promissory claims) rest. To return once more to Wang,

Transforming AI systems is not just confined to the realm of engineering ethics conferences. Working from the mosaic of cure and care, we can recognize the ways individuals and industries are situated in broader systems of capitalism, and that there are numerous entry points into transforming the ways AI is shaped, made, and deployed (Ibid.).

Know Thyself

When I spoke to CEO Marcus in the last months of 2020, his team was developing an app that would make use of their proprietary facial expression recognition model to help users monitor and track changes in their mood. While mood and mental health trackers are a well-documented phenomenon and part of the larger cultural trend of the “quantified self” (Lupton 2016), Marcus had a more specific vision for his technology. As he explained to me, the app was intended for the “pregnancy and fertility market”, as its ultimate purpose was to help women monitor their stress and mood levels to facilitate conception and provide mental health support both during and

after pregnancy. The tool was originally meant as a “health and wellbeing” app, rather than a clinical device. This way, the company would not be required to undergo a clinical trial regulation and could launch the product sooner in the next year. However, Marcus mentioned that through his company’s partners, there would be opportunities to connect the app to a midwife, a counsellor, or a fertility consultant.⁷³

To collect mood data, the prospective user would interact with the app by talking to a virtual agent, reading out loud prompts and playing a series of psychometric games. These interactions would be recorded using the user’s smartphone camera for data labelling, model training, and mood prediction. By designing the app in the form of a “mood diary”, with daily tasks and prompts, Marcus’ company would be able to collect longitudinal data about the user and refine the ability of the algorithm to provide accurate predictions. To classify and analyse facial expression data, they would adopt both FACS and VAD coding, in the ways described in chapter 6. Here, practitioners would not be concerned with the accurate prediction of categorical dimensions of affective experience (‘happy’ or ‘sad’) but, rather, with the ability to detect changes in behavior and promote a healthier lifestyle, as illustrated in chapter 5.

Marcus’ project was premised on two underlying assumptions: first, it embraced the Affective Computing and HCI motto of “*know thyself*”, which suggests that by quantifying and analysing real-time data about our health and behavioral experience we can improve our wellbeing (Riva, Calvo, and Lisetti 2015). As another participant described, “things like mood management or emotion management are good and powerful and they can help people become happier and better people. Ah, of course they can be used for bad purposes as well.”⁷⁴ This behaviorist push towards better health and happiness can be read as part of the positive psychology trend that

⁷³ For legal and commercial reasons, at this stage Marcus was not allowed to name the partners involved in the project.

⁷⁴ Interview conducted with Cedric on November 23rd, 2021.

permeates Western societies and often underlies neoliberal imperatives of hyper productivity (Davies 2015). Similarly, feminist scholar Sara Ahmed has pointed to the “moral quality” of happiness, that is, the association of happiness with the “normal” and “good” functioning of society. Key to this “happiness duty” is the idea that happiness (and other feelings) can be measured and manipulated (Ahmed 2010).

Second, Marcus’ app relies on the assumption that, by tracking changes in facial emotional expressions, users can obtain an accurate and helpful portrait of their own mental health. Building on Marx’s economic theory, Jessica Feldman has called this phenomenon “computerized alienation”, the marketization of the self “for evaluation and recognition, according to the terms of the computer” (Feldman 2016: 18). Inscribed in these practices are theories of human identity that have a predictive and prescriptive effect, in that they are grounded on what the technology wants to discover, be that “motivating drives (for advertising), unconscious discomfort (for lie-detection and investment planning), and mental health tendencies (for benefits administration)” (9). As Feldman points out, the combination of advances in machine learning and affective computing has enabled a shift from computers’ mere attempt to recognize users’ affective response to the idea that the computers can reveal something that even the users don’t know: their own affective experience. Consistent with affective computing’s promise, FER systems seem to offer a privileged and unique view into the user’s inner life. A final point of consideration is the notion of (self-)care encoded in these projects. As I discuss more in depth in the following sections, tools that track and predict behaviors presumed to correlate with mental states can reinforce normative assumptions about mental health and affective expressions.

The search for objectivity

While the app was not originally intended as a medical device, Marcus's long-term goal was to turn it into a "diagnostic aid" that could support healthcare providers in making "more accurate" and "more objective" diagnoses of mental health conditions such as depression and anxiety. At this point, our conversation seemed to have reached a conceptual impasse: until a few minutes earlier, Marcus had been very cautious in describing his work on facial expression recognition as exactly and only that: recognition of facial expressions, not internal states. The idea that his system could be used to detect mental states such as depression and anxiety seemed to me counterintuitive. He promptly explained to me that his model would of course be unable to 'see' internal states but would be capable to pick up on changes in facial muscle activity typically associated with depressive states. This approach is grounded on a theory of clinical psychology called "psychomotor retardation", a condition that can be observed in depressed patients and that is characterized by decreased speed and range of muscle activity including slowed speech, reduced facial expressions, eye movements, and overall diminished body movements (Buyukdura, McClintock, and Croarkin 2011; Bennabi et al. 2013). Psychologists often locate the first, and most notable, observed instances of this symptomatology in Charles Darwin's 1872 *Expression of the Emotions in Man and Animals*. In illustrating the physical reactions of those who have experienced loss and grief, he described:

As soon as the sufferer is fully conscious that nothing can be done, despair or deep sorrow takes the place of frantic grief. The sufferer sits motionless, or gently rocks to and from; the circulation becomes languid; respiration is almost forgotten, and deep sighs are drawn. All this reacts on the brain, and prostration soon follows with collapsed muscles and dulled eyes. As associated habit no longer prompts the sufferer to action, he is urged by his friends to voluntary exertion, and not to give way to silent, motionless grief. Exertion stimulates the heart, and this reacts on the brain, and aids the mind to bear its heavy load (Darwin 1998: 80-81).

In 1976, psychologist Gary Schwartz and colleagues set out to test Darwin's observations by recording the changes in facial movements of patients diagnosed with depression and compare them to those of non-diagnosed ones. Since facial micro-movements, as postulated by Ekman, take place too quickly for the naked eye to accurately detect them, and because computer vision techniques were not yet widely available, Schwartz opted for electromyography (EMG) to record facial muscle activity. As shown in Figure 8, the procedure – which recalls Duchenne's electrical stimulation of facial activity described in chapter 4 – involves the application of electrodes to the subject's face to measure muscle contraction. According to Schwartz, this technique allowed researchers “to discriminate when subjects were self-generating happy, sad, or angry thoughts and feelings, even when no differences were readily apparent in the overt face” (338). The study results, conducted with “12 depressed subjects and 12 matched normals” (337), concluded that participants diagnosed with depression showed less intense expressive patterns, particularly those associated with positive expressions. Since 1976, Schwartz's study has not been repeated. However, in 2000, psychomotor retardation was included in the fourth edition of the *Diagnostic and statistical manual for mental disorders* (DSM-IV) as one of the nine core symptoms of major depressive disorder (Buyukdura, McClintock, and Croarkin 2011).

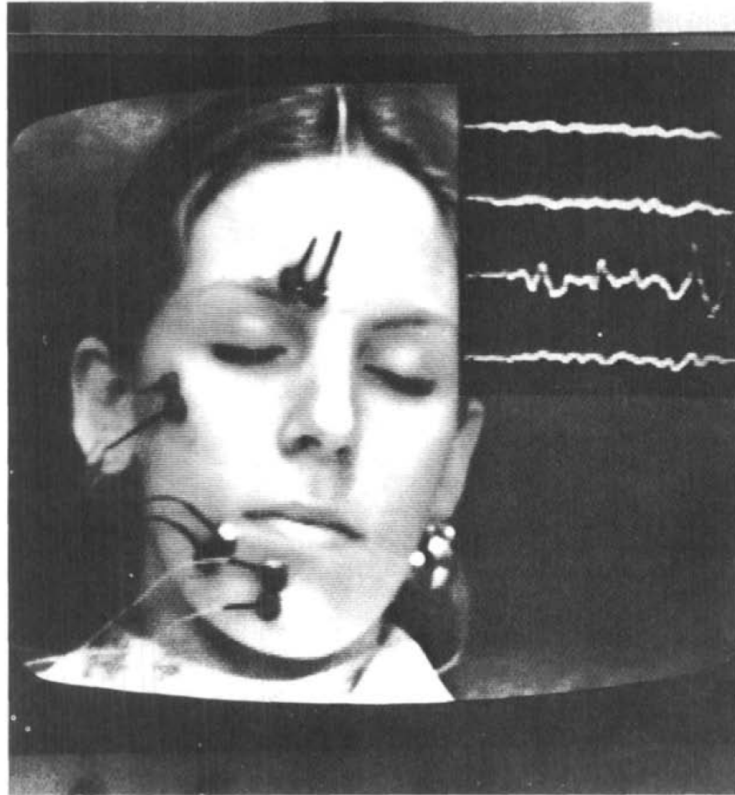


Figure 8 Electrical activity of four facial regions detected by EMG sensors, as shown in (Schwartz et al. 1976).

Today's computer vision techniques seem to offer a more efficient, cost-effective alternative to Schwartz's EMG techniques. In 2013, the third edition of the Audio/Visual Emotion Challenge and Workshop (AVEC) – an international competition aimed at the advancement of audio-visual emotion analysis techniques – opened with a set of pressing questions:

Are differences between depressed and non-depressed persons specific to depression or are they common to the types of people most likely to become depressed? Depression is strongly related to stable individual differences in neuroticism, introversion, and conscientiousness. Differences in nonverbal behavior between those with and without depression could indicate personality differences rather than depression. Do they? What can non-verbal behavior tell us about possible functions or mechanisms of depression? (Cohn 2013: 1).

As the keynote speaker announced, the AVEC challenge was, for the first time, entirely dedicated to the task of detecting depression from non-verbal behaviors. Since then, research on the automated detection of depression has received growing attention, with publications in this area rapidly increasing between 2017 and 2020 (He et al. 2022). This seemingly sudden fascination for automated depression detection can be seen as part of a larger project of digitisation of psychiatry that was consolidated, as many have noted, by a 2015 paper published in *Nature*, titled *The digital phenotype* (Jain et al. 2015). In the paper, the authors borrowed from evolutionary biologist Richard Dawkins who had argued that the concept of phenotype (a person’s observable characteristics) should be extended to include “all effects that a gene has on its environment inside or outside of the body of the individual organism” (462). Similarly, Jain and colleagues applied the concept of extended phenotype to all interactions that users have with technology. Parsing these digital traces – the activity registered through keyboard interactions, smartphone sensors, and so on – could help monitor and detect the early onset of health conditions in an unprecedented fashion. In 2017, former director of the National Institute of Mental Health (NIMH), Thomas Insel published a widely cited paper, *Digital Phenotyping: Technology for a New Science of Behavior* (Insel 2017), extending the concept of digital phenotyping to mental health.⁷⁵ As he suggested,

After 40 years of psychiatry becoming more mindless than brainless, perhaps digital phenotyping will help the pendulum swing back toward a fresh look at behavior, cognition, and mood. It has been said that new directions in science are launched by new tools much more often than new concepts. In this case, a tool that is inexpensive and ubiquitous may change the direction of the field (1216).

The rapid proliferation of mental health tracking apps and services gave rise to the field of “mHealth” (or mobile health), “a key facet of this matrix of illness, optimisation and

⁷⁵ As media scholar Luke Stark has noted, Insel had left his leading role at the NIMH to work at Google’s parent company Alphabet (Stark 2020).

technology” (Pickersgill 2019: 21), which extended its reach, as many of my participants bear witness, to facial expression recognition. The data captured through camera sensors would allow psychiatry to “return to the body for information about the mind” (Stark 2020: 304), in an ever going search for objective markers for mental health and illness (Harrington 2019; Pickersgill 2010).

Over the years, researchers in both the computer and psychological sciences – an interdisciplinary new field called “psycho-informatics” (Markowetz et al. 2014; Baumeister and Montag 2023) – have developed various deep learning models to assess the severity of depressive episodes or the efficacy of anti-depressant treatments by measuring the muscle activity of facial expressions. In research settings, FER systems measure reductions in facial expressivity of undiagnosed subjects against subjects diagnosed with depression or, within the same subject sample, by monitoring changes in facial activity over time (such as before, during, and after anti-depressant treatments). Like in Schwartz’s original study, existing diagnoses of depression represent the baseline against which the system’s accuracy is measured. The commercialization of these methods, however, seem to rely on the assumption that these studies’ findings can be extended to the general population. As Nash, a machine learning engineer working at a start-up, described:

[psychomotor retardation] that's the theory. What we have seen in the existing scientific literature, there is a strong correlation in terms of the apparent expression and the state of mind, how the person is feeling, in terms of the actual diagnosis of depression itself. So, that's what we want to see, how well the models will perform in actual practice. Whatever studies have been done till now it's in the lab, so we'll need to see how well we can do it and in a much, much wider population, so that's what our clinical trials will help in studying them.⁷⁶

⁷⁶ Interview conducted on July 16th, 2021.

Underlying these efforts there seem to be assumptions about mental health that leave differences in affective behavior largely unaccounted for. Intuitively, facial expressivity can be shaped by individual and cultural differences (people who appear generally “less expressive” when compared to other individuals or social groups), disability and chronic pain, and other types of medications. However, psychomotor retardation seems to promise to finally establish that causal relation between facial expressions and mental states that FACS alone could not fully guarantee. As Marvin assured me,

[...] certain behaviors, when you observe them, will have a direct link to an illness. For them, you can look at that behavior out of context, not knowing what the person was talking about, what their age, who they are, who are they talking to? Like, very out of context, and still learn something or see some signs potentially of either an illness or symptom of an illness. And a good example is that depression has psychomotor retardation. You naturally don't move as much. So, then your vowels, when you pronounce your vowels, they will sound more alike because you don't open your mouth as much. Possibly your vowels will sound more alike. So, then you can measure that, and that becomes the marker.⁷⁷

Disability scholars have noted how automated systems that aim to detect signs of mental illness can reinforce and reproduce normative models of disability and mental health. In particular, algorithmic tools that detect or assess mental health and disability based on out-of-context data often overlook how other aspects of identity (e.g. race, class, gender, and other chronic health issues) can affect people's experience of illness or disability (Mills and Whittaker 2019). In addition, even when measuring against existing diagnoses of depression, many have noted how available data on mental health might not be representative of the general population. In many countries, public stigma

⁷⁷ Interview conducted on February 3rd, 2022.

around mental health has been linked to under-reporting and under-coding of mental health disorders, resulting in datasets which are not representative of the true extent of prevalence of these conditions in the overall populations (Walsh et al. 2020). Other social determinants can affect access to diagnosis and treatment. Recent research has demonstrated that Western mental health assessment tools and protocols fail to account for non-Western expression of distress (Straw 2020). By reducing mental health to its assumed physical manifestations, practitioners risk overlooking the socio-cultural contexts in which illness is experienced.

If I am allowed, once again, to draw a comparison between these practices and the historical controversies over the validity of phrenological claims, it is perhaps possible to observe a similar rhetorical move: much like in 19th century phrenology and its focus on skull anatomy, the emphasis on an anatomical explanation for the inference of mental states – with the added benefit of psychiatry’s validation – helps practitioners “secure social credibility” (Shapin 1979: 168). As historian of science Steven Shapin noted in his own account of the historical disputes over the validity of phrenology,

Credibility may be secured by the production of apparently naturalistic knowledge, in which it is impossible for opponents (or historians) to discern social interest, but not *any* naturalistic account will best further a particular social interest. Certain natural realities are better for some purposes than others (Ibid.).

The theory of psychomotor retardation becomes here the naturalistic explanation conferring scientific legitimacy to FER practice. However, multiple logics drive the development and commercialization of these projects. As Martyn Pickersgill has noted, digital psychiatry received its legitimacy from the constitution of “promissory statements” with regards to its therapeutic applications (Pickersgill 2019). Through the deployment of these promissory claims – the strategic mobilization of what Pickersgill calls “biomedical virtues” (Ibid.) – practitioners strive to establish both their professional identity and their field of practice. The next sections examine participants’

visions of affective behaviors, mental health, therapeutic treatments, and the market logics that complement them.

“It ticks all the boxes”

When I asked what personal or professional trajectory had led him to create a FER model for maternal mental health, Marcus first offered me a somehow unsurprising response:

In many ways, it's not so much that the commerce.. it's not so much the monetization, but it's the ability to take the technology that we've created to the next level where it's actually used by people, and clearly I wanted it to be used for good, for good purposes. But perhaps even more important for me is just the fact that it is being used to the best possible use by as many people as possible.⁷⁸

I do not intend to imply that Marcus was not acting out of genuine concern. Yet, this type of responses reflects the techno-optimism that permeates the field of machine learning already described in chapter 5 and well documented by others (Tutton 2021; Pfister 2017). In addition, scholars have noted how the health sector represents a particularly fertile ground for practitioners' declared pursuit of social good. As Lehoux et al. have argued,

Innovation in the health sector is endowed with a particular moral meaning. Using the powers of a corporate entity to create and commercialise a medical technology that is intended to benefit patients and society more broadly involves a value-laden action that may, as a whole, be seen as contributing to the common good (Lehoux et al. 2014: 739).

⁷⁸ Interview conducted on December 10th, 2020.

However, these claims are complicated by additional economic and organizational logics – which the authors call “regimes of engagement” (Ibid.) – that orient practitioners’ action. Indeed, in recounting his experience, Marcus added some significant detail:

And then we went on a journey of seeing what we could do with our face analysis and that led us to identify that the area of medical conditions that change your expressive behavior. We had the largest USB [unique selling benefits], unique selling points in that area. And that's where it went. And then we had a period of time exploring that further. And then we ended up focusing on perinatal mental health and health and well-being leading up to pregnancy as well. And, of course, for some period after pregnancy. So that's a key area.⁷⁹

This type of response shows that practitioners’ choice with regard to technology design and development are shaped by pragmatic and situated judgments, “modulated by the reality checks of the corporate context in which their collective action unfolds” (Lehoux et al. 2014: 751). Another participant, Nathan, described a similar calculation, uncovering both the economic and technical drivers that led his team to choose the mental health sector over other application domains. As he described, guiding the company’s direction was primarily the readiness of available tools and techniques (algorithms that “work on depression”), followed by the identification of a specific market segment where these tools can generate the most economic value:

We came up with depression *because we had reasonably good evidence that the algorithms and things we've got work on depression* [emphasis added]. But then you start looking out there and you kind of think "Ok, well, do we want to build a product that measures depression? OK, so what's its use case? Where's its use? Where's its value?".

⁷⁹ Ibid.

Because while it is applicable to the broad spectrum, you kind have to prove it and regulate in a particular area first that we had to narrow down, we couldn't just build a tool that was.. Broad. So, we have to choose a segment. And we considered lots of different segments. What was considered quite a lot was the depression in young people and in terms of like a target market, because they're much more technologically savvy. A lot of people are very interested in it. But unfortunately, economic case didn't stack up for that one.⁸⁰

This response reflects a view of technology that sees social and health issues primarily as viable business-cases, in a re-configuration of the classic engineering “problem → solution” approach into a “solution → problem” one, where available tools and technical competences orient and direct practitioners’ action. Underlying this view are economic incentives and financial constraints that compel practitioners to focus on application areas that can more easily generate revenue. As Nathan explained, small start-ups require business models that can support the company’s growth:

Because we're building business, we need to grow, we need more people. We've got to get to a point where we can fund...it [the technology] self-funds. And when we did that analysis, which areas were probably quickest to generate revenue, in the health sector. It was obviously a good one to go for.⁸¹

This pursuit of profit is not in conflict with practitioners’ professed moral principles. As Lehoux et al. have noted, these “orders of worth” are not necessarily mutually exclusive: “As long as technology developers genuinely believe in their efforts serve health, striking market-oriented trade-offs makes sense to them” (Lehoux et al. 2014:

⁸⁰ Interview conducted on April 21st, 2021.

⁸¹ Ibid.

750). As described in Nathan's response, "health and wealth" (Ibid.) were two pieces of the same puzzle:

And so, *kind of ticks all the boxes* [emphasis added], it's a huge problem and it's got a huge cost attached, but it's known that if you can intervene early, you get better outcomes. And then I suppose the last piece of the puzzle was actually knowing that quite a lot of it was missed. So unless it gets very.. The more serious, the higher the detection rate gets. But there's quite a lot of kind of low to medium level depression that isn't treated at all...When we did the analysis, it just, *everything just kind of clicked* [emphasis added]. So that's why we ended up with that.⁸²

Some participants, however, reported an inversion of these dynamics. Quentin, for example, was developing a piece of hardware technology that would combine FACS coding, electromyography (described in the previous section), and virtual reality (VR). This system would track changes in facial movements (elicited through VR scenarios) to help patients who suffered from facial paralysis re-train their facial expressions. As he described, however, the projected costs for each single headset were too high and would not generate the revenue required to support his company. For this reason, his business model focused on selling a general-purpose FER system to a wider range of clients. He would then use a portion of the profits to support the R&D for his healthcare applications. As he described,

My core is still patients with facial paralysis, but recognizing that the unit cost of each system will be very, very high. But if I can broaden the market for it and create value by people being able to use it, then that basically means that I'll be able to kind of reverse back to that core problem I was looking to solve at the beginning by having...Through economy of scale, the system is much cheaper, and also because in order to get developing

⁸² Ibid.

hardware is very expensive and to get investment to develop the technology to help those patients. Investors aren't interested in a very narrow use case. They're interested in a broader market.⁸³

As I discuss in the next section, participants mobilize this “promissory bioeconomy” of healthcare (Pickersgill 2010) in various ways. These dynamics can be observed in the original inception of the affective computing market, born – as described in chapter 4 – out of Rosalind Picard and Rana el Kaliouby fortuitous encounter in the early 2000s. Central to affective computing’s origin story is a seemingly benign (and almost humanitarian) medical application of FER technology that promised to revolutionize psychiatric care for neurodevelopmental conditions. Underlying these techno-optimistic visions of affective technology are, however, normative ideas of health and illness, as well as gendered assumptions about expressive affective behaviors. The next section traces these tensions before concluding with a reflection on the notion of care encoded in these projects.

Normative visions of health and gender

As described in chapter 4, both Picard and el Kaliouby wanted to apply their research to the health sector and had chosen autism as their primary area of intervention. Building on el Kaliouby’s doctoral work, they developed a FER system called MindReader that would help autistic people interpret their interlocutors’ affective cues.⁸⁴ In a 2006 paper co-authored with the Director of the Cambridge Autism

⁸³ Interview conducted on February 16th, 2022.

⁸⁴ I follow here disability scholars and critical scholarship on disability and technology in adopting an identity-first approach to discuss autism and disability (Mills and Whittaker 2019; Spiel et al. 2019).

Research Centre, Simon Baron-Cohen, the researchers made the case for the use of affective computing for social interactions in autism by comparing autistic people to computers:

While much of the work in affective computing has been motivated by the goal of giving future robots and computational agents socioemotional skills, its researchers have also recognized that they face similar challenges to those who try to help people with autism improve such skills. Computers, like most people with autism, do not naturally have the ability to interpret socioaffective cues, such as tone of voice or facial expression. Similarly, computers do not naturally have common sense about people and the way they operate (el Kaliouby, Picard, and Baron-Cohen 2006: 3).

Interdisciplinary research on autism and affective computers, the authors argued, could help advance our understanding of affective behaviors, communication, and social interactions and improve both computer and autistic people's ability to read and respond to socioaffective cues. Computers and autistic people, they reasoned, systematize information and recognize patterns in similar ways. As they described,

Persons diagnosed with ASC [autism spectrum condition] are extreme systemizers, showing intact or superior systemizing abilities, such as excellent attention to detail, islets of ability in topics like prime numbers, calendrical calculation, or classification of artifacts or natural kinds (4).

According to this framing, by adopting “systematic approaches to teaching empathy” (Ibid.), the researchers hoped to improve both social interactions in autistic people and computers' affective capabilities. This emphasis on systematization is grounded in the theory of the “extreme male brain”, a notion that Baron-Cohen had developed in his book, *The Essential Difference: Male and Female Brains And The Truth About Autism*, famous for its deterministic views on gender differences and how these affect autism (Baron-Cohen 2003). This theory posits that, due to different in-utero exposures to hormones, male brains are predisposed to systematization and logical

reasoning, while female brains are more inclined to empathetic behaviors and social interactions. As disability scholars have noted, however, this gendered vision of autism ignores the ways women and non-binary people experience autism and how this might lead to underdiagnosis and unequal access to treatment (Spiel et al. 2019). In addition, as psychologist Cordelia Fine has argued, these stereotypical views on gender differences contribute to a form of longstanding scientific sexism that she termed “neurosexism” (Fine 2005). Stereotypes about the “gendered brain” are indeed rooted in historical attempts to translate social hierarchies into scientific differences such as gendered variations in brain and skull size (Schiebinger 1990), and are still reflected in today’s disparate medical treatments of women and BIPOC people (Williams and Rucker 2000).

Picard and el Kaliouby seemed to acknowledge the normative risks of their assistive technology. As they noted, “one issue to consider is whether individuals with autism need treatment or technology ‘fixes’ at all” (el Kaliouby, Picard, and Baron-Cohen 2006: 14). However, practitioners’ therapeutic promises continue to be permeated by similar assumptions and expectations about autism and disability. For Marcelo, affective computing’s role in this context was twofold: akin to Picard’s original vision, assistive technologies would help disabled people to “communicate more effectively” and, at the same time, it would allow neurotypicals to better interpret autistic people’s behaviors and intentions:

People with autism, as well as other disabilities, sometimes they lack the capability of communicating how they are feeling, right? So if you have a wearable that allows you to communicate more effectively, let's say with your significant other or your stakeholders, your family members, that can be very helpful to.. For example, we have seen in some of our research that people with autism, you may see huge spikes on their physiology, but you don't see anything outside. You can't see any facial expressions, any discomfort. But then after a few seconds, after a few minutes, they start engaging in self-injury behavior

and all of these things that are very problematic. Right? So, what the data shows is that these episodes of high intense stress starts much earlier than you can actually see them. It may be that a plane is flying by, and they are very sensitive to the noise, or it may happen that they don't like something that it's in the room, but they cannot communicate it. So that helps you to better understand what is going on.⁸⁵

In this regard, disability scholars have argued that technologies that target autistic people can reinforce the expression of affective behaviors and modes of interaction aligned with dominant socio-cultural norms of expressions and, despite practitioners' initial cautiousness, can reproduce "normative expectations of a neurotypical society, which predominantly views autism as a medical deficit in need of 'correction'" (Spiel et al. 2019: 1). Affective computing perpetuates those same social norms of interaction that autistic people are presumed not to be able to adhere to. This prescriptive and performative element of affective computing was even more explicit in some of my participants' responses. As Otis recounted, one of the projects he was working on adopted FER technology to improve bipolar women's interactions with their own children:

One good example of that is that we've seen a lot of research done in autism. And, also, bipolar disorder. And one of the challenges there is the inability to communicate your feelings to others. So, for example with autism, they have a really, really hard time decoding what other people are trying to express. Another very great example.. I'm currently running a training study here in [European city] with a group of researchers interested in bipolar disorder. More specifically, with mothers that have been previously diagnosed with a bipolar disorder and have just given birth to a child. So, in that scenario, it's really important that if you are the mom, that you can mimic and that you can communicate to the child what are your feelings so that there is.. *Because that's the only*

⁸⁵ Interview conducted on July 17th, 2021.

way of communicating [emphasis added]. That's actually by facial expressions trying to comfort the child or just trying to mimic 'You're feeling sad and I'm going to do something about it'. So, you see it in this mother-to-child interactions that facial expressions are the way that you communicate with the child before you learn to speak. And one of the challenges with bipolar disorder is that they are actually very poor at communicating. So, we've developed a training paradigm where we're actually showing them different videos. And the idea is that they try to mimic the child's reaction and we're giving them feedback on whether or not they did a good job. So, I think in that scenario, it makes a lot of sense as a form of social communication.⁸⁶

The long quote unearths two distinct, but related, discursive repertoires: first, the emphasis on facial expressions and mimicry presumes the primacy of visuality over other forms of communication and care. In this scenario, mothers are expected to care for their children primarily through vision, rather than other forms of sensory perception such as touch or hearing. In her exploration of care, feminist and STS scholar Maria Puig de la Bellacasa has suggested that touch, for example, (the “haptic”) offers an alternative form of thinking about care as it “holds promises against the primacy of detached vision, a promise of thinking and knowing that is ‘in touch’ with materiality, touched and touching” (Bellacasa 2017: 104). Further, vision is here mediated by FER models – with their own ways of ‘seeing’ affective information. As I have shown here and in chapter 6, these systems’ vision rests on infrastructures of knowledge (the various theories of the mind, affects, and disability), human interpretation (the work of meaning-making performed by annotators), and power (the organizational and social structures that shape these data practices). As feminist studies of science have long argued (and as mentioned in chapter 2 of this thesis), the act of vision is not neutral but always practiced within a social and political framework. To

⁸⁶ Interview conducted on June 14th, 2021.

borrow again Donna Haraway's words, we must ask "Where to see from? Who interprets the visual field?" (Haraway 1990: 587).

Second, the emphasis on the maternal and the necessity to coach women into "better" and "healthier" parents uncovers concurrent assumptions around motherhood and disability. As disability scholar Harriet Cooper notes, "we don't expect a mother to be a disabled person, still less a disabled person to be a mother (to draw a subtle distinction). But perhaps the struggles of each intersect in more ways than we would like to think" (Cooper 2020: 6).⁸⁷ Once again, underlying these tools are normative ideals that can reinforce existing discriminations and forebode equal access to care. In the final section, I offer some reflections on the notion of care engendered in these tools.

What kind of care?

A recurring concern shared by participants was the fact that searching for indications of depression and other mental states by looking at people's facial expressions meant observing them in their most vulnerable moments. What if a user was showing worrying signs of distress? Or what if they were harming themselves? As Marcus described, his company had to weigh these concerns against its legal obligation to privacy:

⁸⁷ Building on Foucault's notion of biopolitics (M. Foucault, Davidson, and Burchell 2008), philosopher Paul Preciado has argued that "it is in the field of reproduction—sexual, social, cultural—that we confront the most crucial dimension of contemporary power" (Preciado 2018), with reference to recent biomedical developments in genetics and artificial reproductive technologies. For Preciado, these technologies reproduce neoliberal forms of governmentality as they perpetuate heteronormative assumptions around sex and gender. In a similar manner, we could read the projects described here by Marcus and Otis (with their emphasis on the gendered performativity of affects and normative assumptions around motherhood and disability) as part of a broader array of technologies that aim to regulate the reproduction and maintenance of life.

We take a massive privacy-by-design and security-by-design approach to developing all our products, but this has led to interesting situations where actually we are unable to contact somebody if we see something, anything happening in a video, we would not be able to link that video to a person. So, because of GDPR, we cannot take a, let's say, a healthy decision. So, we are now wondering whether there needs to be a discussion about the trade-offs between GDPR on the one hand and health and safety on the other hand.⁸⁸

This response reflects the tensions between care and privacy – the duty of care as opposed to the duty of confidentiality – amplified by the Covid-19 pandemic, where privacy and data protection regulations have at times been upended by both governments and corporations in the name of public health and crisis management (Zwitter and Gstrein 2020). The pandemic has also drawn attention to the growing central role that tech corporations play in the distribution of welfare and care resources. As many have noted, tech companies have positioned themselves as increasingly essential to the functioning of society (Couldry and Mejias 2019). In the healthcare context, this can often mean that proprietary and commercial tools become “obligatory points of passage” (Callon 1984) to access diagnoses and treatments. When tracking devices become an essential part of patient care plans, privacy is “something that disabled people aren’t able to choose” (Mills and Whittaker 2019: 23). Companies can extract value from user data that disabled and ill people cannot refuse to share. Related to this is the question of “bystander privacy”, that is, privacy concerns that arise with the use of wearable and mobile computer vision technology that can record bystanders’ activity without their consent. For Marcus, however, this presented again as an issue related to the duty of care:

⁸⁸ Interview conducted on December 10th, 2021.

Let's say we see something happening in the background, right? Well, I think.. You know, with Zoom at the moment, we've all had interesting things happening in the background, and we are basically unable to follow up because we have promised our users that we would not link their identity to that data. And so, those are interesting new challenges that we hadn't considered in the first place. And the question is, is that ethical?⁸⁹

Entering the algorithm's frame of vision is not just the prospected user, but all those exposed to its totalizing view. I share here the concerns of the many scholars who have argued how pervasive data collection practices can lead users to accept potentially harmful surveillance practices simply because they are widespread and routinely implemented by tech corporations (Mills and Whittaker 2019).

As argued in this chapter's introduction, the technologies described here seem to replicate a model of *cure* rather than *care*. In addition to normative ideals around health and gender, they advance a notion of care that sees technical interventions as solutions to wider socio-economic inequities in healthcare access and treatment. This is best illustrated in Nash' response, where he points to the shortcomings of public healthcare as a reason for the automation of mental health diagnosis and treatment:

I think it [the depression detection model] definitely can be quite helpful because right now our health system is like, it's not able to give enough time to everyone. If our system can help in the early detection of mental health conditions, that would be also quite useful because the earlier you could detect these conditions, the greater the chances that you can treat these conditions medically. The later you wait, and the intensity of that mental health condition increases a lot, the more difficult it gets to find a way to basically find a cure for it or to manage it.⁹⁰

⁸⁹ Ibid.

⁹⁰ Interview conducted on July 16th, 2021.

As many have noted, however, automation is seldom the appropriate response to complex structural issues (such as the lack of adequate care infrastructures). As argued throughout the chapter, these tools can replicate normative and discriminatory modes of care that are present in non-automated medical practices, as well as intensify surveillance practices that already harm the most vulnerable. To borrow the words of historian of science Hannah Zeavin, “We cannot hope to enfold those who care forgot while repressing this central fact: care is a tool, but it is also, too often, a weapon” (Zeavin 2021).

Conclusion

Affective computing promises to provide objective and meaningful insights into the inner life of its users. Originally intended to support the affective experiences of autistic people, applications of affective computing today include mood tracking, depression and anxiety detection, as well as diagnosis and treatment of broader neurodevelopmental conditions. To produce meaningful knowledge about these various conditions, practitioners rely on a range of theories drawn from Western psychology that posit a correlation between “abnormal” facial movements and particular mental states (such as depressive episodes).

These efforts echo a broader trend, called “digital phenotyping”, that aims to democratize healthcare provision by monitoring and predicting health conditions through digital technologies. Practitioners mobilize a variety of therapeutic promises and societal benefits (such as objective diagnoses and increased accessibility to treatments) to legitimize the development and deployment of these tools. This chapter has brought to the fore the social and cultural representations of the affects, the human, and the mind that underly these efforts. Despite the narratives that frame these technologies as fundamentally benign, tools that discriminate between health and

illness can replicate existing exclusionary practices that already harm the most vulnerable by, for example, reinforcing normative assumptions around illness and disability that can prevent people from receiving adequate care. This chapter also drawn attention to the market logics that drive these projects and see complex social and health issues as viable business-cases. According to this vision, technical interventions are seen as adequate solutions to wider structural problems in the distribution and access to care. I have joined here feminist and STS scholars in arguing that, embedded in these projects, is a notion of *cure* rather *care*. Here, cure is often a point of departure, rather than of arrival. Cure is the array of algorithmic techniques available to practitioners, and the set of theories that fit these tools: solutions in search for a problem.

Conclusion

In this concluding chapter, I bring together my findings and discuss their relevance to scholarship, policy, and practice. I begin by reviewing the development of my argument throughout the thesis and then reflect on the limitations of this work and suggest avenues for future research. Finally, I return to the issues of agnosticism and uncertainty with which this work started. I compare participants' profession of agnosticism to *agnogenesis*, that is, the construction of ignorance within technoscientific contexts to avoid accountability (Proctor and Schiebinger 2008). I then argue for the recognition of uncertainty as constitutive to the project of automation and the need to create space within design pipelines to collectively contest and challenge algorithmic certainty and objectivity.

Seeing affect within knowledge infrastructures

In this thesis, I have investigated the infrastructures of knowledge on which practitioners draw to classify and make sense of affective behaviors. In particular, I have shown how, despite their professed agnosticism with regards to conceptual models of emotions and expressive behaviors, practitioners' ways of 'seeing' affect are shaped by a complex of norms, practices, and normative assumptions about affect, aspects of identity (e.g. gender), the mind, health, and illness. Informed by feminist studies of science and technology, and knitting together historical inquiry with qualitative investigations of participants' local and contingent practices of sense-making, I have sought to uncover the cultural and historical imaginaries that underpin academic and industry FER projects.

In chapter 3, *Performing theories*, I traced the popularization of the “classic view” of emotions – the idea that expressions of emotions are universal, biologically determined, and involuntary – through the work of psychologist Paul Ekman who, between the 1960s and 2000s, set out to establish a scientific and objective framework for the analysis of facial behaviors. Supported and often in service of US intelligence and law enforcement agencies, his lifelong investigation of nonverbal clues of deception has directly shaped early developments of automated FER systems. Central to his efforts was the creation of the Facial Action Coding System (FACS), a scientific method for the classification and analysis of affective behaviors that turned facial expressions into calculable and manageable raw data, regardless of the social and cultural contexts that elicited those expressions. Ekman’s work has been object of much criticism, as a result of which many practitioners have progressively abandoned inferential claims about facial expressions and emotions. They have not, however, abandoned Ekman’s approach to expressive behaviors as something that can be transformed into measurable and objective properties and used to produce some form of knowledge about individuals. This fundamental assumption has directly shaped the research and development of automated FER systems. Despite efforts to re-frame the narrative, this genealogy shows that it is not fully possible to divorce the classification of facial expressions from its origin as a tool to spot deception from nonverbal behaviors.

In chapter 4, *The making of affect*, I situated Ekman’s legacy within the longer history of quantifications of affect and, following historian Otniel Dror, I located the origin of scientific efforts to measure expressive behaviors in 19th century Western physiology, when the creation of new scientific instruments and protocols made it possible to elevate affective behaviors to objects of scientific knowledge and use this knowledge to legitimize projects of social classification. As shown in Paolo Mantegazza’s and Duchenne de Boulogne’s photographic atlases of emotional expressions, these efforts to quantify and display emotions helped reinforce and

normalize assumptions towards affective behaviors aligned with social and cultural norms that posited emotional expressions of white and wealthy Europeans as the apex of the civilization process. However, to capture and display affective behaviors in a scientific and objective manner, physiologists needed to establish standardization practices that could tame the subjective and fleeting nature of emotions: practices such as consistent head orientation, background colors, and the use of captions echo the modern process of data wrangling necessary to collect, clean, and label ground truth data on which FER algorithms are trained. Finally, these scientific representations of emotions served a prescriptive role: by displaying what emotional expressions should look like, these representations facilitated the construction of cultural norms of emotional expression and a hierarchy of feeling that privileged certain expressions and behaviors over others. These two chapters formed the backdrop for the next three empirical chapters (chapters 5 to 7), where I brought these historical dimensions in conversation with modern FER design and development.

In the first empirical chapter, *Mobilizing science*, I drew attention to participants' construction of objectivity and validity through means of discursive and rhetorical practices. As mentioned in chapter 3, despite (and perhaps because of) the public and scholarly controversy over efforts to classify expressive behaviors, many practitioners prefer to maintain an agnostic view with regards to the conceptual underpinnings of FER systems. However, to classify affective expressions, and infer meaningful information from this classification, practitioners must make some fundamental assumption about what constitutes a measurable affective expression and how to extract knowledge from it. In light of this apparent paradox, the chapter foregrounds the forms of boundary work through which participants establish and maintain epistemic authority over their claims. Emerging from these responses are practices and tactics that allow participants to validate their system's predictions without needing to commit to any particular conceptual model of affect: since

algorithmic performance cannot be evaluated in relation to the inner ‘truth’ of emotional experience, participants described how they focus instead on the measurement and prediction of changes in expressive behavior, as long as these are *consistent* with mainstream emotion models.

The following chapter, *Seeing affect*, focused on the role that different data structures and practices play in constructing claims of objectivity and accuracy. Here, I looked at the experience of two annotators performing FACS and VAD coding, regarded respectively as forms of “expert” and “lay” annotation. However, by focusing on the norms, grids, and local negotiation practices of these coding schemes, I showed how both practices require that annotators learn how to ‘see’ affective data, bringing to the fore the subjective and interpretive element that underpin FER projects. Additional factors, such as annotation interfaces and automation tools that aim to tame the subjective element of the labelling process, shape annotators’ ways of seeing and interpreting the data. I then argued that the definition and attribution of skill and expertise in data structures reflects broader power asymmetries in machine learning design, maintained and reinforced through practitioners’ perception and representation of data work. In particular, by reinforcing perceptions of data work and annotation as either objective and scientific or subjective and intuitive, practitioners overlook (and at times obscure) the human work of negotiation and interpretation that underpin both coding practices and, ultimately, the uncertainty that lies at the core of ground truth datasets.

In the final empirical chapter, *The promise of Affective Computing*, I explored the social and cultural imaginaries that inform the development of healthcare applications of FER systems and, particularly, those designed for mental health and neurodevelopmental conditions. Here, I showed how knowledge produced through participants’ techno-scientific discourse and practice can shape and determine bodies and behaviors. In line with what previously observed about the prescriptive dimensions

of historical projects of affect classification, here normative ideas of health and illness, as well as gendered assumptions about expressive affective behaviors that are at play in the development of FER systems can in fact facilitate the expression of certain behaviors aligned with socio-cultural expectations around, for instance, gender and health. Finally, encoded in the promise of affective computing is also a notion of care that, at the same time, treats a) illness and disability as viable business cases, and b) technical interventions as solutions to wider socio-economic inequities in healthcare access and treatment.

Contributions to knowledge and policy

In this thesis, I have offered a feminist STS account of affective computing practice as a way of ‘seeing’ data that is shaped by a multifaceted infrastructure of theories, norms, routines, and normative assumptions about affect, the mind, health, and illness. This work contributes to at least two broad research areas: first, this analysis adds to the interdisciplinary scholarship at the intersection of Critical Data Studies, Fairness, Accountability, and Transparency (FAccT) studies, and Computer Supported Cooperative Work (CSCW) concerned with algorithmic accountability from a qualitative research perspective. In particular, the historical and ethnographic investigation of the social and cultural imaginaries that underpin the development of FER systems allow us to contest and resist claims of objectivity and universality and open up spaces for accountability. Following feminist studies of science and technology, this research locates the practices of sense-making involved in the quantification of affect both locally and historically, as a way to ground technoscientific knowledge about affective behaviors and make it accountable. Second, the investigation of the interpretive and subjective work required for the creation of emotion atlases as well as facial expressions training datasets (as illustrated in chapters

3, 4, and 6) makes a contribution to the nascent field of Critical Dataset Studies which, drawing on archival studies, aims to account for the ethics and politics of datasets by engaging critically with “the weight of history within data set regimes” (Thylstrup 2022: 66). Further, this research foregrounds the practices of data annotation, interpretation, and standardization within FER design as “acts of governance” (665) that define what does and does not constitute as valid affective behavior, thus articulating “‘the law of what can be said’ through distributed moments of interpretation and validation” (Ibid.). However, the analysis emphasizes the role that practitioners’ representations and expectations of data work play on the interpretation of data, adding to CSCW and HCI literature concerned with how power dynamics shape data structures (Miceli, Schuessler, and Yang 2020; Miceli et al. 2021; Miceli, Posada, and Yang 2022; Irani 2015b; Gray and Suri 2019; Raval 2021).

In this thesis, I sought to make a contribution to future policy interventions that aim to curb the potential harms of FER systems and related affective computing technologies. In particular, by tracing the various genealogies of healthcare applications, I move beyond narrow framings of FER systems as emotion recognition technology and instead illustrate the normative and potentially harmful effects that these systems can have even when disconnected from claims of emotional inference. In October 2022, the Information Commissioner’s Office (ICO) – a British independent body advocating for data rights and privacy regulations – issued a warning to organizations, citing the current “immature” state of emotion recognition technology as a risk of bias and discrimination.⁹¹ In particular, Deputy Commissioner Stephen Bonner mentioned the need for technology to be rigorously backed by science before being commercially deployed:

⁹¹<https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/10/immature-biometric-technologies-could-be-discriminating-against-people-says-ico-in-warning-to-organisations/>.

Developments in the biometrics and emotion AI market are immature. They may not work yet, or indeed ever. While there are opportunities present, the risks are currently greater. At the ICO, we are concerned that incorrect analysis of data could result in assumptions and judgements about a person that are inaccurate and lead to discrimination. The only sustainable biometric deployments will be those that are fully functional, accountable and backed by science. As it stands, we are yet to see any emotion AI technology develop in a way that satisfies data protection requirements, and have more general questions about proportionality, fairness and transparency in this area.⁹²

However, as illustrated throughout this research, the boundaries between science and pseudo-science (or science and non-science) can be shifted through discursive and material tactics that validate particular knowledge claims over others such as, for instance, FER systems' supposed ability to provide objective assessments of mental health and neurodevelopmental conditions that seemingly do not involve inferences about emotional states. Further, practitioners' efforts to re-frame their professional and disciplinary identity as disconnected from emotion recognition practice (and sometimes from affective computing altogether) constituted a key tactic for the validation of their claims. Given the efforts of most regulatory proposals to govern emotional inference projects,⁹³ participants seemed to have – at least rhetorically – moved away from discussions around 'emotions' and 'affect' to focus instead on the less controversial measurement of facial and expressive behaviors. Hopefully, extending the remit of investigation of FER systems outside of those contexts and actors explicitly

⁹² Ibid.

⁹³ See for example the 2021 European Commission's Proposal for the Artificial Intelligence Act ('EU AI Act'), available at https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF. The proposal identifies the risks of, on one hand, *emotion recognition systems*, defined as "AI systems for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their biometric data", and, on the other hand, *biometric categorisation systems*, defined as "AI systems for the purpose of assigning natural persons to specific categories, such as sex, age, hair colour, eye colour, tattoos, ethnic origin or sexual or political orientation, on the basis of their biometric data" (p. 42). As illustrated in this thesis, however, many of the projects discussed by the participants evade these definitions in that they do not (explicitly or implicitly) seek to infer emotions nor intentions from facial behaviors.

centred around emotional inference can contribute to policy interventions aiming to reduce harm and discrimination of machine learning systems.

Limitations, future research, and “good” technology

As mentioned in chapter 2, this thesis – as any piece of research – is constrained by some methodological limitations and, particularly, by issues related to sampling and scope. First, choices regarding methods and sampling were greatly influenced by the effect of the Covid-19 pandemic on face-to-face ethnographic research. As mentioned in the methods section, the impossibility to attend in-person social events for most of the data collection period heavily impacted my ability to build trust and long-term relationships with participants. This meant that I could not access contexts of FER research and development as originally envisaged, and that conversations about participants’ projects (particularly commercial ones) were often constrained by concerns about trade-secrets and intellectual property. To circumvent these issues, as explained in chapter 2, I extended my sample to include academic participants and collect a wider range of experiences. While this added depth and coherence to the findings – as both groups offered similar accounts of their practices and assumptions – a potential avenue for future research could tease out more nuanced distinctions between, as well as within, these two sectors. For example, overlooked here are the market dynamics that regulate industry R&D with regards to the different size and scope of participants’ organizations – e.g. how do classification practices differ between large corporations and start-up companies? For instance, there is an expectation within industry contexts that start-ups will be acquired by or integrated in larger organizations (Prado and Bauer 2022). This might shape in important ways participants’ knowledge claims and development practices. Another limitation with regards to sampling is the

relative demographic homogeneity of the participants. As previously discussed, the sample skews white and male, thus reflecting the overall gender and race disparities in tech-related fields within Western countries. Future research could bring these dimensions more to the fore and aim for a more representative sample. A different but somehow related point of consideration is that absent from the analysis is a discussion of whether and how gender and race are annotated within facial expressions datasets (and, particularly, commercial ones). While many have noted the importance of questioning how these dimensions are represented in and codified into machine learning systems (Scheuerman et al. 2020; Scheuerman, Paul, and Brubaker 2019), the proprietary nature of participants' datasets meant that I was often given little detail about the characteristics of the training data. Further research could take these concerns as a point of departure and add to the literature on bias and fairness in HCI, AI, and affective computing.

A second limitation concerns the temporal scope of the research. As mentioned in chapter 2, the data collection process took place between the fall of 2020 and the summer of 2022. While long-term ethnographic studies of technology can provide richer and more nuanced insights into design and development practices, my research was limited by practical constraints (namely, the length of the scholarship that supported this work). Sampsa Hyysalo et al. have warned against the “framing effect” of STS investigations focused on too narrow time frames and how these can lead to simplistic and unrefined understandings of technology development (Hyysalo, Pollock, and Williams 2019). To address this concern, I have however taken caution to situate my empirical research within the longer history of the quantification of affective behaviors, thus extending the investigation longitudinally and increasing the relevance of the findings. Missing from this work are, however, what the authors call research “interstices”, as the “moments and sites in which the various focal actors in the ecology interlink and affect each other and the evolving technology” (7). Future research could

include multi-sited ethnographic investigations (Marcus 1995) to capture points of contact between different relevant actors, including users and regulators, where technology (and its related imaginaries) is co-constructed.

Finally, a note on the research overall orientation: while I have given some space to participants' positive framings of FER systems, in this thesis I have mostly focused on the potential for harm and discrimination produced or reinforced through this technology. Missing from my account is a reflection on whether "good" FER systems are possible – given that we could answer the question "good for whom?" – and, if so, what they would look like. For instance, HCI scholar Kirsten Boehner and her colleagues have argued for a conceptual as well as methodological shift from the dominant view of emotions as biological, universal, and measurable objects to, instead, an interactional model that sees affect as "culturally grounded, dynamically experienced, and to some degree constructed in action and interaction" (Boehner et al. 2007: 276). An interactional approach to affect, the authors argue, "moves the focus from helping computers to better understand human emotion to helping people to understand and experience their own emotions" (277). In this sense, the subjective and interpretive element of emotional expression recognition is not denied or obfuscated but, rather, brought to the fore:

Emotion is a witnessable property of social action, a way in which actions are rendered interpretable and meaningful. The question of the dynamic, situated interpretation (and attribution) of emotional behavior is critical here. Negotiation, interpretation and inference are inextricably intertwined (280).

The authors offer a number of examples of design strategies grounded on the interactional model of emotions. Among these is the "Affecter" (Sengers et al. 2008), a system designed to collect real-time video data of two separate office workers and relay data visualizations of the workers' affective states on a "video window" between

the two adjoining offices. The data is presented in the form of “distortions” – image alterations such as color inversion or pixelization – to enable users’ co-interpretation as well as manipulation of their own emotional output (Boehner et al. 2007). As the authors explain, “instead of the system interpreting the emotional meaning of the input, the users interpret the emotional meaning of the output, and tune its output to support their readings” (282). As an interactional model of emotions, Affecter is designed primarily to facilitate users’ awareness and self-reflection about their-own and others’ affective states.

While offering ways to foreground the interpretive and contextual element of expressive behaviors, this approach is still based on a view that sees technology as an adequate and meaningful tool for the regulation of affect. In commenting on Boehner and colleagues’ interactional approach, scholar Kelly Gates has questioned this very premise: “rather than taking as given the value of emotional self-examination and self-knowledge, it is important to consider why individuals are so persistently encouraged or enjoined to engage in practices of self-assessment” (Gates 2011: 187). In this sense, and consistently with some of the dimensions of FER systems explored in chapter 7, she argues that affective computing applications resonate “with a certain regime of the self that is characteristic of late capitalist societies, one that encourages individuals to continuously examine and reinvent themselves in order to acclimate to their changing social, economic, and technological environment” (Ibid.). In this sense, imagining “good” affective technologies – although the question of “good for whom?” persists – might mean asking whether it is possible to imagine technologies that can evade the ubiquitous datafication and monetization of human experience typical of neoliberal technological regimes. Further, a similar technology should perhaps encode alternative forms of care that differ radically from those described in chapter 7 (a normative notion of care-as-cure seeking to ‘fix’ abnormal bodies), and promote instead care practices that are “ambivalent, contextual, and relational” (Martin, Myers, and Viseu 2015: 631),

that is, practices that pay attention to the local and situated conditions in which care is required and performed. Envisioning these alternative forms of affective technology, perhaps, does not so much require re-imagining their application contexts or knowledge claims (in the form of “more accurate” or “more valid” claims) but, rather, reconfigure the ways in which infrastructures of knowledge, vision, and power are enacted within data structures. The next section offers a final reflection on these dimensions in the context of feminist practice.

“I prefer not to”: From ignorance to uncertainty

In his short story *Bartleby, the Scrivener: A Story of Wall Street*, writer Herman Melville narrates the life of Bartleby, a clerk whose task is to make copies of legal documents for an elderly lawyer on Wall Street. One day, asked to proofread a piece of paper, Bartleby answers with what is perhaps one of the most renowned utterances in American literature: “I prefer not to”. From this point onward, Bartleby won’t perform any task, repeating the same formula over and over, until the police remove him from the office. Thrown in jail, Bartleby finally dies of self-starvation. Many have since speculated on the conceptual and ethical implications of this story and, particularly, on the political connotations of Bartleby’s enunciation. As Gilles Deleuze argued, the scrivener is not simply *refusing* to obey the lawyer’s requests:

If Bartleby had refused, he could still be seen as a rebel or insurrectionary, and as such would still have a social role. But the formula stymies all speech acts, and at the same time, it makes Bartleby a pure outsider to whom no social position can be attributed. This is what the attorney glimpses with dread: all his hopes of bringing Bartleby back to reason are dashed because they rest on a logic of presuppositions according to which an employer “expects” to be obeyed, or a kind friend listened to, whereas Bartleby has invented a new logic, a logic of preference, which is enough to undermine the presuppositions of language as a whole (Deleuze 1997: 4).

I do not intend to dwell excessively on Deleuze's interpretation, nor do I aim to provide an alternative explanation of Melville's short novel. I would, instead, like to take inspiration from *Bartleby* to re-think the problem of agnosticism with which this research started. Indeed, in this work, I have taken the problem of agnosticism in affective computing as an entry point to investigate the tensions associated with the classification of affective behaviors and, more broadly, the potential as well as real harms of machine learning as a discipline and practice that treats all forms of knowledge as amenable to its techniques and objectives. In chapter 5, I have described this phenomenon as the *magpie effect*, that is, the tendency in computer science practice to borrow theoretical frameworks from other disciplines without questioning the epistemological assumptions on which they are grounded. This seeming lack of reflexivity could also be seen as what Robert N. Proctor and Londa Schiebinger called *agnogenesis*: the construction of ignorance within techno-scientific domains to avoid or evade accountability (Proctor and Schiebinger 2008). A notorious example of this is the deliberate fabrication of doubt by the tobacco industry about the dangers of smoking (Ibid.). Similarly, climate change deniers have often argued for the need of more scientific evidence with the aim to stall political action (Oreskes and Conway 2008). In a similar sense, organizations construct ignorance to control "uncomfortable knowledge", that is, information that can "undermine the organizational principles of a society or organization" (Rayner 2012). Organizations, argued Steve Rayner, can construct ignorance through four different implicit strategies: *denial* (refusal to act on known information), *dismissal* (rejection of uncomfortable knowledge as wrong or irrelevant), *diversion* (the establishment of a "decoy activity" (118) to divert attention from known problems), and *displacement* (the creation of a "surrogate" object that, although originally designed to provide and manage information about a particular phenomenon, becomes the object of controversial knowledge) (Ibid.). Examples of

displacement include computational models designed to generate data about pollution that are not accurate representations of local ecosystems but are nonetheless adopted “to simulate the results of policy interventions” (121).

These strategies can perhaps describe some of the rhetorical devices reported here by participants: while practitioners’ agnosticism can be in itself a way to avoid accountability by means of dismissal, the emphasis on seemingly uncontroversial measurements of affective behaviors can be instead seen as a form of diversion or displacement. As described in this thesis, efforts to decouple FACS (or other approaches to facial coding) from higher levels of interpretations of affective behaviors aim to shift the focus away from inferential claims and draw attention instead to the perceived objectivity of facial muscle measurements. However, as argued throughout this work, participants’ agnosticism with regards to the conceptual underpinnings of FER models can obscure, if not obstruct, accountability, in that it removes from public sight the historical and epistemological assumptions on which these models rest.

How can one then challenge agnosticism and uphold accountability in computer science? In line with the theoretical framework that informs this research, I seek to contest agnosticism in a way that doesn’t require to assert a single “truth” but rather acknowledges the situated perspectives that inform computational efforts to quantify and predict human experience. As argued in chapter 6, this could mean foregrounding the situated and partial knowledges that contribute to the creation of ground truths in machine learning: a recognition of uncertainty as constitutive to the project of automation, rather than as a statistical outlier or error. Acknowledging the uncertainty of sense-making practices within machine learning becomes then a political effort in that it can re-assemble the configurations of knowledge, vision, and power explored in this thesis. Borrowing once again from Louise Amoore, rather than seeking grounds for automated decisions, algorithmic accountability – what she termed a “cloud ethics” – “begins from the ungroundedness of all forms of decision, all political claims, human

and algorithmic” (Amoore 2020: 148). In this sense, I seek inspiration from Bartleby’s enunciation – “I prefer not to” – as a way to re-claim uncertainty (rather than agnosticism) as a form of doubtfulness and opening for accountability. As per Deleuze’s interpretation of Melville’s story, uncertainty does not entail a simple *refusal* to act (or, in this case, to design); while some critical HCI scholars have argued for the need to “not design” (Baumer and Silberman 2011) or “undesign” (Pierce 2012) inadequate or inappropriate technological interventions, not all those involved in the design process have the power to disengage without facing some kind of consequences. Conversely, Os Keyes et al. argued for a shift towards an “anarchist HCI”, a practice “dedicated to building *prefigurative counterpower*: creating constantly-justified spaces that embody autonomous, anti-oppressive values as a means to build the revolution in increment” (Keyes, Hoy, and Drouhard 2019: 4). As they described, an anarchist HCI should not be understood as a mere theoretical framework, but rather an encouragement to interrogate and make accountable power relations and systems of oppression at play in local contexts. In this perspective, practitioners’ work should not be guided by “funders’ implicit or explicit expectations” (8), but rather by the needs and desires of those impacted by design work: “work cannot be undertaken without the active consent and participation of these communities” (Ibid.). However, since ethical research in AI and data-driven technologies is now often funded by large tech corporations, Corinne Cath and Os Keyes have more recently advocated for a practice of “feminist refusal” (Cath and Keyes 2022). Building on Bonnie Honig’s *Feminist Theory of Refusal* (Honig 2021), the authors proposed a form of engagement with funders that recognizes “that we are never outside relations of power. As such, we have responsibilities of care to those nestled more deeply within these power structures, to treat refusal as a tactic” (Cath and Keyes 2022: 34). To *refuse* is, in this sense, to remain aware of the limits and opportunities of working within power structures and, when

possible, use those structures to the advantage of the communities impacted by research work.

Perhaps unsurprisingly, Honnig's theory of refusal takes Bartleby as a point of departure, whose utterance "I prefer not to" she described as "a celebrated example of inoperativity" – the refusal to do anything – and a response to "the concern that giving reasons and making demands just enters would-be dissidents into the trade-offs of their opponents and compromises projects of radical transformation" (Honnig 2021: 14). The inoperativity of Bartleby – a concept first articulated by Italian philosopher Giorgio Agamben (Agamben and Heller-Roazen 1999) – rejects instead all "instrumental and teleological approaches to ethics or politics", treating refusal as an end in itself. However, Honnig noted, Bartleby's ultimate death demonstrates that his solitary refusal is insufficient to motivate a political project or contribute to larger political movements. In contrast, other literary characters can offer inspiration for a collective politics of refusal and, particularly, the Greek tragedy of *Bacchae*. In Euripides's play, the women of Thebes respond to the chant of "*Eta Bakkae, Eta Bakkae*" ("Gather round, gather round!"), a call to leave the city and join together in dance to celebrate Dionysus, the Greek god of wine and ecstasy, banished from worshipping by King Pentheus. Here, the women "eat and drink to their hearts' content, unleash desire, generate collective power, and host new practices of mutuality that express care but also unleash violence" (Honnig 2021: xii). In the play, the bacchantes' women's refusal to follow the King's orders is described as a sign of their insanity and the god's manipulation. However, argued Honnig, their refusal should be taken seriously: "Fearlessly, they defy orders and abandon the instruments of work, the loom and shuttle. Locked up, they break out and flee imprisonment to escape the city" (22). Building on Judith Butler's work on the performative power of democratic assemblies (Butler 2015), Honnig described the women's coming together in chants and dances as a form of joyous *praxis*, akin to the

collective chants and dances one can observe in public gatherings: “The chant binds people together and collectivizes their resolve” (Honig 2021: 19).

These reflections add further dimension to the problem of agnosticism: if we ought to resist and contest abstract and ahistorical claims of algorithmic objectivity, and foreground the doubtfulness that lies at the core of computational projects, we must invoke a form of uncertainty that does not rely exclusively on the individual – as some forms of AI ethics would require – but rather on a collective reconfiguration of power dynamics. As suggested in chapter 6, the maintenance of uneven power dynamics within data structures contributes to the inability to treat uncertainty as constitutive of data work and challenge the interpretive grids and norms shaping participants’ sense-making practices. As Cath and Keyes have observed, one must not forget that feminist practices:

are often relational, involving assemblages of people, and that (partly as a result of this) they are *practices*. Answers are contingent on circumstance, and the pursuit of them alters those circumstances and in turn demands a re-evaluation of our ongoing actions (Cath and Keyes 2022: 34).

Within industry contexts, however, this might require a collective reconfiguration of the labor structures on which machine learning projects rest. Although peripheral to the responses analyzed in this thesis, collective action is indeed a key element of political projects of algorithmic justice (Hicks 2018; Stapleton et al. 2018). In this sense, the ability to contest and challenge the certainty of algorithmic predictions should be made integral part of FER and broader machine learning projects, in the form of a collective “I prefer not to”.

Final remarks

As many of the authors cited in this thesis have argued, algorithms' technical opacity and commercial secrecy should not be an obstacle to critical inquiry, nor should the professed detachment of computer science from questions of accountability and social responsibility. Throughout this work, I have sought to investigate FER systems analytically and offer strategies for grounding the historical, social, and cultural norms and assumptions that inform their development. Further, I hope that the concerns and the research methods discussed here can be extended to algorithmic technologies more broadly. As discussed in this thesis' introductory chapter, the constant effort to make ambiguous and elusive objects of experience – in this case, affect – manageable in practice is not unique to affective computing. It is, rather, at the heart of many computational projects. The strategies provided here can hopefully help illuminate this fundamental ambiguity beyond a single piece of technology or computing field. In addition, this work has shown how practitioners circumvent this uncertainty by putting into circulation practical engineering solutions that bypass scientific controversy and that, in the long term, might shape and change human experience.

Bibliography

- Abdar, Moloud, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, et al. 2021. 'A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges'. *Information Fusion* 76 (December): 243–97. <https://doi.org/10.1016/j.inffus.2021.05.008>.
- Abdurahman, J. Khadijah. 2022. 'Birthing Predictions of Premature Death'. *Logic Magazine*, 22 August 2022. <https://logicmag.io/home/birthing-predictions-of-premature-death/>.
- Agamben, Giorgio, and Daniel Heller-Roazen. 1999. *Potentialities: Collected Essays in Philosophy*. Stanford University Press.
- Ahmed, Sara. 2010. *The Promise of Happiness*. The Promise of Happiness. Duke University Press. <https://doi.org/10.1515/9780822392781>.
- Aleksic, Petar S., Gerasimos Potamianos, and Aggelos K. Katsaggelos. 2005. 'Exploiting Visual Information in Automatic Speech Processing'. In *Handbook of Image and Video Processing (Second Edition)*, edited by AL Bovik, 1263–XXXIX. Communications, Networking and Multimedia. Burlington: Academic Press. <https://doi.org/10.1016/B978-012119792-6/50134-0>.
- Alvesson, Mats, and Kaj Sköldbberg. 2017. *Reflexive Methodology: New Vistas for Qualitative Research*. SAGE.
- Amoore, Louise. 2020. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Duke University Press.
- Anderson, Chris. 2008. 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete'. *Wired*, 2008. <https://www.wired.com/2008/06/pb-theory/>.
- Angerer, Marie-Luise, and Bernd Bösel. 2016. 'Total Affect Control. Or: Who's Afraid of a Pleasing Little Sister?' *Digital Culture & Society* 2 (1): 41–52. <https://doi.org/10.25969/mediarep/824>.
- Angerer, Marie-Luise, Bernd Bösel, and Michaela Ott, eds. 2014. *Timing of Affect: Epistemologies of Affection*. DIAPHANES. <https://press.uchicago.edu/ucp/books/book/distributed/T/bo19163301.html>.
- Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. 'Machine Bias'. *ProPublica*, 23 May 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Araujo, Theo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. 2020. 'In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence'. *AI & SOCIETY* 35 (3): 611–23. <https://doi.org/10.1007/s00146-019-00931-w>.

- Arcas, Blaise Aguera y. 2017. 'Physiognomy's New Clothes'. Medium. 20 May 2017. <https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a>.
- Article 19. 2021. 'Emotional Entanglement: China's Emotion Recognition Market and Its Implications for Human Rights'. <https://www.article19.org/emotion-recognition-technology-report/>.
- Baron-Cohen, Simon. 2003. *The Essential Difference: The Truth about the Male and Female Brain*. The Essential Difference: The Truth about the Male and Female Brain. New York, NY, US: Basic Books.
- Barrett, Lisa Feldman. 2017. *How Emotions Are Made: The Secret Life of the Brain*. Pan Macmillan.
- Barrett, Lisa Feldman, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. 'Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements'. *Psychological Science in the Public Interest* 20 (1): 1–68. <https://doi.org/10.1177/1529100619832930>.
- Bartlett, Marian Stewart, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. 2006. 'Automatic Recognition of Facial Actions in Spontaneous Expressions'. *Journal of Multimedia* 1 (September). <https://doi.org/10.4304/jmm.1.6.22-35>.
- Baumeister, Harald, and Christian Montag. 2023. 'Digital Phenotyping and Mobile Sensing in Psychoinformatics—A Rapidly Evolving Interdisciplinary Research Endeavor'. In *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*, edited by Christian Montag and Harald Baumeister, 1–9. Studies in Neuroscience, Psychology and Behavioral Economics. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-98546-2_1.
- Baumer, Eric P.S., and M. Six Silberman. 2011. 'When the Implication Is Not to Design (Technology)'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2271–74. CHI '11. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/1978942.1979275>.
- Beer, David. 2016. *Metric Power*. London: Palgrave Macmillan UK. <https://doi.org/10.1057/978-1-137-55649-3>.
- Bellacasa, María Puig de la. 2017. *Matters of Care: Speculative Ethics in More than Human Worlds*. U of Minnesota Press.
- Bengio, Yoshua. 2020. 'Time to Rethink the Publication Process in Machine Learning'. *Yoshua Bengio* (blog). 27 February 2020. <https://yoshuabengio.org/2020/02/26/time-to-rethink-the-publication-process-in-machine-learning/>.
- Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Wiley.

- Bennabi, Djamila, Pierre Vandel, Charalambos Papaxanthis, Thierry Pozzo, and Emmanuel Haffen. 2013. 'Psychomotor Retardation in Depression: A Systematic Review of Diagnostic, Pathophysiologic, and Therapeutic Implications'. *BioMed Research International* 2013 (October): e158746. <https://doi.org/10.1155/2013/158746>.
- Berger, John. 1990. *Ways of Seeing*. British Broadcasting Corporation and Penguin Books.
- Bietti, Elettra. 2020. 'From Ethics Washing to Ethics Bashing', 10.
- Birhane, Abeba. 2021a. 'Cheap AI'. In *Fake AI*, edited by Frederike Kaltheuner. Meatspace Press.
- . 2021b. 'Algorithmic Injustice: A Relational Ethics Approach'. *Patterns* 2 (2): 100205. <https://doi.org/10.1016/j.patter.2021.100205>.
- Birhane, Abeba, and Olivia Guest. 2020. 'Towards Decolonising Computational Sciences'. arXiv. <https://doi.org/10.48550/arXiv.2009.14258>.
- Birk, Rasmus, Anna Lavis, Federica Lucivero, and Gabrielle Samuel. 2021. 'For What It's Worth. Unearthing the Values Embedded in Digital Phenotyping for Mental Health'. *Big Data & Society* 8 (2): 20539517211047320. <https://doi.org/10.1177/20539517211047319>.
- Bloor, David. 1976. *Knowledge and Social Imagery*. University of Chicago Press.
- Boehner, Kirsten, Rogério DePaula, Paul Dourish, and Phoebe Sengers. 2007. 'How Emotion Is Made and Measured'. *International Journal of Human-Computer Studies*, Evaluating affective interactions, 65 (4): 275–91. <https://doi.org/10.1016/j.ijhcs.2006.11.016>.
- Bogner, A., B. Littig, and W. Menz. 2009. *Interviewing Experts*. Springer.
- Bonilla-Silva, Eduardo. 2019. 'Feeling Race: Theorizing the Racial Economy of Emotions'. *American Sociological Association*, February. <https://www.asanet.org/2018-presidential-address-feeling-race-theorizing-racial-economy-emotions>.
- Bösel, Bernd, and Serjoscha Wiemer, eds. 2020. *Affective Transformations: Politics – Algorithms – Media*. meson press. <https://doi.org/10.14619/1655>.
- Bowker, Geoffrey C., and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. MIT Press.
- boyd, danah, and Kate Crawford. 2012. 'CRITICAL QUESTIONS FOR BIG DATA: Provocations for a Cultural, Technological, and Scholarly Phenomenon'. *Information, Communication & Society* 15 (5): 662–79. <https://doi.org/10.1080/1369118X.2012.678878>.
- Browne, Simone. 2015. *Dark Matters: On the Surveillance of Blackness*. Duke University Press.
- Bryman, Alan. 2016. *Social Research Methods*. Oxford University Press.

- Buolamwini, Joy, and Timnit Gebru. 2018. 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification'. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. PMLR. <https://proceedings.mlr.press/v81/buolamwini18a.html>.
- Burrell, Jenna. 2016. 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms'. *Big Data & Society* 3 (1): 2053951715622512. <https://doi.org/10.1177/2053951715622512>.
- Butler, Judith. 1999. *Gender Trouble*. 2nd ed. New York: Routledge. <https://doi.org/10.4324/9780203902752>.
- . 2011. *Bodies That Matter: On the Discursive Limits of Sex*. London: Routledge. <https://doi.org/10.4324/9780203828274>.
- Buyukdura, Jeylan S., Shawn M. McClintock, and Paul E. Croarkin. 2011. 'Psychomotor Retardation in Depression: Biological Underpinnings, Measurement, and Treatment'. *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 35 (2): 395–409. <https://doi.org/10.1016/j.pnpbp.2010.10.019>.
- Callon, Michel. 1984. 'Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay'. *The Sociological Review* 32 (1_suppl): 196–233. <https://doi.org/10.1111/j.1467-954X.1984.tb00113.x>.
- Calvo, Rafael A., and Sidney D'Mello. 2010. 'Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications'. *IEEE Transactions on Affective Computing* 1 (1): 18–37. <https://doi.org/10.1109/T-AFFC.2010.1>.
- Calvo, Rafael A., Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas. 2015. *The Oxford Handbook of Affective Computing*. Oxford University Press.
- Campolo, Alexander, and Kate Crawford. 2020. 'Enchanted Determinism: Power without Responsibility in Artificial Intelligence'. *Engaging Science, Technology, and Society* 6 (January): 1. <https://doi.org/10.17351/ests2020.277>.
- Campos, Belinda, Dacher Keltner, and Molly Parker Tapias. 2004. 'Emotion'. In *Encyclopedia of Applied Psychology*, edited by Charles D. Spielberger, 713–22. New York: Elsevier. <https://doi.org/10.1016/B0-12-657410-3/00042-8>.
- Cath, Corinne, and Os Keyes. 2022. 'Your Thoughts for a Penny? Capital, Complicity and AI Ethics'. In *Economies of Virtue: The Circulation of 'Ethics' in AI*, edited by Thao Phan, Jake Goldenfein, Declan Kuch, and Monique Mann. Vol. 46. Theory on Demand. Amsterdam: Institute of Network Cultures.
- Cattell, Heather E.P., and Alan D. Mead. 2008. 'The Sixteen Personality Factor Questionnaire (16PF)'. In *The SAGE Handbook of Personality Theory and Assessment: Volume 2 — Personality Measurement and Testing*, 135–59. London: SAGE Publications Ltd. <https://doi.org/10.4135/9781849200479>.

- Cave, Stephen, Claire Craig, Kanta Dihal, Sarah Dillon, Jessica Montgomery, Beth Singler, and Lindsay Taylor. 2018. 'Portrayals and Perceptions of AI and Why They Matter'.
- Chiarelli, Cosimo. 2020. 'L'Atlante Del Dolore : Fotografia Ed Espressione Delle Emozioni in Paolo Mantegazza'. *L'Atlante Del Dolore : Fotografia Ed Espressione Delle Emozioni in Paolo Mantegazza*, 13–36.
<https://doi.org/10.19272/202012401002>.
- Clynes, Manfred, and Nathan S. Kline. 1960. 'Cyborgs and Space'. *Astronautics* 14 (9): 26–27.
- Cohn, Jeffrey F. 2013. 'Beyond Group Differences: Specificity of Nonverbal Behavior and Interpersonal Communication to Depression Severity'. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 1–2. AVEC '13. New York, NY, USA: Association for Computing Machinery.
<https://doi.org/10.1145/2512530.2512537>.
- Collingridge, D G. 1979. 'The Entrenchment of Technology: The Case of Lead Petrol Additives'. *Science and Public Policy* 6 (5): 332–38.
<https://doi.org/10.1093/spp/6.5.332>.
- Cooper, Harriet. 2020. 'The Fantasy of Maternal Autonomy and the Disabled Mother'. *Studies in the Maternal* 13 (1). <https://doi.org/10.16995/sim.296>.
- Costanza-Chock, Sasha. 2018. 'Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice'. In .
<https://doi.org/10.21606/drs.2018.679>.
- Crawford, Kate. 2021. *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
<https://yalebooks.yale.edu/9780300264630/atlas-of-ai>.
- Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, et al. 2019. 'AI Now Report 2019'. New York: AI Now Institute. https://ainowinstitute.org/AI_Now_2019_Report.html.
- Crawford, Kate, and Trevor Paglen. 2021. 'Excavating AI: The Politics of Images in Machine Learning Training Sets'. *AI & SOCIETY*, June.
<https://doi.org/10.1007/s00146-021-01162-8>.
- Crivelli, Carlos, James A. Russell, Sergio Jarillo, and José-Miguel Fernández-Dols. 2016. 'The Fear Gasping Face as a Threat Display in a Melanesian Society'. *Proceedings of the National Academy of Sciences*, October.
<https://doi.org/10.1073/pnas.1611622113>.
- Darwin, Charles. 1998. *The Expression of the Emotions in Man and Animals*. Oxford University Press.
- Daston, Lorraine, and Peter Galison. 2007. *Objectivity*. Princeton University Press.
- Davies, William. 2015. *The Happiness Industry: How the Government and Big Business Sold Us Well-Being*. Verso Books.

- Deleuze, Gilles. 1997. 'Bartleby; or, the Formula'. *Essays Critical and Clinical* 86.
- Denton, Emily, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. 'On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet'. *Big Data & Society* 8 (2): 20539517211035956.
<https://doi.org/10.1177/20539517211035955>.
- Derkatch, Colleen. 2008. 'Method as Argument: Boundary Work in Evidence - Based Medicine'. *Social Epistemology* 22 (4): 371–88.
<https://doi.org/10.1080/02691720802559412>.
- D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. MIT Press.
- Dingwall, Robert. 1997. 'Accounts, Interviews and Observations'. In *Context and Method in Qualitative Research*, by Gale Miller and Robert Dingwall, 52–65. 6 Bonhill Street, London England EC2A 4PU United Kingdom: SAGE Publications Ltd. <https://doi.org/10.4135/9781849208758.n4>.
- Döringer, Stefanie. 2021. "'The Problem-Centred Expert Interview". Combining Qualitative Interviewing Approaches for Investigating Implicit Expert Knowledge'. *International Journal of Social Research Methodology* 24 (3): 265–78. <https://doi.org/10.1080/13645579.2020.1766777>.
- Dourish, Paul. 2016. 'Algorithms and Their Others: Algorithmic Culture in Context'. *Big Data & Society* 3 (2): 2053951716665128.
<https://doi.org/10.1177/2053951716665128>.
- Downey, Gary Lee, and Juan C. Lucena. 2001. 'Engineering Studies'. In *Handbook of Science and Technology Studies*, edited by Sheila Jasanoff, Gerald E. Markle, James C. Peterson, and Trevor J. Pinch. Thousand Oaks, UNITED STATES: SAGE Publications, Incorporated.
<http://ebookcentral.proquest.com/lib/ed/detail.action?docID=996269>.
- . 2004. 'Knowledge and Professional Identity in Engineering: Code - switching and the Metrics of Progress'. *History and Technology* 20 (4): 393–420.
<https://doi.org/10.1080/0734151042000304358>.
- Dror, Otniel E. 1999a. 'The Affect of Experiment: The Turn to Emotions in Anglo-American Physiology, 1900-1940'. *Isis* 90 (2): 205–37.
- . 1999b. 'The Scientific Image of Emotion: Experience and Technologies of Inscription'. *Configurations* 7 (3): 355–401.
- . 2001. 'Counting the Affects: Discoursing in Numbers'. *Social Research* 68 (2): 357–78.
- . 2011. 'Seeing the Blush: Feeling Emotions'. In *Histories of Scientific Observation*, edited by Lorraine Daston and Elizabeth Lunbeck. University of Chicago Press.
<https://press.uchicago.edu/ucp/books/book/chicago/H/bo10303424.html>.

- Duchenne, Guillaume-Benjamin. 1876. *Mécanisme de la physionomie humaine ou analyse électro-physiologique de l'expression des passions*. Librairie J.-B. Baillière et Fils.
- Dwoskin, Elizabeth, and Evelyn M. Rusli. 2015. 'The Technology That Unmasks Your Hidden Emotions'. *The Wall Street Journal*, 28 January 2015, sec. Tech. <http://online.wsj.com/articles/startups-see-your-face-unmask-your-emotions-1422472398>.
- Edwards, Paul. 2002. 'Infrastructure and Modernity: Scales of Force, Time, and Social Organization in the History of Sociotechnical Systems'. In *Modernity and Technology*, 185–225.
- Edwards, Paul N., Steven J. Jackson, Melissa K. Chalmers, Geoffrey C. Bowker, Christine L. Borgman, David Ribes, Matt Burton, and Scout Calvert. 2013. 'Knowledge Infrastructures: Intellectual Frameworks and Research Challenges'.
- Ekman, Paul. 1970. 'Universal Facial Expressions of Emotion'. *California Mental Health Research Digest* 8: 151–58.
- . 1971. 'Universals and Cultural Differences in Facial Expressions of Emotion'. *Nebraska Symposium on Motivation* 19: 207–83.
- . 2016. *Nonverbal Messages: Cracking the Code. My Life Pursuit*. Paul Ekman Group.
- Ekman, Paul, and Wallace V. Friesen. 1969. 'Nonverbal Leakage and Clues to Deception'. *Psychiatry* 32 (1): 88–106. <https://doi.org/10.1080/00332747.1969.11023575>.
- Ekman, Paul, Wallace V. Friesen, and Thomas G. Taussig. 1969. 'VID-R and SCAN: Tools and Methods for the Automated Analysis of Visual Records'. Clearinghouse for Federal Scientific and Technical Information, Springfield, Virginia 22151 (AD-702 635, MF \$).
- Ekman, Paul, Wallace V. Friesen, and Silvan S. Tomkins. 1971. 'Facial Affect Scoring Technique: A First Validity Study'. *Semiotica* 3 (1). <https://doi.org/10.1515/semi.1971.3.1.37>.
- Ekman, Paul, and Erika L. Rosenberg, eds. 2005. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. 2nd ed. Series in Affective Science. New York: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195179644.001.0001>.
- Ekman, Paul, Terrence Sejnowski, and Joseph Hager. 1993. 'Final Report To NSF of the Planning Workshop on Facial Expression Understanding', January.
- Elish, M. C., and danah boyd. 2018. 'Situating Methods in the Magic of Big Data and AI'. *Communication Monographs* 85 (1): 57–80. <https://doi.org/10.1080/03637751.2017.1375130>.

- Epstein, Steven. 1995. 'The Construction of Lay Expertise: AIDS Activism and the Forging of Credibility in the Reform of Clinical Trials'. *Science, Technology, & Human Values* 20 (4): 408–37.
<https://doi.org/10.1177/016224399502000402>.
- Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group.
- Fazi, M. Beatrice. 2018. *Contingent Computation: Abstraction, Experience, and Indeterminacy in Computational Aesthetics*. Rowman & Littlefield.
- Feldman, Jessica. 2016. "'The Problem of the Adjective"1: Affective Computing of the Speaking Voice'. *Transposition*, no. 6 (December).
<https://doi.org/10.4000/transposition.1640>.
- Fine, Cordelia. 2005. *Delusions of Gender: The Real Science Behind Sex Differences*. Icon Books Ltd.
- Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations'. *Minds and Machines* 28 (4): 689–707.
<https://doi.org/10.1007/s11023-018-9482-5>.
- Forsey, Martin, Georg Breidenstein, Oliver Krüger, and Anna Roch. 2015. 'Ethnography at a Distance: Globally Mobile Parents Choosing International Schools'. *International Journal of Qualitative Studies in Education* 28 (9): 1112–28. <https://doi.org/10.1080/09518398.2015.1074754>.
- Foucault, M., Arnold I. Davidson, and Graham Burchell. 2008. *The Birth of Biopolitics: Lectures at the Collège de France, 1978-1979*. Springer.
- Foucault, Michel. 2011. *The Courage of Truth*. Springer.
- . 2013. *History of Madness*. Routledge.
<https://doi.org/10.4324/9780203642603>.
- Frank, Adam J., and Elizabeth A. Wilson. 2020. *A Silvan Tomkins Handbook: Foundations for Affect Theory*. U of Minnesota Press.
- Freeman, Karoline, Julia Geppert, Chris Stinton, Daniel Todkill, Samantha Johnson, Aileen Clarke, and Sian Taylor-Phillips. 2021. 'Use of Artificial Intelligence for Image Analysis in Breast Cancer Screening Programmes: Systematic Review of Test Accuracy'. *BMJ* 374 (September): n1872.
<https://doi.org/10.1136/bmj.n1872>.
- Gates, Kelly A. 2011. *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*. Our Biometric Future. New York University Press.
<https://doi.org/10.18574/nyu/9780814732090.001.0001>.
- Gavrilescu, Mihai, and Nicolae Vizireanu. 2017. 'Predicting the Sixteen Personality Factors (16PF) of an Individual by Analyzing Facial Features'. *EURASIP*

- Journal on Image and Video Processing* 2017 (1): 59.
<https://doi.org/10.1186/s13640-017-0211-4>.
- . 2018. 'Predicting the Big Five Personality Traits from Handwriting'.
EURASIP Journal on Image and Video Processing 2018 (1): 57.
<https://doi.org/10.1186/s13640-018-0297-3>.
- Ghassemi, Marzyeh, Luke Oakden-Rayner, and Andrew L. Beam. 2021. 'The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care'. *The Lancet Digital Health* 3 (11): e745–50.
[https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9).
- Gieryn, Thomas F. 1983. 'Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists'. *American Sociological Review* 48 (6): 781–95.
<https://doi.org/10.2307/2095325>.
- . 1995. 'Boundaries of Science'. In *Handbook of Science and Technology Studies, Revised Edition*, edited by Sheila Jasanoff, Gerald E. Markle, James C. Petersen, and Trevor J. Pinch, 393–443. Thousand Oaks: SAGE Publications, Inc. <https://doi.org/10.4135/9781412990127>.
- . 1999. *Cultural Boundaries of Science: Credibility on the Line*. Chicago, IL: University of Chicago Press.
<https://press.uchicago.edu/ucp/books/book/chicago/C/bo3642202.html>.
- Gitelman, Lisa. 2013. *Raw Data Is an Oxymoron*. MIT Press.
- Gladwell, Malcolm. 2002. 'The Naked Face'. *The New Yorker*, 28 July 2002.
<https://www.newyorker.com/magazine/2002/08/05/the-naked-face>.
- Goodwin, Charles. 1994. 'Professional Vision'. *American Anthropologist* 96 (3): 606–33.
- Gould, Stephen Jay. 1996. *The Mismeasure of Man*. Norton.
- Grasseni, Cristina. 2004. 'Skilled Vision. An Apprenticeship in Breeding Aesthetics'. *Social Anthropology* 12 (1): 41–55. <https://doi.org/10.1111/j.1469-8676.2004.tb00089.x>.
- . 2007. *Skilled Visions: Between Apprenticeship and Standards*. Berghahn Books.
- Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt.
- Green, Ben. 2021. 'The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice'. *Journal of Social Computing* 2 (3): 209–25.
<https://doi.org/10.23919/JSC.2021.0018>.
- Grill, Gabriel. 2022. 'Constructing Certainty in Machine Learning: On the Performativity of Testing and Its Hold on the Future'. OSF Preprints.
<https://doi.org/10.31219/osf.io/zekqv>.

- Hagerty, Alexa, and Livia Garofalo. 2021. 'Mapping Algorithmic Assumptions: Reflections from a Society for Psychological Anthropology Roundtable'. *Somatosphere* (blog). 13 September 2021. <http://somatosphere.net/2021/mapping-algorithmic-assumptions.html/>.
- Haraway, Donna. 1988. 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective'. *Feminist Studies* 14 (3): 575–99. <https://doi.org/10.2307/3178066>.
- . 1990. *Simians, Cyborgs, and Women: The Reinvention of Nature*. New York: Routledge. <https://doi.org/10.4324/9780203873106>.
- Harrington, Anne. 2019. *Mind Fixers: Psychiatry's Troubled Search for the Biology of Mental Illness*. W. W. Norton & Company.
- He, Lang, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, et al. 2022. 'Deep Learning for Depression Recognition with Audiovisual Cues: A Review'. *Information Fusion* 80 (April): 56–86. <https://doi.org/10.1016/j.inffus.2021.10.012>.
- Healey, J., and R.W. Picard. 1998. 'StartleCam: A Cybernetic Wearable Camera'. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215)*, 42–49. <https://doi.org/10.1109/ISWC.1998.729528>.
- Heimerl, Alexander, Tobias Baur, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. 2019. 'NOVA - A Tool for EXplainable Cooperative Machine Learning'. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 109–15. <https://doi.org/10.1109/ACII.2019.8925519>.
- Hicks, Mar. 2018. 'The Long History behind the Google Walkout'. *The Verge*, 9 November 2018. <https://www.theverge.com/2018/11/9/18078664/google-walkout-history-tech-strikes-labor-organizing>.
- Hill, Kashmir. 2020. 'The Secretive Company That Might End Privacy as We Know It'. *The New York Times*, 18 January 2020, sec. Technology. <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>.
- Hoffmann, Anna Lauren, and Luke Stark. 2015. 'Hard Feelings — Inside Out, Silicon Valley, and Why Technologizing Emotion and Memory Is a Dangerous Idea'. *Los Angeles Review of Books*, 11 September 2015. <https://lareviewofbooks.org/article/hard-feelings-inside-out-silicon-valley-and-why-technologizing-emotion-and-memory-is-a-dangerous-idea/>.
- Honig, Bonnie. 2021. *A Feminist Theory of Refusal. A Feminist Theory of Refusal*. Harvard University Press. <https://doi.org/10.4159/9780674259249>.
- Hu, Lily. 2021. 'Tech Ethics: Speaking Ethics to Power, or Power Speaking Ethics?' *Journal of Social Computing* 2 (3): 238–48. <https://doi.org/10.23919/JSC.2021.0033>.

- Hutchinson, Ben, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. 'Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure'. *ArXiv:2010.13561 [Cs]*, January. <http://arxiv.org/abs/2010.13561>.
- Hutson, Matthew. 2018. 'Artificial Intelligence Could Identify Gang Crimes—and Ignite an Ethical Firestorm'. *Science*, 28 February 2018. <https://www.science.org/content/article/artificial-intelligence-could-identify-gang-crimes-and-ignite-ethical-firestorm>.
- Hyysalo, Sampsa, Neil Pollock, and Robin Alun Williams. 2019. 'Method Matters in the Social Study of Technology: Investigating the Biographies of Artifacts and Practices'. *Science & Technology Studies* 32 (3): 2–25. <https://doi.org/10.23987/sts.65532>.
- Iliadis, Andrew, and Federica Russo. 2016. 'Critical Data Studies: An Introduction'. *Big Data & Society* 3 (2): 2053951716674238. <https://doi.org/10.1177/2053951716674238>.
- Insel, Thomas R. 2017. 'Digital Phenotyping: Technology for a New Science of Behavior'. *JAMA* 318 (13): 1215–16. <https://doi.org/10.1001/jama.2017.11295>.
- Irani, Lilly. 2015a. 'Justice for "Data Janitors"'. *Public Books*, 15 January 2015. <https://www.publicbooks.org/justice-for-data-janitors/>.
- . 2015b. 'The Cultural Work of Microwork'. *New Media & Society* 17 (5): 720–39. <https://doi.org/10.1177/1461444813511926>.
- Irani, Lilly, and M. Six Silberman. 2013. 'Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 611–20. CHI '13. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/2470654.2470742>.
- Jacobs, Abigail Z., and Hanna Wallach. 2021. 'Measurement and Fairness'. <https://doi.org/10.1145/3442188.3445901>.
- Jain, Sachin H., Brian W. Powers, Jared B. Hawkins, and John S. Brownstein. 2015. 'The Digital Phenotype'. *Nature Biotechnology* 33 (5): 462–63. <https://doi.org/10.1038/nbt.3223>.
- Jasanoff, Sheila S. 1987. 'Contested Boundaries in Policy-Relevant Science'. *Social Studies of Science* 17 (2): 195–230. <https://doi.org/10.1177/030631287017002001>.
- Joly, Pierre-Benoit. 2015. 'Governing Emerging Technologies – The Need to Think Outside the (Black) Box'. In *Science and Democracy: Knowledge as Wealth and Power in the Biosciences and Beyond*, edited by Stephen Hilgartner, Clark Miller, and Rob Hagendijk. Routledge.

- Kaliouby, Rana Ayman El. 2005. 'Mind-Reading Machines: Automated Inference of Complex Mental States'.
- Kaliouby, Rana el, Rosalind Picard, and Simon Baron-Cohen. 2006. 'Affective Computing and Autism'. *Annals of the New York Academy of Sciences* 1093 (December): 228–48. <https://doi.org/10.1196/annals.1382.016>.
- Kandel, Sean, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. 'Enterprise Data Analysis and Visualization: An Interview Study'. *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2917–26. <https://doi.org/10.1109/TVCG.2012.219>.
- Kapoor, Sayash, and Arvind Narayanan. 2022. 'Leakage and the Reproducibility Crisis in ML-Based Science'. arXiv. <http://arxiv.org/abs/2207.07048>.
- Karasti, Helena, Florence Millerand, Christine M. Hine, and Geoffrey C. Bowker. 2016. 'Knowledge Infrastructures: Part I'. *Science & Technology Studies* 29 (1): 2–12. <https://doi.org/10.23987/sts.55406>.
- Kelleher, John D. 2019. *Deep Learning*. MIT Press.
- Keltner, Dacher, and Paul Ekman. 2015. 'The Science of "Inside Out"'. *New York Times*, 3 July 2015.
- Keyes, Os. 2018. 'The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition'. *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW): 1–22. <https://doi.org/10.1145/3274357>.
- Keyes, Os, Josephine Hoy, and Margaret Drouhard. 2019. 'Human-Computer Insurrection: Notes on an Anarchist HCI'. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. CHI '19. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3290605.3300569>.
- Khatchadourian, Raffi. 2015. 'We Know How You Feel: Computers Are Learning to Read Emotion, and the Business World Can't Wait'. *The New Yorker*, 2015, 19 edition.
- Kitchin, Rob. 2014. 'Big Data, New Epistemologies and Paradigm Shifts'. *Big Data & Society* 1 (1): 205395171452848. <https://doi.org/10.1177/2053951714528481>.
- Latour, Bruno. 2012. 'Visualisation and Cognition: Drawing Things Together'. *Avant: Trends in Interdisciplinary Studies* 3 (T): 207–60.
- Law, John. 2004. *After Method: Mess in Social Science Research*. London: Routledge. <https://doi.org/10.4324/9780203481141>.
- Law, John, and Michel Callon. 1988. 'Engineering and Sociology in a Military Aircraft Project: A Network Analysis of Technological Change*'. *Social Problems* 35 (3): 284–97. <https://doi.org/10.2307/800623>.
- Lehoux, Pascale, Geneviève Daudelin, Myriam Hivon, Fiona Alice Miller, and Jean-Louis Denis. 2014. 'How Do Values Shape Technology Design? An

- Exploration of What Makes the Pursuit of Health and Wealth Legitimate in Academic Spin-Offs'. *Sociology of Health & Illness* 36 (5): 738–55.
<https://doi.org/10.1111/1467-9566.12097>.
- Leys, Ruth. 2011. 'The Turn to Affect: A Critique'. *Critical Inquiry* 37 (3): 434–72.
<https://doi.org/10.1086/659353>.
- Lipton, Zachary C. 2017. 'The Mythos of Model Interpretability'. arXiv.
<https://doi.org/10.48550/arXiv.1606.03490>.
- Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. 'Algorithm Appreciation: People Prefer Algorithmic to Human Judgment'. *Organizational Behavior and Human Decision Processes* 151 (March): 90–103.
<https://doi.org/10.1016/j.obhdp.2018.12.005>.
- Lunden, Ingrid. 2014. 'Emotient Raises \$6M For Facial Expression Recognition Tech, Debuts Google Glass Sentiment Analysis App'. *TechCrunch*, 6 March 2014.
<https://techcrunch.com/2014/03/06/emotient-raises-6m-for-its-facial-expression-recognition-tech-debuts-sentiment-analysis-app-for-google-glass/>.
- Lupton, Deborah. 2016. *The Quantified Self*. John Wiley & Sons.
- MacKenzie, Donald. 1993. *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*. MIT Press.
- MacKenzie, Donald, and Yuval Millo. 2003. 'Constructing a Market, Performing Theory: The Historical Sociology of a Financial Derivatives Exchange'. *American Journal of Sociology* 109 (1): 107–45.
<https://doi.org/10.1086/374404>.
- Madianou, Mirca. 2021. 'Nonhuman Humanitarianism: When "AI for Good" Can Be Harmful'. *Information, Communication & Society* 24 (6): 850–68.
<https://doi.org/10.1080/1369118X.2021.1909100>.
- Malazita, James W., and Korryn Resetar. 2019. 'Infrastructures of Abstraction: How Computer Science Education Produces Anti-Political Subjects'. *Digital Creativity* 30 (4): 300–312. <https://doi.org/10.1080/14626268.2019.1682616>.
- Marcus, George E. 1995. 'Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography'. *Annual Review of Anthropology* 24: 95–117.
- Markowetz, Alexander, Konrad Błazskiewicz, Christian Montag, Christina Switala, and Thomas E. Schlaepfer. 2014. 'Psycho-Informatics: Big Data Shaping Modern Psychometrics'. *Medical Hypotheses* 82 (4): 405–11.
<https://doi.org/10.1016/j.mehy.2013.11.030>.
- Martin, Aryn, Natasha Myers, and Ana Viseu. 2015. 'The Politics of Care in Technoscience'. *Social Studies of Science* 45 (5): 625–41.
<https://doi.org/10.1177/0306312715602073>.
- Martin Moruno, Dolores. 2016. 'Pain as Practice in Paolo Mantegazza's Science of Emotions'. *Osiris* 31: 137–62.

- Mauro, Hayes Peter. 2000. 'Duchenne: Discourses of Aesthetics, Sexuality, and Power in Nineteenth-Century Medical Photography'. *Athanos* 18: 55–61.
- McStay, Andrew. 2018. *Emotional AI: The Rise of Empathic Media*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd.
<https://doi.org/10.4135/9781526451293>.
- McStay, Andrew, and Lachlan Urquhart. 2019. "This Time with Feeling?": Assessing EU Data Governance Implications of out of Home Appraisal Based Emotional AI'. *First Monday* 24 (10). <https://doi.org/10.5210/fm.v24i10.9457>.
- Meyer, Martin. 2003. 'Academic Entrepreneurs or Entrepreneurial Academics? Research-Based Ventures and Public Support Mechanisms'. *R&D Management* 33 (2): 107–15. <https://doi.org/10.1111/1467-9310.00286>.
- Miceli, Milagros, and Julian Posada. 2022. 'The Data-Production Dispositif'. arXiv. <https://doi.org/10.48550/arXiv.2205.11963>.
- Miceli, Milagros, Julian Posada, and Tianling Yang. 2021. 'Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?' <https://www.semanticscholar.org/paper/434d211269bf82cd743b4f3dbbd3c377d763aa27>.
- Miceli, Milagros, M. Schuessler, and Tianling Yang. 2020. 'Between Subjectivity and Imposition'. <https://doi.org/10.1145/3415186>.
- Miceli, Milagros, Tianling Yang, Adriana Alvarado Garcia, Julian Posada, Sonja Mei Wang, Marc Pohl, and Alex Hanna. 2022. 'Documenting Data Production Processes: A Participatory Approach for Data Work'. arXiv. <https://doi.org/10.48550/arXiv.2207.04958>.
- Miceli, Milagros, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. 'Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices'. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 161–72. FAccT '21. New York, NY, USA: Association for Computing Machinery.
<https://doi.org/10.1145/3442188.3445880>.
- Mills, Mara, and Meredith Whittaker. 2019. 'Disability, Bias, and AI'.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. 'Model Cards for Model Reporting'. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, January, 220–29.
<https://doi.org/10.1145/3287560.3287596>.
- Mitruț, Oana, Gabriela Moise, Livia Petrescu, Alin Moldoveanu, Marius Leordeanu, and Florica Moldoveanu. 2019. 'Emotion Classification Based on Biophysical Signals and Machine Learning Techniques'. *Symmetry* 12 (December): 21.
<https://doi.org/10.3390/sym12010021>.

- Narayanan, Arvind, and Sayash Kapoor. 2022. 'Why Are Deep Learning Technologists so Overconfident?' Substack newsletter. *AI Snake Oil* (blog). 31 August 2022. <https://aisnakeoil.substack.com/p/why-are-deep-learning-technologists>.
- Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. *Algorithms of Oppression*. New York University Press. <https://doi.org/10.18574/nyu/9781479833641.001.0001>.
- Nyeko, Kizito Emmanuel, and Ngui Sing. 2015. 'Academic Entrepreneurs and Entrepreneurial Academics: Are They the Same'. *International Journal of Social Science and Humanity* 5 (January): 1050–55. <https://doi.org/10.7763/IJSSH.2015.V5.603>.
- Oakden-Rayner, Luke, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2019. 'Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging'. arXiv. <https://doi.org/10.48550/arXiv.1909.12475>.
- O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Books.
- Ong, Desmond C., Jamil Zaki, and Noah D. Goodman. 2019. 'Computational Models of Emotion Inference in Theory of Mind: A Review and Roadmap'. *Topics in Cognitive Science* 11 (2): 338–57. <https://doi.org/10.1111/tops.12371>.
- Oreskes, Naomi, and Erik M. Conway. 2008. 'Challenging Knowledge: How Climate Science Became a Victim of the Cold War'. In *Agnotology: The Making and Unmaking of Ignorance*, edited by Robert N. Proctor and Londa Schiebinger. Stanford, California: Stanford University Press.
- Pandzic, Igor S., and Robert Forchheimer. 2002. 'The Origins of the MPEG-4 Facial Animation Standard'. In *MPEG-4 Facial Animation*, 3–13. John Wiley & Sons, Ltd. <https://doi.org/10.1002/0470854626.ch1>.
- Pearl, Sharrona. 2010. *About Faces: Physiognomy in Nineteenth-Century Britain*. Harvard University Press. https://scholar.google.com/citations?view_op=view_citation&hl=it&user=7mCt8B0AAAAJ&citation_for_view=7mCt8B0AAAAJ:RGFaLdJalmkC.
- Perrotta, Manuela, and Alina Geampana. 2020. 'The Trouble with IVF and Randomised Control Trials: Professional Legitimation Narratives on Time-Lapse Imaging and Evidence-Informed Care'. *Social Science & Medicine* 258 (August): 113115. <https://doi.org/10.1016/j.socscimed.2020.113115>.
- Pfister, Damien Smith. 2017. 'The Terms of Technoliberalism'. In *Theorizing Digital Rhetoric*. Routledge.
- Picard, Rosalind W. 1995. 'Affective Computing'. M.I.T Media Laboratory Perceptual Computing Section Technical Report 321.
- . 2000. *Affective Computing*. MIT Press.

- . 2010. 'Affective Computing: From Laughter to IEEE'. *IEEE Transactions on Affective Computing* 1 (1): 11–17. <https://doi.org/10.1109/T-AFFC.2010.10>.
- . 2015. 'The Promise of Affective Computing'. In *The Oxford Handbook of Affective Computing*, edited by Rafael Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas, 0. Oxford University Press. <https://doi.org/10.1093/oxfordhob/9780199942237.013.013>.
- Pickersgill, Martyn. 2010. 'From Psyche to Soma? Changing Accounts of Antisocial Personality Disorders in the American Journal of Psychiatry'. *History of Psychiatry* 21 (3): 294–311. <https://doi.org/10.1177/0957154X09102800>.
- . 2019. 'Digitising Psychiatry? Sociotechnical Expectations, Performative Nominalism and Biomedical Virtue in (Digital) Psychiatric Praxis'. *Sociology of Health & Illness* 41 (S1): 16–30. <https://doi.org/10.1111/1467-9566.12811>.
- Pierce, James. 2012. 'Undesigning Technology: Considering the Negation of Design by Design'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 957–66. CHI '12. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/2207676.2208540>.
- Porter, Theodore M. 1996. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press. <https://doi.org/10.1515/9781400821617>.
- Porter, Theodore M. 2009. 'How Science Became Technical'. *Isis* 100 (2): 292–309. <https://doi.org/10.1086/599552>.
- Prado, Tiago S., and Johannes M. Bauer. 2022. 'Big Tech Platform Acquisitions of Start-Ups and Venture Capital Funding for Innovation'. *Information Economics and Policy* 59 (June): 100973. <https://doi.org/10.1016/j.infoecopol.2022.100973>.
- Preciado, Paul B. 2018. 'BAROQUE TECHNOPATRIARCHY: REPRODUCTION'. *Artforum*, January 2018. <https://www.artforum.com/print/201801/baroque-technopatriarchy-reproduction-73189>.
- Proctor, Robert N., and Londa Schiebinger. 2008. *Agnotology: The Making and Unmaking of Ignorance*. Stanford University Press Stanford, California. <https://philarchive.org/rec/PROATM>.
- Raji, Inioluwa Deborah, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2022. 'AI and the Everything in the Whole Wide World Benchmark'. In . <https://openreview.net/forum?id=j6NxpQbREA1>.
- Raji, Inioluwa Deborah, and Genevieve Fried. 2021. 'About Face: A Survey of Facial Recognition Evaluation'. *ArXiv:2102.00813 [Cs]*, February. <http://arxiv.org/abs/2102.00813>.
- Raji, Inioluwa Deborah, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. 'Saving Face: Investigating the Ethical Concerns

- of Facial Recognition Auditing'. *ArXiv:2001.00964 [Cs]*, January.
<http://arxiv.org/abs/2001.00964>.
- Raval, Noopur. 2021. 'Interrupting Invisibility in a Global World'. *Interactions* 28 (4): 27–31. <https://doi.org/10.1145/3469257>.
- Rayner, Steve. 2012. 'Uncomfortable Knowledge: The Social Construction of Ignorance in Science and Environmental Policy Discourses'. *Economy and Society* 41 (1): 107–25. <https://doi.org/10.1080/03085147.2011.637335>.
- Rhue, Lauren. 2018. 'Racial Influence on Automated Perceptions of Emotions'. SSRN Scholarly Paper. Rochester, NY. <https://doi.org/10.2139/ssrn.3281765>.
- Ribes, David, Andrew S. Hoffman, Steven C. Slota, and Geoffrey C. Bowker. 2019. 'The Logic of Domains': *Social Studies of Science*, May. <https://doi.org/10.1177/0306312719849709>.
- Riva, Giuseppe, Rafael Calvo, and Christine Lisetti. 2015. 'Cyberpsychology and Affective Computing'. In *The Oxford Handbook of Affective Computing*, edited by Rafael Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas, 0. Oxford University Press.
<https://doi.org/10.1093/oxfordhb/9780199942237.013.017>.
- Robinson, Dawn, and Lynn Smith-Lovin. 2006. 'Affect Control Theory'. In , 137–64. https://doi.org/10.1007/978-0-387-30715-2_9.
- Rose, Nikolas. 1988. 'Calculable Minds and Manageable Individuals': *History of the Human Sciences*. <https://doi.org/10.1177/095269518800100202>.
- Rosenberg, Erika L., and Paul Ekman. 2020. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press.
- Ross, Joel, Andrew Zaldivar, Lilly Irani, and Bill Tomlinson. 2009. 'Who Are the Turkers? Worker Demographics in Amazon Mechanical Turk'. *Social Code Report 2009-01*, January.
https://soundideas.pugetsound.edu/faculty_pubs/1013.
- Rouast, Philipp V., Marc T. P. Adam, and Raymond Chiong. 2021. 'Deep Learning for Human Affect Recognition: Insights and New Developments'. *IEEE Transactions on Affective Computing* 12 (2): 524–43.
<https://doi.org/10.1109/TAFFC.2018.2890471>.
- Russell, J. A. 1994. 'Is There Universal Recognition of Emotion from Facial Expression? A Review of the Cross-Cultural Studies'. *Psychological Bulletin* 115 (1): 102–41. <https://doi.org/10.1037/0033-2909.115.1.102>.
- Russell, James A. 1980. 'A Circumplex Model of Affect'. *Journal of Personality and Social Psychology* 39: 1161–78. <https://doi.org/10.1037/h0077714>.
- Russell, James A, and Albert Mehrabian. 1977. 'Evidence for a Three-Factor Theory of Emotions'. *Journal of Research in Personality* 11 (3): 273–94.
[https://doi.org/10.1016/0092-6566\(77\)90037-X](https://doi.org/10.1016/0092-6566(77)90037-X).

- Sambasivan, Nithya, and Rajesh Veeraraghavan. 2022. 'The Deskilling of Domain Expertise in AI Development'. In *CHI Conference on Human Factors in Computing Systems*, 1–14. CHI '22. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3491102.3517578>.
- Scheuerman, M., Jacob M. Paul, and J. Brubaker. 2019. 'How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services'. <https://www.semanticscholar.org/paper/8140f2a44d3bb698c6008def848c4d1f7f80b1de>.
- Scheuerman, M., Kandrea Wade, Caitlin Lustig, and J. Brubaker. 2020. 'How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis'. <https://doi.org/10.1145/3392866>.
- Schiebinger, Londa. 1990. 'The Anatomy of Difference: Race and Sex in Eighteenth-Century Science'. *Eighteenth-Century Studies* 23 (4): 387–405. <https://doi.org/10.2307/2739176>.
- Schwartz, G. E., P. L. Fair, P. Salt, M. R. Mandel, and G. L. Klerman. 1976. 'Facial Expression and Imagery in Depression: An Electromyographic Study'. *Psychosomatic Medicine* 38 (5): 337–47. <https://doi.org/10.1097/00006842-197609000-00006>.
- Seaver, Nick. 2018. 'What Should an Anthropology of Algorithms Do?' *Cultural Anthropology* 33 (3): 375–85. <https://doi.org/10.14506/ca33.3.04>.
- Sedgwick, Eve Kosofsky. 2003. *Touching Feeling: Affect, Pedagogy, Performativity*. *Touching Feeling*. Duke University Press. <https://doi.org/10.1515/9780822384786>.
- Sengers, Phoebe, Kirsten Boehner, Michael Mateas, and Geri Gay. 2008. 'The Disenchantment of Affect'. *Personal and Ubiquitous Computing* 12 (5): 347–58. <https://doi.org/10.1007/s00779-007-0161-4>.
- Seo, Sungyong, Hau Chan, P. Jeffrey Brantingham, Jorja Leap, Phebe Vayanos, Milind Tambe, and Yan Liu. 2018. 'Partially Generative Neural Networks for Gang Crime Classification with Partial Information'. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 257–63. AIES '18. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3278721.3278758>.
- Setyadi, Ardintya Diva, Tri Harsono, and Sigit Wasista. 2015. 'Human Character Recognition Application Based on Facial Feature Using Face Detection'. In *2015 International Electronics Symposium (IES)*, 263–67. <https://doi.org/10.1109/ELECSYM.2015.7380852>.
- Shapin, Steven. 1979. 'The Politics of Observation: Cerebral Anatomy and Social Interests in the Edinburgh Phrenology Disputes'. *The Sociological Review* 27 (S1): 139–78. <https://doi.org/10.1111/j.1467-954X.1979.tb00061.x>.

- Sloane, Mona, Emanuel Moss, and Rumman Chowdhury. 2022. 'A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo-Science, and the Quest for Auditability'. *Patterns* 3 (2): 100425. <https://doi.org/10.1016/j.patter.2021.100425>.
- Slota, Stephen C, Andrew S Hoffman, David Ribes, and Geoffrey C Bowker. 2020. 'Prospecting (in) the Data Sciences'. *Big Data & Society* 7 (1): 2053951720906849. <https://doi.org/10.1177/2053951720906849>.
- Song, Miri. 2020. 'Rethinking Minority Status and "Visibility"'. *Comparative Migration Studies* 8 (1): 5. <https://doi.org/10.1186/s40878-019-0162-2>.
- Spiel, Katta, Christopher Frauenberger, Os Keyes, and Geraldine Fitzpatrick. 2019. 'Agency of Autistic Children in Technology Research—A Critical Literature Review'. *ACM Transactions on Computer-Human Interaction* 26 (November): 1–40. <https://doi.org/10.1145/3344919>.
- Stapleton, Claire, Tania Gupta, Meredith Whittaker, Celie O'Neil-Hart, Stephanie Parker, Erica Anderson, and Amr Gaber. 2018. 'We're the Organizers of the Google Walkout. Here Are Our Demands'. *The Cut*, 1 November 2018. <https://www.thecut.com/2018/11/google-walkout-organizers-explain-demands.html>.
- Star, Susan. 2002. 'Infrastructure and Ethnographic Practice: Working on the Fringes'. *Scandinavian Journal of Information Systems* 14 (2). <https://aisel.aisnet.org/sjis/vol14/iss2/6>.
- Star, Susan Leigh. 1999. 'The Ethnography of Infrastructure'. *American Behavioral Scientist* 43 (3): 377–91. <https://doi.org/10.1177/00027649921955326>.
- Stark, Luke. 2020a. 'Empires of Feeling: Social Media and Emotive Politics'. In *Affective Politics of Digital Media*. Routledge.
- . 2020b. 'The Emotive Politics of Digital Mood Tracking'. *New Media & Society* 22 (11): 2039–57. <https://doi.org/10.1177/1461444820924624>.
- Stark, Luke, and Jesse Hoey. 2020. 'The Ethics of Emotion in AI Systems'. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. OSF Preprints. <https://doi.org/10.31219/osf.io/9ad4u>.
- Stark, Luke, and Jevan Hutson. 2021. 'Physiognomic Artificial Intelligence'. SSRN Scholarly Paper. Rochester, NY. <https://doi.org/10.2139/ssrn.3927300>.
- Stearns, Peter N., and Carol Z. Stearns. 1985. 'Emotionology: Clarifying the History of Emotions and Emotional Standards'. *The American Historical Review* 90 (4): 813–36. <https://doi.org/10.2307/1858841>.
- Stewart, James, and Robin Williams. 2005. 'The Wrong Trousers? Beyond the Design Fallacy: Social Learning and the User'. In *Handbook of Critical Information Systems Research*, edited by Debra Howcroft and Eileen M. Trauth. Edward Elgar Publishing. <https://www.elgaronline.com/display/9781843764786.00017.xml>.

- Stop LAPD Spying Coalition. 2018. 'Before the Bullet Hits the Body - Dismantling Predictive Policing in Los Angeles'. <https://stoplapdspying.org/before-the-bullet-hits-the-body-dismantling-predictive-policing-in-los-angeles/>.
- Straw, Isabel. 2020. 'The Automation of Bias in Medical Artificial Intelligence (AI): Decoding the Past to Create a Better Future'. *Artificial Intelligence in Medicine* 110 (November): 101965. <https://doi.org/10.1016/j.artmed.2020.101965>.
- Subramanian, Ramanathan, Julia Wache, Mojtaba Khomami Abadi, Radu L. Vieriu, Stefan Winkler, and Nicu Sebe. 2018. 'ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors'. *IEEE Transactions on Affective Computing* 9 (2): 147–60. <https://doi.org/10.1109/TAFFC.2016.2625250>.
- Suchman, Lucy. 1995. 'Making Work Visible'. *Communications of the ACM* 38 (9): 56–64. <https://doi.org/10.1145/223248.223263>.
- . 2006. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511808418>.
- . 2012. 'Configuration'. In *Inventive Methods*, edited by Celia Lury and Nina Wakeford, 48–60. Routledge. <https://doi.org/10.4324/9780203854921-4>.
- Suchman, Lucy A., and Randall H. Trigg. 1993. 'Artificial Intelligence as Craftwork'. In *Understanding Practice: Perspectives on Activity and Context*, edited by Jean Lave and Seth Chaiklin, 144–78. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511625510.007>.
- Sundar, S. Shyam, and Jinyoung Kim. 2019. 'Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information'. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9. CHI '19. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3290605.3300768>.
- Thompson, Courtney E. 2021. *An Organ of Murder: Crime, Violence, and Phrenology in Nineteenth-Century America*. Rutgers University Press.
- Thylstrup, Nanna Bonde. 2022. 'The Ethics and Politics of Data Sets in the Age of Machine Learning: Deleting Traces and Encountering Remains'. *Media, Culture & Society* 44 (4): 655–71. <https://doi.org/10.1177/01634437211060226>.
- Timmermans, Stefan, and Iddo Tavory. 2012. 'Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis'. *Sociological Theory* 30 (3): 167–86. <https://doi.org/10.1177/0735275112457914>.

- Tucker, William H. 2009. *The Cattell Controversy: Race, Science, and Ideology*. Baltimore, UNITED STATES: University of Illinois Press.
<http://ebookcentral.proquest.com/lib/ed/detail.action?docID=3413990>.
- Tutton, Richard. 2021. 'Sociotechnical Imaginaries and Techno-Optimism: Examining Outer Space Utopias of Silicon Valley'. *Science as Culture* 30 (3): 416–39. <https://doi.org/10.1080/09505431.2020.1841151>.
- Ueno, Naoki. 2000. 'Ecologies of Inscription: Technologies of Making the Social Organization of Work and the Mass Production of Machine Parts Visible in Collaborative Activity'. *Mind, Culture, and Activity* 7 (1–2): 59–80.
<https://doi.org/10.1080/10749039.2000.9677648>.
- Vertesi, Janet. 2020. *Shaping Science: Organizations, Decisions, and Culture on NASA's Teams*. University of Chicago Press.
- Wache, Julia, Ramanathan Subramanian, Mojtaba Khomami Abadi, Radu-Laurentiu Vieriu, Nicu Sebe, and Stefan Winkler. 2015. 'Implicit User-Centric Personality Recognition Based on Physiological Responses to Emotional Videos'. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 239–46. ICMI '15. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/2818346.2820736>.
- Walsh, Colin G, Beenish Chaudhry, Prerna Dua, Kenneth W Goodman, Bonnie Kaplan, Ramakanth Kavuluru, Anthony Solomonides, and Vignesh Subbian. 2020. 'Stigma, Biomarkers, and Algorithmic Bias: Recommendations for Precision Behavioral Health with Artificial Intelligence'. *JAMIA Open* 3 (1): 9–15. <https://doi.org/10.1093/jamiaopen/ooz054>.
- Wang, Xiaowei. 2021. 'A New AI Lexicon: Care'. In *A New AI Lexicon*.
<https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-care-82c9031c98c4>.
- Whittaker, Meredith. 2021. 'The Steep Cost of Capture'. *Interactions* 28 (6): 50–55.
<https://doi.org/10.1145/3488666>.
- Williams, David R., and Toni D. Rucker. 2000. 'Understanding and Addressing Racial Disparities in Health Care'. *Health Care Financing Review* 21 (4): 75–90.
- Wilson, Elizabeth A. 2011. *Affect and Artificial Intelligence*. University of Washington Press.
- Woolgar, Steve. 1990. 'Configuring the User: The Case of Usability Trials'. *The Sociological Review* 38 (1_suppl): 58–99. <https://doi.org/10.1111/j.1467-954X.1990.tb03349.x>.
- Wulff, Kristin, and Hanne Finnestrand. 2022. 'It Is like Taking a Ball for a Walk: On Boundary Work in Software Development'. *AI & SOCIETY* 37 (2): 711–24.
<https://doi.org/10.1007/s00146-021-01175-3>.
- York, Emily. 2018. 'Doing STS in STEM Spaces: Experiments in Critical Participation'. *Engineering Studies* 10 (1): 66–84.
<https://doi.org/10.1080/19378629.2018.1447576>.

- Zeavin, Hannah. 2021. 'A New AI Lexicon: CARE'. In *A New AI Lexicon*.
<https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-care-a1243f0e2bad>.
- Zhang, Ting, Ri-Zhen Qin, Qiu-Lei Dong, Wei Gao, Hua-Rong Xu, and Zhan-Yi Hu. 2017. 'Physiognomy: Personality Traits Prediction by Learning'. *International Journal of Automation and Computing* 14 (4): 386–95.
<https://doi.org/10.1007/s11633-017-1085-8>.
- Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019*. Profile Books.
- Zwitter, Andrej, and Oskar J. Gstrein. 2020. 'Big Data, Privacy and COVID-19 – Learning from Humanitarian Expertise in Data Protection'. *Journal of International Humanitarian Action* 5 (1): 4. <https://doi.org/10.1186/s41018-020-00072-6>.

Annex A: Consent form



THE UNIVERSITY *of* EDINBURGH
School of Social and
Political Science

INFORMED CONSENT FORM

Research project title: **Seeing Affect: Knowledge Infrastructures in Facial Expression Recognition Systems**

Research investigator: Benedetta Catanzariti

Please initial each box

If you are happy to participate in the research, please initial each box as appropriate (leave blank any box for which you prefer not to give consent) and then sign this form at the end:

- | | | |
|-------|--|--------------------------|
| 1. | The researcher has given me my own copy of the Participant Information Sheet, and I have had the opportunity to read and consider the information. | <input type="checkbox"/> |
| <hr/> | | |
| 2. | I have been given the opportunity to ask any further questions and have had these questions answered to my satisfaction. | <input type="checkbox"/> |
| <hr/> | | |
| 3. | I have been given information about how my data will be stored and used during and after the end of the research, and I have read and understood this. | <input type="checkbox"/> |
| <hr/> | | |
| 4. | I understand that the interview will be recorded, and my words will be anonymized so that I cannot be identified. These may be quoted in academic publications, articles, books, reports, web sites and presentations related to the research project. | <input type="checkbox"/> |

-
5. I agree that the researcher can re-contact me at a future date should she wish to follow up on this research.

☐

-
6. I understand that my taking part is voluntary; I can withdraw from the project later, and I do not have to give any reasons for why I no longer want to take part. I have read and understood the Participation Information Sheet about the implications of withdrawing at different points during the life of the project.

☐

-
7. I understand I can ask for specific quotes or statements not to be used (or to be redacted from the data) if I wish.

☐

-
8. I understand that if I want to withdraw from the project, I can contact Benedetta Catanzariti () or her supervisors Dr Morgan Currie () and Dr Ewa Luger (), who will discuss with me how existing data will be managed, as outlined in the Participant Information Sheet.

☐

I agree to take part in this research project

Name of Research Participant:

Date:

Signature:

This research project has been approved through the ethical review process in the School of Social and Political Sciences at the University of Edinburgh. If you have any further questions or concerns about this study, please contact:

Benedetta Catanzariti

Or supervisors

Morgan Currie

Ewa Luger

Annex B: Participant information sheet



THE UNIVERSITY of EDINBURGH
School of Social and
Political Science

PARTICIPANT INFORMATION SHEET

Project title

Seeing Affect: Knowledge Infrastructures in Facial Expression Recognition Systems

Research investigator: Benedetta Catanzariti

Science, Technology and Innovation Studies, University of Edinburgh

ABOUT THE PROJECT

Emotion Artificial Intelligence (EAI) is a subfield of artificial intelligence that detects, measures, reproduces, and reacts to human emotions. There are now a range of important health-related and commercial applications of EAI, from monitoring emotional and mental states of patients for improved health care delivery, to illuminating consumer behaviour in the retail sector and monitoring drivers' attention to enhance road safety. AI systems have also, in recent years, made the headlines due to an increasing awareness in civil society of their technological errors and the potential for biases that discriminate against vulnerable groups. This study seeks to understand how designers of EAI anticipate both the societal gains of this powerful technology, as well as its potential limitations and biases, in order that they be developed to avoid social harms. Through interviews and fieldwork with companies developing EAI, I will explore the theories and methodologies guiding designs of EAI, including concerns these designers take for mitigating bias and harm. One major goal of the project is to promote greater public understanding of how EAI research employs methods to address societal concerns, such as those around ethics and fairness.

Why have I been invited to participate?

You have been invited to participate in this study because your experience in the field of emotional AI would be of great benefit to the study.

Do I have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part, you will be given this Participant Information Sheet to keep and be asked to sign an Informed Consent Form. If you decide to take part you are still free to withdraw from the study at a later date,

without giving a reason and without any impact on any services you are using. Only the researcher will have access to the research data.

What does taking part involve?

You will be asked a number of questions regarding the design process and applications of EAI. The interview will take online and at a time that it is convenient for you and will last approximately 30/60/90 minutes (as agreed with the researcher). The interview will be recorded, and a transcript will be produced. You can request a copy of the transcript and correct any factual errors. Access to the recording and the interview transcript will be limited to Benedetta Catanzariti. The recording and the transcript will be stored securely, safely and in accordance with Data Protection Act 2018, the General Data Protection Regulation (2018) and with University of Edinburgh data protection policy.

Any summary interview content, or direct quotations from the interview, that are made available through academic publication or other academic outlets will be anonymized so that you cannot be identified, and care will be taken to ensure that other information in the interview that could identify yourself is not revealed. Any variation of these conditions will only occur with your further explicit approval.

What are the possible benefits of taking part?

By sharing your experiences, you will be helping me and the research community to better understand the theories and methodologies underpinning the development and applications of EAI technologies.

What if I want to withdraw from the project?

Agreeing to participate in this project does not oblige you to remain in the study or to have any further obligations to the research project. If at any stage you no longer want to be part of the study, you can withdraw from the project by contacting Benedetta Catanzariti (benedetta.catanzariti@ed.ac.uk). You should note that your data may be used in the production of formal research outputs (e.g. journal articles, conference papers, reports) prior to your withdrawal and so you are advised to contact the researcher at the earliest opportunity should you wish to withdraw from the study.

If you withdraw from the project all the information and data collected from you, to date, will be destroyed and your name removed from all the project files.

How will my data be looked after during the project?

All your data will be processed and stored in accordance with the General Data Protection Regulation (GDPR) along with the Data Protection Act 2018 (DPA). The project will be also be guided by and adhere to the University of Edinburgh's data protection guidance and regulations. All personal details, including contact details, addresses, phone numbers etc., will be kept strictly confidential with the researcher, stored on password-protected and encrypted devices in accordance with the General Data Protection Regulation, and the latest University of Edinburgh data security protocols.

Electronic project data and paper records will be transferred to password-protected and encrypted devices, only accessible to the researcher. Your Consent Form will be stored separately from your responses.

What will happen to my data after the end of the project?

Data will be retained by the researcher for possible future (re)use and stored on the University secure server.

What will happen with the results of the research project?

The results of this study will be used for my PhD thesis and might be published in academic articles, books, reports and presentations. I may also use data from this project in teaching. Quotes from your interview may be used in these outputs, and these will be anonymous unless you give permission for your real name to be used in the Consent Form.

Who is organising the research?

I am conducting the research as a PhD student in Science, Technology and Innovation Studies, School of Social and Political Sciences at the University of Edinburgh.

Who has approved this project?

This research project has been approved through the ethical review process in the School of Social and Political Sciences at the University of Edinburgh.

Contact for Further Information

If you have any further questions about this project, please contact
Benedetta Catanzariti
PhD student in Science, Technology and Innovation Studies
University of Edinburgh

If you have any concerns about the way in which the project has been conducted, or you wish to make a complaint, you can contact the Deputy Director of Research (Ethics and Integrity) in the School of Social and Political Sciences:

Deputy Director of Research (Ethics and Integrity)
School of Social and Political Sciences
University of Edinburgh
Edinburgh EH8 8LN
e-mail: ethics-ssps@ed.ac.uk

Thank you

Date

Annex C: Interview guide

General interview guide

1. Contextual information

- Can you tell me something about your background? (What in your background led you to work on human behavior?)
- / How did you come to the field?

2. Understanding your organization

- Could you tell me about your organization and its aims?
- Who are your current clients/users? Who are your future or imagined clients/users? Please feel free to answer at the level of specificity you are comfortable with.
- What is your role in the organization? What responsibilities do you have? Do you report to anyone? Who reports or is accountable to you?

3. Understanding Design Practice

- What is the team trying to do?
- How is the team structured? How is work distributed within the team?
- Could you describe a recent task or challenges you have been working on?
- What different fields of expertise are involved? / How do you communicate with each other? How do you identify when there's a piece of information you need from another team member for something you're working on, and how do you figure out whom to ask?
- Can you think of a recent example, even if it's a little thing, of getting information from a colleague that you need for your work? What kinds of information do colleagues come to you to get when they need it?

[I am now going to ask a series of questions about data and classification in your work]

- What datasets do you use? Can you explain how you create a dataset for a facial expression recognition system? If you use off-the-shelf datasets, how do you choose which ones to use, how do you bring them into your system, clean them, etc.?
- Who labels the images? How do you label the images?
- How do you define affect/emotions/mental states? are there other definitions that are used on your team / in this firm? how do you reconcile these?

- how do you think about accuracy/reliability? what counts as a good-enough classification? can you give an example?

4. Issues and challenges

- What do you think is the biggest issue that your field/company is facing?
- how do you think about prospective users in your research? Where do their perspectives enter in your design process?
- Have you looked into the psychology and related literature on emotion and facial expression recognition? Do you consider yourself knowledgeable / expert in the relevant science? What do you consider to be the relevant science?
- Have you heard about the debate around the reliability or appropriateness of emotion recognition science? (debates about the assumptions behind emotion recognition suggest that it rests on a questionable scientific foundation).

5. Going forward

- How do you think these challenges might be addressed? And how can applications of affective technology be validated? (e.g. Experience of use and users? Stakeholders e.g. professional users, regulators, end users)
- how has your work pattern changed under Covid? Are there any new challenges or opportunities? (e.g. facial detection and face masks or the massive increase in use of video conferencing in new groups of people and settings)

6. Space for any additional topics that you might think important to mention

- Who else should I talk to?
- Are there aspects of what we talked about that would be especially good to follow up with particular people?

Interview guide for FACS coders and annotators

- What in your background led you to become a FACS coder/annotator?
- Could you tell me a little bit about your experience with learning the FACS/relevant coding scheme?
- How do you employ FACS/relevant annotation scheme in your work? (in everyday life?)
- Have you worked with other FACS coders/annotators? Can there be disagreement sometimes? How do you deal with disagreement?

- Can you tell me more about your experience coding facial expressions?
- FACS is often embedded in projects that ultimately aim at detecting “emotions”, “mood”, sometimes it’s used for the automated detection of internal states such as depression or anxiety that are associated with “abnormal” facial movements. What are your thoughts on these applications?