

**Reshaping Privacy Norms in the Age of Emotion AI: Socio-Technical Pathways for
Emotional Privacy, Human Agency, and Dignity**

by

Kat Roemmich

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2025

Doctoral Committee:

Associate Professor Florian Schaub, Chair
Assistant Professor Matthew Bui
Professor Kirsten Martin
Professor Emily Mower Provost

Kat Roemmich
roemmich@umich.edu
ORCID iD: 0000-0003-2730-1586

© Kat Roemmich 2025

DEDICATION

For the ones this work is about—and the ones it's for.

ACKNOWLEDGEMENTS

To the many hearts and minds who have guided, challenged, and sustained me through this journey—thank you. I cannot name you all, but I offer these acknowledgments to a few whose contributions made the work what it is, and to the many others whose care and support made it all possible.

To my advisor, Florian Schaub—your support throughout this research journey carries special meaning for this dissertation. When I entered the program, I had little sense of what academic research was: how knowledge is produced, how it accumulates and advances, and what it makes possible for society. Early on, you shared two lessons that have stayed with me: that the work of a PhD is to press against the edge of what is known, even if only slightly, and to cultivate the ability to learn whatever must be learned to answer a question well. Thank you for the support and space to discover the value of that work, and to hold fast to its integrity.

To Kirsten Martin—one of the happiest surprises of my research career was the chance to meet you just after I had come to admire and engage with your work (thanks, Florian!). They say never meet your heroes, and I’m thankful that your mentorship has turned out to be a rewarding exception. You’ve revived my motivation and sharpened my thinking in ways I cannot thank you enough for.

To Emily Mower Provost—your intellectual enthusiasm and care are qualities I deeply admire, and I’m grateful you brought both to your role on my doctoral committees. Your insights have pushed me to make my work more accessible, more relevant, and better attuned to the communities it aims to reach. Your support in getting it there means a lot. Thank you.

To Matthew Bui—I appreciate you for serving on my committee and for your thoughtfulness in each interaction. Though our windows for connection have been brief, your insights on how research can intersect with civil rights and advocacy continue to thread through how I think about the broader stakes of this work. I’m grateful for the presence you bring.

To colleagues who contributed to the research presented in this dissertation— Karen Boyd, Tillie Rosenberg, Nazanin Andalibi, Serena Fan, Mark Ackerman—thank you for your input and the resources you shared. I’m especially grateful to Paul Resnick for your support during the drafting process.

A special shout-out to Shanley Corvite, Nadia Karizat, and Cassidy Pyle—this work wouldn’t be the same without you. Thank you not just for collaborating on the research, but for your friendship

and steady support through all the complexities of navigating it. It made surviving a PhD not just possible, but worthwhile.

To my dear friends and labmates Kaiwen Sun, Abraham Mhaidli, and Yixin Zou—whether it was a technical puzzle, a half-formed idea, personal advice, or just showing up with an open ear, you've always been there, and it's meant the world to me. To Tanisha Afnan, Byron Lowens, Justin Petelka, Lu Xian, Jackie Hu, Allison McDonald, and the rest of the incredible spilab crew over the years—and PhD friends Jane Im, Rahaf Alharbi, Tyler Musgrave, Ben Zhang, Pelle Tracy, Jake Chanenson, Carolyn Guthoff—thank you for the feedback, solidarity, and sense of belonging that never failed to make the hardest stretches lighter. (To those I've failed to mention, please forgive me!)

To David Wallace—teaching information ethics with you was one of the great highlights of this journey. Filled with what brings me joy—lively discussion, collaborative inquiry, and intellectual generosity—that time left a lasting imprint on me, and on this dissertation. Thank you.

To my husband, Joseph—for keeping philosophy alive in my life. You were never drawn to academia—it's one thing I love about you—and yet you made sure it stayed lit for me: always a new book, a quiet nudge to stay close to what matters. It's what brought me back to this path, and what carried me through it. As we move forward into the unknown, thank you for making it possible for me to reach this dream, and for holding space for whatever comes next.

Above all, to my children, who have been with me through all of it. To my bright star, Daphne—your radiant sense of what matters in the here and now has kept me grounded, and your gift for really seeing others has been a steady well of compassion and clarity when my own ran low. To my inquisitive bear, Orson—your big heart and boundless curiosity have kept me inspired and open to the wonder that keeps the work honest, deep, and alive. Your vast interests in how things work and why people do what they do are constant reminders for me to keep asking better questions—and to answer them with the kind of attention that carries its own form of care. And to my kindred spirit, Preston—what began as bedtime readings from the likes of Plato and Aristotle, Hegel and Heidegger, during my undergraduate years became, years later, a return to those same texts as you began reading them in earnest yourself while I finished this dissertation. It was a rare kind of experience that let us see both the philosophers—and each other—anew. Thinking and growing alongside you has been an unforeseen gift, and the clearest reminder that the best thinking is never done alone. For each of you, growing up with a mom getting her PhD could not have been easy. The questions I've asked, the commitments behind them, the shape this work has taken—all bears traces of your patience, your presence, and your love. The meaning of this dissertation, for me, begins and ends with you.

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	ix
LIST OF TABLES	x
LIST OF APPENDICES	xi
ABSTRACT	xii
CHAPTER	
1 Introduction	1
Part I: Conceptual Foundations	3
2 When Human Emotions Meet Machines	4
2.1 The Role of Emotions in Human Flourishing	4
2.2 How Emotions are Made	8
2.2.1 Individual and Social Differences	12
2.2.2 A Unified Theory of Emotions as Evaluative Judgments	14
2.3 What is Emotion AI?	15
2.3.1 System Model	16
2.3.2 Theoretical Assumptions	19
2.3.3 Emotion Labels	21
2.3.4 Algorithmic Goals	23
2.3.5 Emerging Systems	24
Part II: The Empirical Case for Emotional Privacy	24
3 Emotion AI on Social Media	25
3.1 Introduction	27
3.2 Background and Related Work	30
3.2.1 Emotion and Risk Profiling through Social Media Traces	30
3.2.2 Human Values and Ethical Stakes	31

3.3 Methods	31
3.4 Findings	36
3.4.1 What Only Humans Can Offer	37
3.4.2 Emotion AI Uses for Social Good	40
3.4.3 Emotion AI's Potential for Harm	44
3.4.4 Conditional Trust and Acceptance	46
3.5 Discussion	48
3.5.1 What Makes an Ethical and Trustworthy System?	49
3.5.2 What if Individuals Consent?	50
3.5.3 Harm to Vulnerable Populations	52
3.5.4 Trading Privacy for Safety	54
3.6 Conclusion	55
4 Emotion AI in Hiring	57
4.1 Introduction	57
4.2 Background and Related Work	58
4.2.1 Emotions in Hiring	59
4.2.2 AI in Datafied Organizations	59
4.2.3 Workplace Talent Management	60
4.2.4 Criticisms of Algorithmic Hiring	61
4.3 Methods	62
4.4 Findings	67
4.4.1 Hiring (In)accuracy	67
4.4.2 Hiring (Mis)fit	73
4.4.3 Hiring (In)Authenticity	78
4.5 Discussion	82
4.5.1 Designing for Perceptible Fairness	83
4.5.2 Enforcing Fairness	83
4.6 Conclusion	87
5 Emotion AI at Work	88
5.1 Introduction	88
5.2 Background and Related Work	90
5.2.1 Surveilling Workers' Interiority	90
5.2.2 Workplace Applications of Emotion AI	91
5.2.3 Risks to Workers	92
5.3 Methods	93
5.4 Findings	98
5.4.1 Crossing Emotional Lines	98
5.4.2 Emotional Labor, Coerced and Claimed	101
5.4.3 Beyond the Usual Harms	103
5.5 Discussion	113
5.5.1 Naming Emotional Privacy	114
5.5.2 Governing Emotional Privacy	115
5.5.3 Designing for Emotional Privacy	117

5.6 Conclusion	118
Part III: Measuring Emotional Privacy Across Contexts	118
6 Emotional Privacy Judgments and Vulnerabilities in Work and Health	119
6.1 Introduction	121
6.2 Background and Related Work	126
6.2.1 Contextual Factors	127
6.2.2 Socio-Demographic Variations	128
6.2.3 Privacy Belief Influences	130
6.3 Methods	132
6.4 Comparing Normative Judgments of Emotional Privacy	148
6.4.1 Contextual Vulnerability: Setting, Data Input, and Purpose	154
6.4.2 Identity-Based Effects	168
6.4.3 The Role of Privacy Beliefs, Trust, and Data Sensitivity	171
6.5 Relational and Contextual Impacts: Perceived Benefits and Risks	174
6.5.1 Emotion AI in the Workplace	175
6.5.2 Emotion AI in Healthcare	189
6.6 Discussion	204
6.6.1 Bounding Inference Purpose	205
6.6.2 Emotion Data Sensitivity	206
6.6.3 Contextual Vulnerability and Emotional Privacy	207
6.7 Conclusion	209
Part IV: Drawing the Dignity Line in Privacy Theory and Governance	211
7 Inviolate Personhood	212
7.1 Introduction	214
7.2 Background and Related Work	215
7.2.1 The Moral Origins of Privacy	216
7.2.2 Instrumental Privacy Approaches and Fragmentation	220
7.2.3 Contextual Integrity and the Pluralist Turn	222
7.3 Dignity as a Moral Minimum Standard in Contextual Integrity	227
7.3.1 The Basic Norm of Human Dignity	228
7.3.2 Defining and Measuring Human Dignity	234
7.4 Tracing the Dignity Line in Practice	244
7.4.1 Crisis Text Line: The Limits of Internal Governance	245
7.4.2 Clearview AI: The Fragility of Rights Without Dignity	249
7.4.3 Replika: When Harm Metrics Neglect Contextual Vulnerability	257
7.5 Discussion	264
7.6 Conclusion	266
Part V: Concluding Discussion	268
8 From Vibes to Thresholds: Specifying Dignity	269

8.1 Privacy as a Fundamental Right: Do We Need Emotional Privacy as Another Enumeration?	272
8.2 Dispositional and Emotional Privacy	273
8.2.1 Deception Detection	274
8.2.2 Emotion Detection	275
8.2.3 From Information Type to Capability Threat	277
8.3 Emotion Structures, Data Brokers, and Polarization	277
8.4 Closing Reflection	278
 APPENDICES	279
A.1 Employment Context Factorial Vignettes	279
A.2 Healthcare Context Factorial Vignettes	284
A.3 Open-ended Survey Questions, Separately Answered for Each Context	289
A.4 Post-test Socio-Demographic Questions	290
A.5 Post-test Individual Belief Questions	292
A.5.1 General Privacy Concerns	292
A.5.2 Risk Beliefs	293
A.5.3 Trust Beliefs	293
A.5.4 Perceptions of Data Sensitivity	294
A.6 Qualitative Analysis Sample Breakdown	297
A.7 Quantitative Results: Plotted Coefficients with Error Bars	297
B.1 Pre-Screening Survey	298
B.2 Interview Protocol	299
 BIBLIOGRAPHY	304

LIST OF FIGURES

FIGURE

2.1	Emotion AI Building Blocks	17
2.2	3D and Hourglass of Emotions Model	21
6.1	Presentation of Vignettes	139
6.2	Predicted Comfort by Data Input	156
6.3	Predicted Comfort by Purpose	158
6.4	Predicted Comfort by Socio-demographics	168
6.5	Predicted Comfort by Privacy Beliefs	171
6.6	Perceived Data Sensitivity Comparisons – Employment Context	173
6.7	Perceived Data Sensitivity Comparisons – Healthcare Context	173
7.1	Capabilities–Contextual Integrity (CA–CI) Theoretical Framework – Integrating Fixed Dignity Thresholds, Purpose Parameter	243
7.2	Capabilities–Contextual Integrity (CA–CI) Evaluation of Crisis Text Line.	247
7.3	Capabilities–Contextual Integrity (CA–CI) Evaluation of Clearview AI.	252
7.4	Capabilities–Contextual Integrity (CA–CI) Evaluation of Replika.	260
A.1	MacArthur Scale of Subjective Social Status	292
A.2	Coefficient Plot with Error Bars	297

LIST OF TABLES

TABLE

5.1	Participant Demographic Table	96
6.1	Emotional Privacy Norm – Fixed CI Parameters	134
6.2	Vignette Variables	138
6.3	Descriptive Sample Statistics by Socio-demographic Level	142
6.4	Analysis Factors and Levels	143
6.5	Alpha Binary Table – Workplace Context	147
6.6	Alpha Binary Table – Healthcare Context	147
6.7	Regression Results – Employment Context	150
6.8	Regression Results – Healthcare Context	152
6.9	Summary of Key Quantitative Findings Across Factors and Sample Comparisons . . .	153
6.10	Summary Statistics – Mean and Estimated Comfort Levels by Context and Sample .	154
6.11	Emotion Inference Purpose Groupings and Levels	159
7.1	Comparative Overview of Contextual Integrity and Capabilities Approach	242
7.2	CA-CI Evaluation of Crisis Text Line	249
7.3	CA-CI Evaluation of Clearview AI	257
7.4	CA-CI Evaluation of Replika	263
A.1	Socio-demographic Sample Breakdown	296

LIST OF APPENDICES

A Supplemental Materials: Factorial Vignette Survey with Open Ended Qualitative Responses	279
B Supplemental Materials: Worker Recruitment and Interview Protocol	298

ABSTRACT

Technologies that automatically detect, interpret, respond to, and interact with human emotion make bold promises about their potential to improve human life—to *humanize* technology [290]. But in deployment, these promises confront the reality of lived experience—where power, agency, and dignity are not theoretical variables but fragile conditions of social life. These same technologies can enhance or erode human flourishing. This dissertation argues that where the needle turns depends on a single, urgent question:

Where do we draw the justificatory line between acceptable and unacceptable data flows?

Grounded in empirical studies of privacy perceptions, judgments, and harms, this dissertation develops the concept of *emotional privacy* and defends its protection as a moral and political imperative in AI-mediated societies. Linking mixed methods studies in social media, the workplace, and healthcare settings to broader philosophical debates on the value of emotions and the foundations of privacy, this dissertation shows that safeguarding emotional privacy is not just about regulating emotional information—it is about securing the human capacity to live with agency, dignity, and meaning in interdependent social life.

Part I lays the theoretical and technical foundations, surveying philosophical and scientific accounts of emotion alongside the mechanics of emotion AI. Part II builds the empirical case for emotional privacy as a distinct interest through qualitative studies with data subjects [731, 735] and a content analysis of commercial vendors [734], revealing high-risk emotional data flows that reconfigure what it means to be human.

Part III reports a mixed-methods factorial vignette study that quantitatively measures emotional privacy judgments using Helen Nissenbaum's theory of privacy as Contextual Integrity (CI) [732], supported by qualitative analysis of perceived risks and benefits in employer [219] and healthcare [733] scenarios. While the findings validate CI's insight—that privacy judgments are grounded in context-relative informational norms and teleological aims—they also surface a deeper, context-independent expectation: to maintain dignity.

In response, Part IV introduces the Capabilities–Contextual Integrity (CA–CI) framework, which integrates CI's ecological mapping of data flows with Martha Nussbaum's Capabilities

Approach. CA-CI retains CI's analytic strengths but anchors its justificatory logic in a shared threshold of human dignity. It formalizes *purpose* as a sixth constitutive parameter and treats the set of core capabilities constitutive of a life with dignity—agency in matters of life; bodily health; bodily integrity; emotions; practical reason; senses, imagination, and thought; affiliation; play; other species; control over one's environment, up to their minimum thresholds [658]—as fixed transmission principles. Under CA-CI, a data flow is presumptively appropriate only if its foreseeable impact preserves each person's dignity threshold; if it predictably erodes even one below bar, it fails the dignity test, prohibited unless and until it can be redesigned with the the essential constraints in place. I demonstrate CA-CI's practical utility through three case studies, showing how the framework surfaces dignity-eroding flows that procedural or classification-based approaches alone can miss.

By drawing the normative line in the sand at dignity, CA-CI offers policymakers, regulators, technologists, and other system evaluators a principled roadmap for building and governing socio-technical systems that enhance, rather than undermine, the human capacities on which a just and flourishing society depends.

CHAPTER 1

Introduction

Human emotions are essential components of human intelligence: personal judgments that reflect our values and beliefs—the lens through which we see the world—revealing what matters to us and its salience to our lives, often before we fully comprehend or articulate it ourselves [659]. Technologies designed to automatically infer and respond to human emotion thus at once promise significant advancements to human welfare and well-being in key social domains, including mental health, workplaces, and social media [495, 434, 555, 827, 862, 454, 665, 366, 593] while presenting significant risks to individuals and society. The capacity for machines to reach into and shape inner life challenges foundational assumptions—our ability to think, feel, and act freely; to participate in social and institutional life; and to maintain just balances of power within and across social domains. These challenges demand scrutiny of the conditions under which emotion inferences are made and acted upon.

What is the moral and political significance of emotional information and how it is collected and used? What do we call the sense when machines mine, extract, interpret, act upon, and commodify inferences about our emotional lives—our personality, moods, dispositions, values, and beliefs? What is at stake when an algorithm detects a worker’s anger during a customer interaction, flags a patient as needing emotional support, or labels a teenager a suicide risk? How should we weigh potential benefits against the risks such technologies pose—and by what standards should we draw the line between appropriate use and unacceptable harm? Part empirical inquiry, part critique, part theory development, this dissertation demonstrates that existing frameworks in privacy and AI ethics cannot fully answer these emerging dilemmas, and advances a new theoretical model for evaluating when technologies cross an unacceptable line.

Part I surveys leading theories of emotion and how they have been made tractable for technical systems, Parts II and III present the results of four empirical studies that examine the ethical and privacy implications of emotion AI technologies in three key domains: social media, the workplace, and healthcare. Part II draws on qualitative methods to center the experiences of emotion AI’s data subjects, surfacing the benefits, risks, and harms they encounter or anticipate. I name and frame these stakes as a problem of *emotional privacy*.

Part III extends these insights through a cross-contextual factorial vignette survey study, applying Helen Nissenbaum’s theory of privacy as Contextual Integrity (CI) [646] to empirically measure emotional privacy judgments and the factors by which they vary. Using mixed methods, I locate the perceived appropriateness of emotional data flows in the intuitive moral judgments of those situated within the relevant context. My findings confirm that CI’s theory of privacy effectively explains participants’ privacy judgments in many cases: approving of emotion AI uses aligned with a context’s core purposes, and disapproving of those that strained or disrupted them—consistent with CI’s normative heuristic [649]. Yet, the data also revealed a deeper throughline: judgments turned not only on breaches of contextual norms and goals, but also on breaches of a more basic moral threshold: one’s dignity.

Part IV turns to theory development. It traces the historical trajectory of privacy as a dignitary interest in law and philosophy to argue that the moral source of privacy claims lies not only in the integrity of a context but in the integrity of the self: rooted in our capacity for agency, to maintain the moral boundaries required to live with dignity in the face of external intrusion and constraint. These conditions of human dignity are not derivative to context, but what make context possible, functioning precisely because individuals retain the capacity to think, feel, and act as whole persons across them [872]. When machines infer our emotions, they gain access not merely to another enumerated type of sensitive information—they gain access to the constitutive components of moral personhood [656, 659].

By drawing on theoretical scholarship defending privacy as essential to developing, maintaining, and protecting dignity and the moral self, Chapter 7 begins by tracing privacy’s trajectory as a dignitary interest and argues for a second justificatory line within CI: recognizing flows as appropriate where they uphold respect for context *and* human dignity. This final chapter of the dissertation bridges empirical and normative insights to advance a novel governance model, the Capabilities-Contextual Integrity (CA-CI) framework. CA-CI integrates Martha Nussbaum’s Capabilities Approach (CA) [658] to establish concrete dignity thresholds as fixed transmission principles within CI. Through three case studies, I demonstrate how CA-CI’s evaluation of whether a flow aligns with contextual aims (CI) and the core capabilities essential to human dignity (CA) offers a robust standard for evaluating when data flows are appropriate (and when they are not) where current governance procedures may fall short.

Chapter V’s concluding discussion engages these concerns in light of AI’s accelerating trajectory. I assess how current and near-future advances strain existing regulatory logics and informational privacy paradigms, and outline key open questions for future governance and research.

This dissertation is about privacy, but also about dignity. To insist on the moral significance of emotional privacy is not merely to protect personal data or mitigate technology-enabled harm. More fundamentally, it is to defend the conditions under which people can think, express, and

become themselves—to pursue their own vision of the good. As AI systems increasingly engage with our inner lives—what we think, feel, believe, and value—that security has never been more urgent.

Part I: Conceptual Foundations

CHAPTER 2

When Human Emotions Meet Machines

This chapter establishes why emotions and privacy are not niche concerns but pre-conditions for human dignity and agency in an age of emotion AI. It proceeds in three moves. First, drawing on philosophy and the affective sciences, it frames emotions as *evaluative judgments* at the heart of human flourishing. Second, it shows how privacy functions as the condition of emotional agency and why machine inference threatens that function. Third, it sketches how prevailing theories of emotion translate (often problematically) into technical assumptions inside emotion AI systems. The result is a concise intellectual runway for Parts II–IV that follow, where empirical studies and the CA–CI framework take up the normative challenge.

2.1 The Role of Emotions in Human Flourishing

What exactly are emotions, and why do they matter for a dissertation on privacy and AI? Across two millennia of inquiry—stretching from Confucius and Aristotle to contemporary affective neuroscience—scholars have contested the very nature, structure, and function of emotion [70, 568, 756, 89]. Today the debate spans virtually every field that studies human (and non-human) life: law, biology, sociology, linguistics, economics, psychology, neuroscience, anthropology, computing, and more.

Despite disciplinary rifts, a growing consensus now treats emotions not as raw feelings or irrational eruptions but as *evaluative judgments*: appraisals about what is salient, valuable, or threatening in relation to our goals, commitments, and identities. Martha Nussbaum crystallises this view, describing emotions as “intelligent responses to the perception of value”—affect-laden cognitions that disclose what we care about and why [659]. On this eudaimonistic account, emotions

⁰Adapted and condensed from my preliminary-field milestone at the University of Michigan School of Information.

are inseparable from practical reason: they guide attention, shape belief, motivate action, and thus become constitutive ingredients of human flourishing.

Empirical research aligns with this philosophical insight. Cognitive studies show that emotions steer perceptual focus [384], bias reasoning and memory formation [101], and energise goal pursuit [614]. In social domains, emotional expressions signal intentions, coordinate cooperation, and forge durable bonds [567]. Developmentally, emotions scaffold identity and moral learning [913]. At the societal level, they underwrite shared norms and institutional arrangements [83]. In short, from the micro-phenomenology of feeling to the macro-structures of culture, emotion is the connective tissue of human life.

Because emotions perform these evaluative and coordinative functions, interference with emotional life is not a trivial matter. Misreading, manipulating, or forcibly exposing emotions can distort a person’s value landscape, undermining the very capacities—reflection, affiliation, practical reason—that Nussbaum’s Capabilities Approach identifies as thresholds of a dignified existence [658]. Emotional information is therefore not just another type of sensitive data; it is morally freighted. Any technology that detects or predicts emotional states acquires leverage over how individuals see the world and see themselves—leverage that can sustain, skew, or shatter autonomy.

This dissertation proceeds on that premise. Throughout the following pages I ask how emotion-sensing systems intersect with these evaluative underpinnings, and where the line must be drawn to protect the conditions of flourishing. The interdisciplinary tour that follows is necessarily selective, but it clarifies the stakes: understanding emotion theory is prerequisite to judging what emotion AI can validly infer, how those inferences shape dignity, and why emotional privacy must be treated as a first-order concern in data governance.

2.1.0.1 Social Norms

If emotions are evaluative judgments about what matters to us—as Nussbaum argues—then it follows that emotional expression and regulation are shaped not only by individual cognition, but by the cultural and moral norms that teach us what *should* matter. Social norms, in this sense, do not merely influence the performance of emotion; they encode shared beliefs about which emotional experiences are desirable, appropriate, virtuous, or taboo, and thereby mediate our efforts to live well.

Philosophical traditions across cultures have long embedded emotions within broader visions of the good life. In the West, Aristotle’s ethical system of *eudaimonism* casts flourishing (*eudaimonia*) not as hedonic pleasure but as the fulfillment of one’s inner potential or *daimon* [19, 877]. Happiness, in this view, is not a feeling to be pursued directly, but a higher-order satisfaction that emerges from living in accordance with virtue. This idea—that emotional maturity and self-cultivation are integral to moral excellence—permeates modern positive psychology and Western social norms

that enjoin individuals to “live their best life” and pursue happiness as both a personal and civic duty [402, 60, 597].

Yet, as Sara Ahmed has shown, this imperative has become entangled with gendered, racialized, and heteronormative scripts that weaponize happiness as a condition of belonging and a tool of social regulation [28]. The command to be cheerful, resilient, or “emotionally intelligent” under conditions of structural harm can obscure injustice and stigmatize resistance. Against this backdrop, Ahmed calls for a “happiness turn”: an embrace of emotional dissent, discomfort, and the “freedom to be unhappy” as modes of critique and alternative worldmaking.

Importantly, this Western fixation on happiness as the telos of emotional life is not universal. In Confucian moral philosophy, for instance, the ideal is not self-actualization through positive emotion, but relational harmony achieved through moderation, mutual responsiveness, and virtue expressed in affective restraint [816, 523]. Emotions are not fleeting feelings to be managed for optimal outcomes; they are cultivated dispositions—stable, internalized traits that reflect moral character. Affective balance, not expressive positivity, is the mark of ethical maturity. Sharing in another’s grief, for instance, is not something to be “fixed” or avoided, but a necessary expression of benevolence and co-suffering [816].

These divergent emotional norms have direct implications for the design and deployment of emotion AI. Technology deployments that aim to detect or regulate affect inevitably embed normative assumptions about which emotional states are desirable—and for whom. Systems designed to nudge users toward happiness, calm, or engagement risk reproducing the Western ideals from which those goals derive, along with their associated exclusions. The moral weight of these systems lies not just in what they detect, but in what they normalize. As Nussbaum’s framework helps illuminate, emotion AI’s evaluative force does not come from identifying internal states in the abstract, but from how those states are interpreted, ranked, and acted upon within a given social order.

Despite these cultural divergences, a shared insight emerges: emotions matter because they disclose what is meaningful to human life. Whether one locates moral excellence in happiness, balance, or benevolence, emotional life remains foundational to human flourishing. For this reason, systems that sense, infer, or manipulate emotions are not merely analyzing behavioral signals; they are operating at the heart of what it means to live well, and thus must be held to standards that protect that moral terrain.

2.1.0.2 Personal and Social Development

Across cultures and disciplines, emotions are recognized not only as internal experiences but as relational and social forces that shape who we are and how we live together. Regardless of how distinct emotional expressions are socially valued, emotions as both a phenomenon and process

are foundational to social life [83, 367, 546]. They undergird the development of relationships, social structures, and selfhood alike, enabling us to navigate moral meanings, construct identities, and form beliefs [913, 912]. Emotions are also motivational engines for intellectual curiosity and creative thought [235], with affective states shaping how individuals engage with ideas, learn from others, and act in the world [911].

At both micro- and macro-social levels, emotional processes constitute the fabric of collective life: institutions are not only governed by norms but animated by feelings such as trust, loyalty, shame, or resentment [83]. As Section 2.1.0.1 underscored, emotions reflect and reinforce shared ideals, while also enabling individuals to pursue personal and communal visions of flourishing. Whether in family life, education, work, or civic participation, emotional intelligence and expression are central to the realization of meaningful, agentic lives.

2.1.0.3 Privacy

Less often acknowledged in this literature is that emotional life depends upon the preservation of privacy. Emotions are not only about social connection but also about boundary-setting—what is disclosed, to whom, and under what conditions. Emotional privacy, in this sense, is not incidental; it is constitutive of our ability to flourish—an insight this I develop further in Part IV.

Emotions and privacy are closely interwoven. Privacy has emotional and affective dimensions [825], and emotions have deeply private ones. Altman’s theory of privacy regulation describes how individuals manage desired boundaries around access to the self—including emotional access—through psychological, spatial, and communicative cues [45]. These boundaries are dynamic: when privacy is lacking, we may withdraw from interaction to re-establish emotional equilibrium; when overly isolated, we may open ourselves more than usual [45]. The turbulence caused by breached emotional boundaries is often experienced as a privacy violation, producing psychological harms such as stress, shame, or loss of control [194].

Indeed, such harms can be intensified when emotional information is exposed, inferred, or acted upon without consent. Stark has noted that emotional distress is a recurrent theme in public responses to controversial privacy violations, including Facebook’s emotional contagion experiment and surveillance capitalism more broadly [801]. As Parts II and III show, people perceive emotion data as particularly sensitive—not only because of what it reveals, but because of how it feels to be seen and interpreted against one’s will.

These emotional dimensions also influence privacy behavior. People are more likely to disclose sensitive information online when they associate a site with positive emotional experiences—such as trust or safety [527, 866]—and less likely when they experience anxiety or discomfort [528, 828]. While emotional disclosures can foster support and intimacy, they also expose individuals to risk, especially when disclosures extend beyond their intended audience and become commodified by

third parties . Platforms often exploit these dynamics by fostering trust to increase engagement and data richness [867, 588], transforming emotional experiences into extractive assets.

Crucially, emotional privacy is not only about what we share, but what we reserve. The ability to feel one thing and express another—to manage emotional expression—is itself a privacy act [40]. A parent may reassure their child in a moment of crisis while privately feeling fear; a worker may express satisfaction while masking frustration. These practices rely on a basic assumption: that what we feel internally is ours to control, and not automatically transparent to others. As Chapter 5 shows, emotion AI threatens this assumption by bypassing expressive management altogether.

The ability to preserve emotional privacy—both individually and collectively—enables autonomy, safeguards vulnerability, and underpins social relationships built on mutual recognition and respect [564]. For instance, employees who can shape how their emotional state is perceived may enjoy greater job security or avoid punitive consequences [782, 703]. Privacy also enables emotional rest and reflection, providing a retreat from the performative demands of public life [417, 157]. Without this retreat, individuals face cumulative affective strain, undermining both wellbeing and dignity.

Thus, emotional privacy makes possible the selective self-disclosure through which intimacy is built, the affective regulation through which autonomy is preserved, and the expressive space through which dignity is realized—a foundational human interest central to both privacy and the capabilities that underwrite a flourishing life.

2.2 How Emotions are Made

Human emotion is a complex and contested concept, with definitions, functions, and structures varying across disciplines and cultures [426, 326]. The very term “emotion” did not emerge as a category for systematic study until the late 19th century, overlapping with earlier concepts such as “passions” and “sentiments” [269]. Scientific debates continue over the nature of emotion, partly due to the lack of standardized terminology and conceptual coherence within and across fields [442]. Despite this, two dominant meta-theoretical paradigms have shaped contemporary scientific and applied understandings: discrete and dimensional models [397].

The core distinction between these models concerns whether emotions are best understood as categorically distinct states (e.g., fear, anger, joy) or as continuous experiences that vary along multiple dimensions (e.g., valence, arousal, dominance). Discrete models typically align with biologically essentialist views, positing that emotions are universal, evolutionarily ingrained, and expressed similarly across cultures. Dimensional models, by contrast, accommodate more flexible, socially and culturally embedded accounts of emotion, often treating emotional states as fluid, overlapping, and context-sensitive.

Though these paradigms originate in scientific theory, they implicitly inform applied systems in affective computing and emotion AI. Commercial systems often operationalize emotion within either the discrete or dimensional framework, assigning labels like “happy” or “angry,” or estimating degrees of arousal or valence—while largely bypassing deeper theoretical debates about what emotions are or what it means to infer them [166, 621].

This section organizes a non-exhaustive review of emotion theories around this discrete/dimensional divide, reflecting the dominant schema adopted in affective computing. However, other classificatory approaches exist. In the social sciences, for instance, emotion theories are often grouped into three categories: categorical, dimensional, and appraisal-based models [372, 383]. Appraisal-based approaches define emotions in terms of dynamic evaluations of situational meaning—how events align or conflict with one’s goals, values, or expectations. While rich in explanatory power, such models remain underutilized in commercial emotion AI due to their contextual complexity and challenges in automated implementation [383, 802].

While appraisal theories are often underutilized in commercial applications due to their contextual and cognitive complexity, they offer critical insight into the moral and evaluative dimensions of emotion. Among them, Martha Nussbaum’s theory of emotions as value-laden judgments stands out as a philosophically rigorous and normatively rich account [659]. Though not typically cited in technical literature, it provides a crucial foundation for this dissertation’s treatment of emotional privacy. I therefore return to Nussbaum’s account in Chapter 7 to develop a normative model that centers emotional meaning, dignity, and human flourishing.

Importantly, the boundaries between these theoretical traditions are neither rigid nor mutually exclusive. Many contemporary models draw from both categorical and dimensional frameworks, and disciplinary affiliations do not map neatly onto specific paradigms. Nevertheless, understanding how these theories shape the assumptions built into emotion AI systems is essential for evaluating their validity, scope, and normative implications. The following subsections examine these two dominant models and their influence on the design and deployment of emotion AI.

Discrete, or basic, theories of emotion conceptualize emotions as a set of distinct, mutually exclusive categories—such as anger, fear, happiness, or sadness—that are biologically programmed and universally expressed. One of the most prominent proponents of this view is Paul Ekman, whose Basic Emotion Theory posits that certain emotions are evolutionarily hardwired and triggered by environmental stimuli, producing a cascade of automatic physiological responses that can be reliably observed across individuals and cultures [286].

This discrete perspective is often paired with biological and physiological accounts of emotion, which emphasize the evolutionary and embodied functions of affective states. While not all biological theories adopt the discrete view, many share core assumptions about the universality, innateness, and recognizability of emotion. To situate the rise of discrete emotion models, it is

useful to first trace the biological lineage of emotion theory—from Darwin’s naturalistic account to the James-Lange theory of bodily feedback, and later developments in neuroscience and embodied cognition.

Darwin’s *The Expression of the Emotions in Man and Animals* framed emotions as evolved adaptations, shared across species and serving communicative and survival functions [239]. Building on this, the James-Lange theory proposed that emotions arise from the perception of bodily change, suggesting that we feel afraid *because* we tremble. These ideas laid the groundwork for later models that treated emotional experience as a byproduct of physiological activation, further advanced by contemporary neuroscience and embodied cognition research. Such perspectives emphasize the measurable, biological underpinnings of emotion, often operationalized through heart rate, facial musculature, or brain activity.

Ekman’s Basic Emotion Theory drew heavily from these foundations, asserting that certain emotions are universally expressed through facial expressions and can be detected reliably across cultures. His cross-cultural studies, often involving photographs of posed facial expressions, were widely influential in establishing the idea that emotional categories are fixed and identifiable. This view has heavily shaped emotion recognition systems in computing, which often rely on categorical labeling of affective states based on presumed universal features.

However, critics have long challenged these assumptions. Empirical evidence shows substantial variation in how emotions are expressed and interpreted across individuals and cultures, and in many cases, the same facial expressions can correspond to different emotions depending on context [89]. Lisa Feldman Barrett, for example, calls this the “emotion paradox”: despite strong intuitive beliefs that we can recognize emotions when we see them, scientific evidence reveals that emotional categories are not fixed biological facts but constructed interpretations based on context, culture, and learned associations [90].

Nonetheless, the intuitive appeal of discrete models—particularly their alignment with visual and behavioral cues—continues to shape the design of many emotion AI systems. As the next section discusses, continuous and dimensional theories offer an alternative to this fixed-categorical approach, challenging the assumptions of universal expression and inviting richer, more socially embedded accounts of emotion.

Biological Functions

Some theories view emotion as biologically produced and adaptive, evolved to help organisms respond to salient environmental stimuli. These perspectives generally posit that emotional states are mediated by physiological mechanisms that provoke patterned responses—“inherited, reflex-like modules that cause a distinct and recognizable behavioral and physiological pattern” [94]. While such biological views often underpin discrete theories, they need not assume universal

categories or mutually exclusive emotions.

Darwin’s naturalism Darwin’s comparative work on humans and animals is widely regarded as foundational to the scientific study of emotion [238, 91]. He viewed emotional expressions as biologically inherited adaptations—raised eyebrows, for example, both widen the visual field and communicate surprise to others.

James–Lange theory Building on Darwin, the James–Lange theory reframed emotion as a bodily phenomenon: we do not tremble because we are afraid; we are afraid because we tremble [445]. That is, physiological responses to stimuli precede—and give rise to—emotional experience, through one’s awareness of those changes [512].

Embodied emotion Later embodied approaches retained the primacy of the body but rejected the requirement of conscious awareness. Instead, they posit that emotions are enacted through lived, subjective bodily states—emotion is not something we have, but something we do [208, 207].

Neuroscientific models The James–Lange view also informed neuroscientific approaches, which investigate correlations between emotional experience and neural dynamics [109, 900]. The Cannon–Bard theory challenged James–Lange by proposing that physiological arousal and emotion occur simultaneously and independently [587]. Later models distinguished affective experience from expression—as in the Papez–MacLean circuit, where feelings arise in the limbic system while expressions are governed by the hypothalamus [881]. Contemporary work emphasizes the amygdala [519], though much remains unresolved. As Barrett notes, the precise mapping of neural activation to discrete categories or affective dimensions remains a central puzzle [94]. Still, emotion-AI projects increasingly aim to bypass expression altogether by directly translating neural patterns into affective data [540].

Biological function models Functional theories emphasize what emotions do. They regulate arousal [496], guide attention [785], and facilitate communication and affiliation [294]. Simon, for instance, saw emotion as an interrupt system—a fast, adaptive override of cognition to enable re-prioritization [785]. His view echoes Dewey’s earlier pragmatist theory of emotion, which synthesized Darwin’s emphasis on expression and James’s focus on bodily feeling by describing emotion as emerging from goal conflict or internal tension [260]. Both challenged the Cartesian legacy that treats emotion as antithetical to intelligence [567]. While Simon’s work helped shape early AI, mainstream approaches continued to treat emotion as noise, a stance now under renewed scrutiny [121].

This pragmatist-functional lineage contrasts with appraisal theories, which emphasize cognition and meaning-making rather than physiological origin.

2.2.0.1 Common Factors Across Cultures

Discrete, or “basic,” emotion theories posit a small set of biologically hardwired categories. The lineage begins with Darwin, is developed by Silvan Tomkins, and crystallized in Ekman’s Basic Emotion Theory (BET), which identifies six core emotions—anger, fear, joy, sadness, disgust, surprise—each tied to dedicated neural programs and stereotyped expression patterns [289, 284]. These patterns are presumed to be universally legible, making BET an attractive off-the-shelf label set for emotion-recognition systems [802].

Empirical studies did find above-chance recognition of posed expressions across cultures [287, 582]. But two limitations have persisted. First, recognition accuracy is typically higher within cultural groups than across them, indicating that local “display rules” shape perception. Second, similar expressions may carry divergent meanings. A smile in Tokyo, Lagos, or Toronto may express compliance, irony, or warmth, respectively.

Ekman later conceded some of these complexities. Not all emotions have unique facial markers, and prototypical cues like smiling span multiple categories [285]. Meta-analyses reinforce these concerns. Durán et al. report weak correlations between muscle movements and self-reported affect [274]. Barrett et al. highlight that training datasets often consist of actors’ staged photos labeled by third-party guesses—with little input from first-person reports or diverse cultural samples [92]. This produces brittle models prone to context errors. Barrett dubs this the emotion paradox: we trust what we see, yet the science lacks a ground truth [90].

These concerns do not deny cross-cultural regularities—the wide-eyed startle and nose-wrinkled disgust still recur globally—but they caution against assigning deterministic meanings to expressive cues. Systems trained on categorical mappings risk reverse inference errors: assuming that expression x signals emotion y , regardless of context [463]. Such errors may be tolerable in entertainment but carry serious stakes in domains like hiring or policing.

Discrete models remain appealing for their simplicity and compatibility with classification algorithms. But their limitations—cultural bias, contextual insensitivity, and over-compression of affect—have reenergized interest in dimensional and appraisal accounts that better reflect emotion’s variability and situational logic.

2.2.1 Individual and Social Differences

Challenges to the basic-emotion template come from a broad family of theories that regard emotions as *continuous, context-sensitive, and socially embedded*. Rather than slotting experience into fixed

bins, these accounts emphasize gradation, appraisal, and enculturation—thereby complicating the “common view” that a facial configuration or heartbeat pattern maps cleanly onto a universal label [94, 92].

Dimensional accounts. Building on Wundt’s early triad of pleasure—arousal—tension, contemporary models typically locate emotions in a two- or three-dimensional space: valence (pleasant ↔ unpleasant), arousal (activated ↔ calm), and sometimes dominance or control [744, 761]. This geometry captures blends, shifts, and ambiguous feelings that categorical schemes gloss over. Because points in a space are language-agnostic, dimensional models also travel more easily across cultures and domains where emotion words do not align one-to-one [167]. For affective computing, the approach offers a richer signal—yet still requires contextual interpretation to say *which* worry, thrill, or melancholy a vector actually signifies [383].

Appraisal theories. Appraisal frameworks add the missing layer of *meaning*. Emotions arise, they argue, when an individual evaluates an event’s novelty, goal relevance, agency, or norm compatibility [759]. Such evaluations can be swift and pre-reflective or slow and deliberative, but in either case emotions are intelligent commentaries on “what matters here.” Neo-Stoic and Aristotelian variants—e.g., Nussbaum’s view of emotions as value-laden judgments bound up with flourishing—underscore how culture and moral ideals shape what is appraised, and why [658]. The same racing pulse might be coded as righteous anger, stage fright, or competitive excitement depending on those evaluative frames.

Psychological constructionism. Constructionist accounts push the argument further: bodily fluctuations in valence and arousal become “anger,” “fear,” or “awe” only when interpreted through learned emotion concepts and situational cues [94]. Hence, there is no universal “anger circuit” waiting to be detected; emotions are assembled in the moment from core affect plus conceptual knowledge. This helps explain why similar scowls can signal moral disgust in one setting and playful teasing in another—and why emotion AI trained on surface cues alone often stumbles across cultures [231].

Brunswik’s lens and interactionist views. Brunswik framed emotion perception as a probabilistic inference: observers sample noisy cues and apply their own priors [153]. Interactionist theorists extend the point, arguing that emotion is co-constructed *in situ* through norms, roles, and reciprocal feedback [131]. A furrowed brow means something different in a courtroom than in a comedy club, not because the muscle movement changes, but because the social lens does.

Implications. Taken together, these perspectives do not deny that cross-cultural regularities exist—startle responses, soothing tones, and the like appear with striking frequency [292]. They do contend, however, that the *meaning* of an emotion episode is never fully determined by physiology or expression alone. For systems that aim to sense, classify, or manipulate emotion, this insight is consequential: without attention to appraisal, concept, and context, technical accuracy can diverge from felt reality and, by extension, from dignity-based concerns explored later in this dissertation.

2.2.2 A Unified Theory of Emotions as Evaluative Judgments

Philosopher Martha Nussbaum offers a unified theory of emotion that bridges historical philosophical traditions with contemporary scientific insights across biology, psychology, sociology, and cultural studies. Rather than privileging one disciplinary vantage point, Nussbaum integrates these perspectives to present emotions as universally human experiences—rooted in embodiment and shaped by context—yet also deeply tied to agency, self-determination, and human flourishing.

At the heart of her account is the view that emotions are evaluative judgments: intelligent appraisals of the world in relation to what matters to us. These judgments are not passive reactions but meaningful interpretations—ways of making sense of people, events, and ideas through the lens of our values, beliefs, attachments, and life plans. Fear arises when we perceive a threat to something we hold dear; love reflects the recognition of a person as central to our purpose and identity. While physiological changes may accompany these experiences, Nussbaum insists that what gives an emotion its meaning is not its bodily signature but the evaluative act—the interpretation shaped by personal history, social context, and cultural norms. This richness does not undermine universality; rather, it affirms that the emotional life is both deeply particular and fundamentally human [658]. Indeed, this universality extends beyond humans to other animals, whose emotions reflect their species-specific perceptual and evaluative capacities [658, 661].

Crucially, emotions are not separable from practical reasoning but are integral to it. On this view, emotions guide value judgments, orient moral attention, and animate one’s vision of a flourishing life. They are eudaimonistic—structured by what we take to be good, and indispensable to grasping what is at stake in living well. Emotions thus have not just psychological but normative significance. They enable ethical engagement with others by disclosing their humanity and our own. As such, they carry *intrinsic* moral and political value, not merely instrumental worth. Respecting emotional life becomes a condition of dignity: we owe it to one another not because emotion is a fragile substrate, but because it is the substance of what makes us intelligible as agents striving for meaning.

This is no minor conceptual claim. Nussbaum’s account is a direct challenge to the tradition—across philosophy, psychology, and computing—that has treated emotion as marginal, irrational,

or epistemically unstable. Her theory reframes emotion as a site of intelligence. It is also a *feminist intervention*, confronting longstanding attempts to discredit emotional life as too opaque or politicized to warrant serious normative inquiry.

This marginalization is not accidental. It mirrors the same logic that Warren and Brandeis rejected when they defended the “inviolate personality”—an ideal rooted in Romantic literature that understood emotional and intellectual life as the twin foundations of dignity [876, 739]. Nussbaum retrieves that insight, offering a model that makes emotional experience both analyzable and normatively indispensable.

Moreover, her account takes embodiment seriously. It does not subordinate bodily processes to cognition or treat physical responses as epiphenomenal. Emotions, for Nussbaum, are *embodied appraisals*—interpretive perceptions that reorient the whole self, reshaping how we see and act in the world. They are the *lens through which we encounter reality*, not mere byproducts of cognition. This stance protects her theory against critiques of early cognitive accounts that neglected the role of the body—as noted by Boehner and others in HCI [130]. Nussbaum’s model treats physical and evaluative aspects as mutually constitutive, not hierarchically ordered. She thus transcends the cognitive–bodily divide that has long distorted both empirical science and philosophical interpretation.

Her account is, I argue, the most comprehensive we currently have: one that supports empirical inquiry while grounding the moral and political stakes of emotion, privacy, and agency. For any serious treatment of emotional life in HCI, data governance, or privacy theory, adopting Nussbaum’s framework is not just theoretically justified—it is normatively essential.

2.3 What is Emotion AI?

There are two longstanding challenges in the study of affect: what Prinz calls “the problem of parts” and “the problem of plenty” [701]. The former concerns which aspects of emotion—physiological, phenomenological, behavioral, mental, expressive—are essential to its definition and detection in a given context. The latter refers to the surplus of divergent models and definitions that resist straightforward integration. These conceptual tensions become especially consequential when emotion theories are encoded into computational systems.

Emotion AI refers to a set of technologies designed to detect, classify, and sometimes simulate human emotional states from observable data—facial expressions, vocal tones, body posture, physiological signals, text, and more. Yet these systems rely on partial, reductive models of emotion that raise serious epistemic and normative concerns. Stark argues that because the very nature of emotion remains contested, emotion AI systems risk reifying narrow assumptions—about what emotions are, how they are expressed, and how they ought to be interpreted—into technical

infrastructures that shape decisions and distribute resources [802]. These reductive models encode specific judgments about subjective experience, motivation, and belief. As a result, what counts as valid evidence of emotion, which emotions are considered desirable or deviant, and how they are tied to fairness, transparency, or accountability in AI systems are all shaped by the conceptual scaffolding of the models themselves.

More than a decade earlier, Boehner et al. voiced similar concerns regarding affective computing, the precursor to contemporary emotion AI. While they acknowledged affective computing’s potential to challenge overly rationalist views of emotion, they warned that many implementations still treated emotion as an internal, individual phenomenon—failing to engage the interactional and socio-cultural dimensions that give emotions their meaning [130]. By clinging to a cognitivist paradigm that sees emotion as information to be processed, rather than a situated mode of valuation and coordination, affective computing risked reinforcing the very limitations it sought to overcome.

It would be easy to read Boehner’s warning as a blanket indictment of all cognitivist theories; yet not every cognitivist account collapses emotion into de-contextualized information. Martha Nussbaum’s appraisal framework is a case in point. For Nussbaum, emotions are indeed *judgments*—but they are *embodied, historically layered, and culturally meaningful* judgments that reveal what a person values in that very situation [659]. The bodily surge of fear, the culturally tutored script for anger, and the narrative self that interprets both are inseparable parts of a single evaluative act. In other words, cognition here is not cold computation; it is a socially embedded way of seeing the world through a felt lens. Precisely because of this integration, Nussbaum avoids the reduction Boehner critiques: her theory treats emotion neither as disembodied signal nor as timeless universal, but as lived appraisal whose content and form are co-produced by biology, culture, and personal history. That richer conception provides a firmer normative baseline for judging what emotion-AI systems may legitimately infer, expose, or manipulate—and what would constitute a violation of dignity.

2.3.1 System Model

Emotion AI can be segmented into three main areas or “building blocks”: automatic emotion recognition, artificial emotion augmentation, and artificial emotion generation [765]. Most commercial systems focus on recognizing human emotion or augmenting machine responses to it, with emotion generation occupying a smaller but aspirational role in the broader trajectory toward affective artificial general intelligence (AGI). Figure 2.1 summarizes these components, adapted from Schuller and Schuller [765].

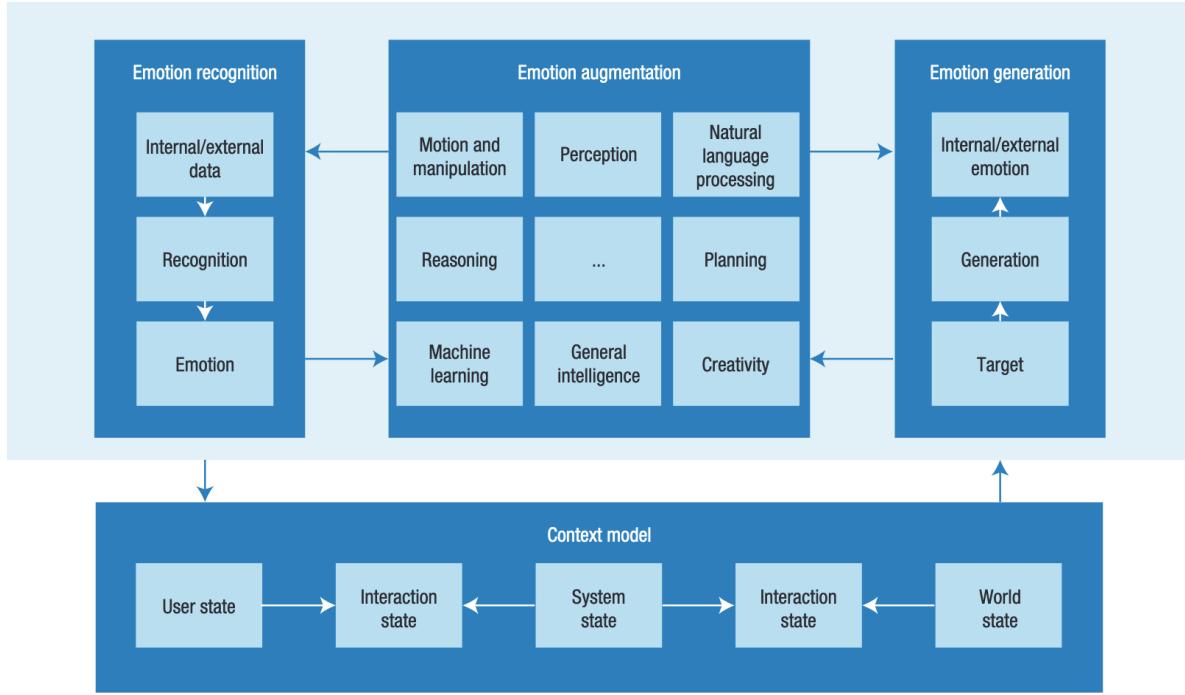


Figure 2.1: Emotion AI Building Blocks (Schuller and Schuller, 2018)

Emotion Recognition

Emotion recognition systems analyze input signals to infer an individual’s emotional state, typically using supervised learning approaches. Common inputs include visual, auditory, textual, and physiological data: facial expressions, voice tone and prosody, body posture and movement, gaze patterns, written language, and biosignals such as heart rate variability or skin conductance [910]. Increasingly, these are combined with contextual and behavioral data such as location history, device usage, social media activity, or user demographics to augment emotion inference models [919, 234, 910].

Inferences are typically modeled using either discrete emotion categories (e.g., anger, joy, fear) or dimensional representations (e.g., valence-arousal space). Input modality often influences model choice—facial expression data tends to be used with discrete models, whereas biosignals are more commonly analyzed using dimensional ones [796]. In either case, the process depends on mapping externally visible or measurable features onto internal affective states, via assumptions embedded in emotion theory [621].

Earlier systems relied on modular machine learning pipelines (e.g., SVMs, HMMs, or shallow neural nets), with separate stages for pre-processing, feature extraction, and classification [765]. More recent approaches favor deep end-to-end learning, which enables raw multimodal data to be processed directly by a unified architecture that learns relevant features internally [849]. This

shift allows for multimodal emotion recognition, which combines inputs such as facial expression and voice for improved prediction accuracy [477, 849]. Deep learning also circumvents earlier limitations related to sparse or low-quality annotations [396], though it demands significantly larger and more diverse datasets to avoid overfitting and bias [726].

Notably, the theoretical model assumed by the system governs what emotions are recognized, how they are defined, and what counts as valid evidence. As discussed above, appraisal theories like Nussbaum’s avoid the reductive tendencies critiqued by Boehner et al. by recognizing emotions as embodied, socially-situated judgments. Yet most systems still treat emotion as a discrete label or latent vector—abstracted from the self-understandings or value commitments that give them meaning. This conceptual gap underscores the risks of normative overreach when such systems assign emotional states without a theory of personhood or context.

Emotion Augmentation

Emotion augmentation systems go beyond recognition by incorporating emotional information into system behavior. They synthesize or simulate emotion in interaction, adapting outputs to enhance user experience, persuasion, learning, or regulation. This often involves architectural modules for perception, planning, reasoning, and decision-making, building on affective state estimates from the recognition pipeline [690, 765].

Augmentation tasks may aim to *express* emotion in machines (e.g., empathic chatbots), or to *influence* emotion in users (e.g., calming a distressed caller). Many rely on rule-based decision models for state transitions or dialogue management [435, 682]. Others incorporate machine learning, including reinforcement learning, to optimize adaptation over time [671].

Dimensional models are especially useful here due to their continuous representation of affective states and ability to handle blends and ambiguity [383]. Hybrid systems combining discrete and dimensional features are also common. As with recognition, deep learning is increasingly applied in end-to-end augmentation systems, folding expression and adaptation into a unified pipeline [488].

Emotion augmentation also plays a role in broader AI/ML systems that model user motivation, social behavior, or decision-making. These systems incorporate affect not merely as output but as a factor in learning, planning, or optimization—for example, by integrating emotional variables like anxiety or confidence into backpropagation algorithms to enhance adaptability [471, 620, 245].

However, most augmentation tasks stop short of generating authentic affective states. Their aim is not to induce emotions in the machine itself but to simulate affective behavior or influence human users. This distinction is critical for assessing both functional limits and normative implications.

Emotion Generation

The most speculative building block, emotion generation, seeks to endow machines with internally coherent emotional states—not just to simulate emotion, but to *possess* it. This includes generating both outward expression and inner experience, akin to emotional sentience [765, 171]. Though often conflated with augmentation, generation marks a stronger claim: that emotion can be instantiated, not merely mimicked.

To date, most generation models remain rule-based and task-specific [765]. However, advances in generative deep learning are reviving research in this area [171]. Still, generation is rare in deployed systems and raises fundamental debates about whether machines can—or should—experience emotions at all [691].

The aspiration toward emotionally intelligent AGI, present since early AI research [642, 586], remains largely theoretical. As such, this dissertation does not treat emotion generation in depth, but acknowledges its role as a longer-term trajectory in emotion AI’s evolution—and as a frontier with profound ethical stakes.

2.3.2 Theoretical Assumptions

Most emotion AI systems adopt one of two theoretical paradigms: *discrete* emotion models, which posit a biologically grounded set of universal categories, and *dimensional* models, which represent emotion as positions in a continuous affective space. These foundational assumptions shape every aspect of model development: which emotional signals are treated as valid input, how emotional states are computationally represented, and what normative claims are embedded in outputs [166]. While discrete models have historically dominated both academic and commercial research—due to their simplicity, intuitive appeal, and compatibility with observable signals like facial expressions—they are increasingly critiqued for their theoretical limitations and lack of cross-cultural validity [802, 765, 792].

By contrast, dimensional models, which allow for emotional complexity and blended affective states, are gaining favor in the research community [792, 645]. These approaches model emotions along continuous scales (e.g., arousal, valence, dominance) and are often better suited to physiological and multi-modal input data. Yet their increased complexity—both computational and interpretive—can make them less attractive for practical deployment [463, 460]. Some researchers adopt hybrid approaches that attempt to bridge this divide, such as by jointly modeling categorical and dimensional outputs [915], or by transforming discrete categories into coordinates in affective space [765].

A pragmatic stance has emerged among some developers: that discrete emotion categories may still be computationally useful even if their theoretical assumptions (e.g., emotion universality) are

not fully endorsed [690, 463]. According to this view, bodily expressions and vocal cues can still convey information relevant for inferences—even if the system’s outputs are coarse or probabilistic approximations. However, this instrumental framing risks concealing the underlying limitations of emotion recognition, particularly in commercial settings where system assumptions, modeling choices, and performance constraints are rarely made transparent to users or affected populations.

Some emotion AI models aim to escape this theoretical bind altogether by rejecting narrow definitions of emotion in favor of more pluralistic or interactionist accounts. For example, hybrid models like Cambria et al.’s *Hourglass of Emotions* combine multiple theoretical perspectives: emotions are modeled as states along four affective dimensions (sensitivity, aptitude, pleasantness, attention), which are further categorized into hierarchical clusters of twenty-four discrete emotions, including compound and secondary emotions [167]. This approach allows for non-linear emotional relationships and co-occurring affective states, better reflecting human experience while retaining structural clarity.

Still, even sophisticated hybrid models face challenges. Their multidimensional representations can be difficult to interpret without culturally specific grounding [223], and they often lack standardized metrics for validating outputs across populations. Such difficulties reflect deeper issues in affective science: semantic ambiguities in emotion language [870], lack of taxonomic consensus [810], and enduring disagreements over the function and structure of emotion itself. These epistemic challenges make it unlikely that emotion AI can ever offer purely “objective” or universal inferences across settings and populations.

Yet this is not an argument for abandoning algorithmic modeling of emotion. On the contrary, computational implementations can serve as tools for conceptual clarification. As Paiva et al. and Marsella et al. argue, realizing a theory in code can force tacit assumptions into the open, make psychological models testable, and generate artifacts that permit empirical scrutiny [567, 568]. Simulation environments have already contributed to new insights about the role of affect in attention, memory, motivation, and behavior [142, 263, 145, 373, 64]. These epistemic gains, however, do not erase the ethical risks introduced when such systems are deployed in real-world, high-stakes contexts—especially when their theoretical limitations are hidden behind persuasive narratives of objectivity or personalization.

Ultimately, if emotion AI is to be integrated into social systems, its theoretical assumptions must be treated not as technical design decisions but as normative claims. Whether a model treats emotions as universal categories, as socially constructed dimensions, or as hybrid structures implicates judgments about what emotions *are*, what people *feel*, and what constitutes legitimate inference about another’s internal state. These are not only scientific matters—they are matters of justice, dignity, and human understanding.

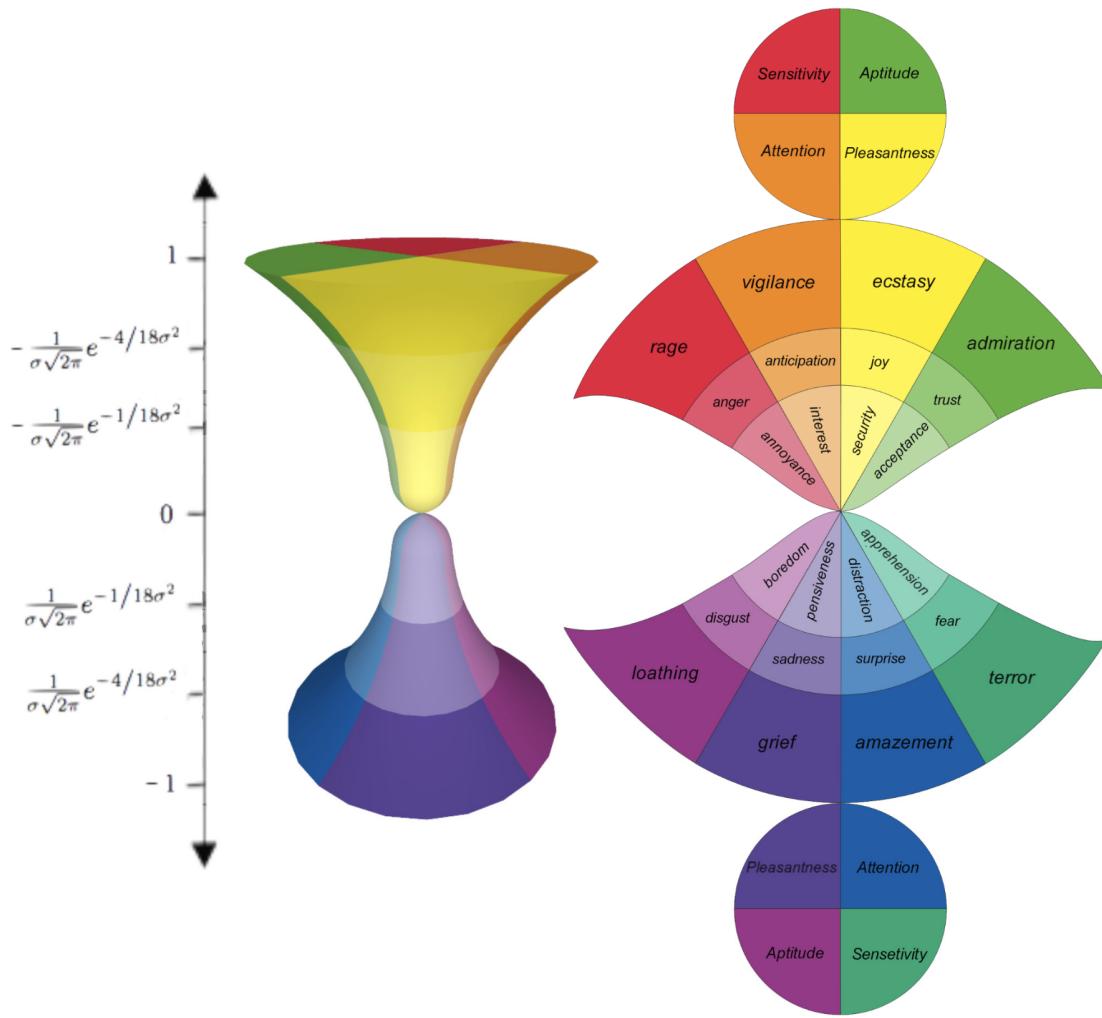


Figure 2.2: 3D model and net of the Hourglass of Emotions model (Cambria et al., 2012)

2.3.3 Emotion Labels

The emotion labels used in emotion AI systems typically align with the system's underlying theoretical model—discrete, dimensional, or hybrid. In discrete approaches, labels often draw from basic emotion theories, such as Ekman's six universal emotions, and may involve hierarchical distinctions (e.g., positive, neutral, negative), or attributes such as intensity (fixed or graded) and purity (pure or mixed) [383]. While these latter variables do not strictly adhere to discrete theory's assumption of mutual exclusivity, they are routinely incorporated in practice, reflecting a pragmatic loosening of theoretical commitments [383, 166, 463].

Dimensional approaches typically model arousal and valence, with some incorporating a third dimension such as dominance or power [383]. Because dimensional models represent affective

states as coordinates in a continuous space, they inherently support the classification of mixed, blended, or evolving emotions. In practice, these models can also be extended to capture contextual and environmental variables [863].

Regardless of model type, emotion labels are derived from training data annotated by humans. These annotations are based on either (1) the self-reported emotions of participants, or (2) observer perceptions of the emotion expressed by others. Often both are collected and compared, though system development tends to prioritize the latter. For example, Ekman’s foundational studies used both self-reports and observer ratings in response to emotion-inducing film stimuli [288]. Annotation tools vary by model type: ELAN and ANVIL are commonly used for discrete labels [69, 480], while tools like ANNEMO are tailored to dimensional data [724].

Yet discrepancies between perceived and self-reported emotions are well-documented [162], presenting a core challenge to the validity of emotion AI. Mismatches between observer ratings and actual emotional experience limit the reliability of inferred labels across both discrete and dimensional approaches [463]. To address this, various strategies have been proposed: averaging ratings across multiple observers [843]; incorporating multimodal inputs [161]; using feature learning or multi-task learning to jointly model perceived and self-reported emotion [915]; or improving temporal modeling by identifying spikes, baselines, and transitions in affect [463]. Some have emphasized the need for more inclusive training datasets, especially incorporating self-reports from autistic individuals and culturally diverse populations [511, 491]. Yet the effectiveness of these methods varies across emotion types, labeling schemas, and demographic subgroups [915, 843, 491].

Emotion labels can also distinguish between *felt* emotions—those that arise spontaneously—and *expressed* emotions—those that are deliberately performed for interpersonal or social reasons [283]. For instance, a person may display a smile to meet social expectations (a “Duchenne smile”), even if not genuinely experiencing joy [337, 687, 498]. Studies show that human observers can often distinguish between felt and expressed emotion [634, 498, 817], and emotion AI research increasingly aims to do the same [463]. Approaches to this include temporal modeling [203, 762, 857, 264, 419], empirical studies of divergence between internal state and external display [218, 291, 274, 92], and theoretical arguments against treating self-reports as ground truth [463]. The latter suggest reframing labels not as veridical signals of internal states but as behaviorally meaningful cues—especially in “in-the-wild” contexts such as social media, where individuals are not involved in the inference process [463].

Additional layers of differentiation can include degrees of prototypicality (i.e., how well the expression conforms to stereotypical features of a given emotion), levels of regulation (e.g., suppressed, exaggerated), or whether the expression is acted or spontaneous [765].

Emotion AI remains widely commercialized. The gap between the system’s capabilities and

its marketed claims often stems from a deeper ambiguity about its epistemic target. Most current models are better understood as inferring *perceived* emotion—that is, how a target’s expression is interpreted by observers—rather than directly detecting what a person truly feels [92, 409]. Nonetheless, many commercial systems, and the broader aspiration of emotion AI itself, remain committed to closing this gap: to infer not only what people express, but what they mean to express—and ultimately, what they actually feel [162].

2.3.4 Algorithmic Goals

The goals of emotion AI algorithms are determined by both the model’s function and its underlying theoretical assumptions—reflected in the type of emotion labels used and the kinds of outputs they are optimized to produce.

For systems focused on automatic emotion recognition [765], goals typically fall into three categories: (1) to infer the stimulus that elicited an emotional response (e.g., music [278], opinions [168, 298]); (2) to predict how an observer might perceive a person’s emotional expression (e.g., facial expressions labeled by third-party raters [92]); and (3) to approximate the emotional state of the person themselves (e.g., models trained on self-reports [409] or designed to distinguish “genuine” from “fake” emotions [437]). These goals may be pursued using discrete, dimensional, or hybrid approaches [765].

However, emotion AI does not directly measure emotions. Rather, it infers them indirectly from observable signals—speech, facial expressions, physiological cues—mapped onto pre-labeled datasets. In most commercial and academic applications, these labels are derived from observer judgments, not ground-truth internal states [92]. This introduces a core epistemic limitation: most emotion AI systems are effectively trained to replicate social perceptions of emotion rather than capture felt experience [93]. Some proposed methods aim to bridge this gap by incorporating self-reports or human-in-the-loop feedback during system interaction [465, 906], though such methods remain rare in commercial systems.

Emotion AI technologies, when deployed, pursue broader contextual goals. For instance, conversational agents and workplace tools may be used to modulate the emotional states of workers, and been shown to influence affective responses in users [479, 465]. Such manipulation can have both moral and political consequences: Facebook’s emotional contagion experiment demonstrated how algorithmic control over emotional content in News Feeds could influence users’ moods without their knowledge or consent [495].

These affective interactions are often justified as usability goals: increasing trust, engagement, or believability [435, 516, 253]. More ambitiously, emotion simulation is pursued as a pathway toward artificial general intelligence (AGI), under the assumption that affective reasoning is a prerequisite

for adaptive, human-like intelligence [765]. Emotional intelligence in this context includes detecting emotions in oneself and others, adapting emotional responses, and using emotional information for decision-making and goal pursuit.

Yet even as emotion AI systems pursue increasingly sophisticated goals, including aligning inferred emotion with stimuli, perception, or felt experience, these goals often outpace the ecological validity of the systems themselves. Most models are trained on constrained, lab-based datasets that reflect limited demographic, cultural, and situational diversity [910]. As a result, real-world interpretability and validity remain uncertain: the same emotion label may signal fear, excitement, or disinterest depending on context, with no systemically reliable means of disambiguation.

2.3.5 Emerging Systems

Emerging architectures such as those using transfer learning or multi-task optimization blur the line between emotion recognition, personalization, and behavioral modulation—shaping future system behavior across tasks and domains. Trained on one type of data and then applied to another (e.g., tailored responses, content recommendation), emotion recognition feeds directly into adaptive system design—modulating experiences, steering user engagement, and optimizing predicted outcomes across applications [917, 36].

Unlike earlier models that relied on discrete labels or explicit dimensional coordinates (e.g., valence–arousal–dominance), these newer systems may forgo interpretable emotion spaces altogether. Instead, they extract latent affective representations from high-dimensional data [918]—producing downstream effects without requiring human-readable emotion categories. This shift makes it more difficult to track what is being inferred, manipulated, or optimized—intensifying epistemic and normative concerns.

With these conceptual foundations in place, the following chapters turn to empirical investigation. Building on the technical and theoretical groundwork developed thus far, Parts II and III center the lived perspectives of those most affected by emotion AI: data subjects. Through interviews, content analysis, and a mixed-methods factorial vignette study, I examine how emotion AI is designed, promoted, and perceived in real-world domains—surfacing the ethical and privacy implications these systems pose as they move from laboratories into social life.

Part II: The Empirical Case for Emotional Privacy

A recurring theme across the studies presented in Part II—from participants’ conceptualizations of and attitudes toward automated wellbeing interventions on social media in Chapter 3, to workers’ perceptions of and experiences with emotion AI in the workplace in Chapter 5—is concern over how emotion AI, whether deployed in institutional (i.e., workplace) or consumer (i.e., social media) settings, enables the commodification of emotion by exploiting human affect for corporate gain. These studies highlight a persistent worry among data subjects that the unchecked sensing and circulation of emotional information serves corporate interests at the expense of their own wellbeing.

In Chapter 3, interviews with social media users reveal predominantly negative attitudes toward AI-enabled well-being interventions. Participants express deep skepticism about whether automated emotion recognition can genuinely support users’ health or psychological safety, or whether such interventions ultimately manipulate emotional states for platform optimization. At the heart of this skepticism lies a concern not about technical adequacy, but moral legitimacy. Participants emphasized the irreducibility of human qualities including morality, lived expertise, and shared humanity in contexts of care. These qualities, they doubted, could be adequately replicated by machines—rendering automated emotional interventions not only as insufficient, but ethically misaligned with the forms of emotional reciprocity and support that people would welcome.

Chapter 5 turns to the workplace, where interviews with workers suggest that emotion AI functions as a tool of emotional surveillance—exposing workers’ inner lives to managerial oversight in contexts where consent is often experienced as coercive. In these accounts, privacy intrusions are essential functions of the system, viewed as primary drivers of harm. Participants describe the erosion of emotional autonomy, intensified affective labor, and the suppression of dissent—particularly among workers already marginalized by gender, race, or role. These findings underscore the need to recognize *emotional privacy* as a distinct and foundational privacy interest—one that demands robust and targeted protections commensurate with the stakes introduced by emotion AI.

Chapter 4 reinforces these concerns through a content analysis of promotional materials from emotion AI hiring services, revealing an engineering and design logic that commodifies workers’ emotional lives in service of organizational control. Vendors market emotional conformity—achieved through emotion AI-driven pre-employment screening—as a means to automatically weed out hiring “misfits” and cultivate a workforce that “lives and breathes” corporate values. By translating affective metrics into hiring filters, these systems do more than measure emotion—they seek to recalibrate workers’ inner lives to align with corporate and managerial ideals. My analysis

shows that, once folded into labor management pipelines, emotion AI crystallizes corporate affective preferences into automated decision-making systems, institutionalizing new forms of exploitation and stratification. A worker’s prospects no longer hinge solely on qualifications or demonstrated soft skills, but on the ability to render enthusiasm, authenticity, motivation, docility, and cultural fit in ways that are legible to opaque algorithmic infrastructures.

Echoing Elizabeth Anderson’s critique of the “tyranny” of the U.S. workplace [58], the findings in Part II illustrate how emotion AI, when deployed in coercive environments, extends employers’ unchecked power to shape stratifying social norms, bypass meaningful consent, and restrict social participation—reinforcing structural asymmetries and narrowing agency over personal emotional lives. Such risks remain insufficiently recognized, addressed, or remediated by existing privacy frameworks and harms taxonomies.

Together, these findings raise critical moral and political questions about the limits of emotion AI as a solution to social problems, its role in reinforcing institutional power, and the ethical boundaries of its deployment. By establishing emotional privacy as a distinct interest, Part II highlights the need for a robust normative framework capable of evaluating the unique challenges posed by emotion AI across diverse socio-technical contexts—technologies that offer both extraordinary promises but also carry the potential for profound harm.

This section connects a range of ethical concerns associated with emotion AI—including meaningful consent, algorithmic bias, surveillance and disciplinary technologies, information asymmetries, power imbalances, and labor commodification—to the underlying flows of emotional information. In doing so, it sets the stage for understanding emotional privacy as a critical lens for identifying and mitigating harmful data practices that expose, instrumentalize, and reshape the experience of human emotion.

CHAPTER 3

Emotion AI on Social Media: Data Subjects’ Conceptualizations of and Attitudes Toward Automatic Wellbeing Interventions¹

3.1 Introduction

Social media platforms provide distinctive contexts for people to share personal and emotional content, while also being personally and emotionally affected by interactions mediated on these platforms [776]. Social media companies have been implicitly or explicitly interested in emotions. For example, to better understand their users’ emotions and whether and how they could shape them, in 2014, Facebook researchers conducted a large-scale experimental study to examine whether “emotional states could be transferred to others via emotional contagion” [495]. The public backlash to this study of emotion manipulation was widespread and severe [606, 769, 433]. Bottom-up criticism derived from news article commentators has demonstrated that the public had a variety of concerns about Facebook modifying their News Feed for emotional content and analyzing their subsequent engagement with the platform to infer its emotional impact, including concerns about being manipulated, being subject to research without consent, violation of expected use of data, and lack of trust in Facebook generally [391].

Despite negative public sentiment regarding technology companies engaging in the manipulation of and making inferences about an individual’s emotion, as evidenced perhaps notably by Facebook’s emotional contagion study, researchers and technology companies continue to deepen and expand the application of the growing emotion AI market. As costs for computing power continue to decline while advances in computational power rise [5], and sharing of personal and

¹This chapter is based on: Kat Roemmich and Nazanin Andalibi. 2021. Data Subjects’ Conceptualizations of and Attitudes Toward Automatic Emotion Recognition-Enabled Wellbeing Interventions on Social Media. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 308 (October 2021), 34 pages. <https://doi.org/10.1145/3476049>. This material is based upon work supported by the National Science Foundation under Grant No. 2020872.

revealing information on social media continues to grow [685], development of emotion recognition applications on social media have spread across public and private sectors. Computationally inferred emotions can be processed to make inferences and predictions about individual behaviors, medical and mental health conditions, and emotional states [360]. There has been increasing enthusiasm for population scale research and monitoring, particularly in the medical and computational social science fields [555, 787, 839, 38, 214]. Academic researchers are particularly interested in harnessing emotion data for public health purposes, while corporations use it to gauge opinion about their products and consumer preferences, and governments find it useful to understand public sentiment and assess security risks, to name a few [190, 608].

In the US, medical and mental health data is protected by federal and state legislation, but such digital inferences fall outside the scope of these protections [566]. As such, inferences of a person's emotions can be packaged and sold to insurance companies, advertisers, and other interested parties, often without that person's knowledge or consent [360, 177]. Emotion inferences are of special interest to psychiatry and psychology—fields that have traditionally relied on self-reported patient surveys to diagnose conditions—as an intervention tool with potential to increase accuracy in patient diagnosis, detect conditions early, and identify people in moments of crisis or relapse [434, 118, 715].

Social media platforms in particular are a uniquely rich source of emotional data, as individuals use these sites to disclose and disseminate sensitive information such as personal experiences and media content [50], as well as receive social support from friends within their network [54] with the potential to improve wellbeing [389].

Automated wellbeing interventions on social media have been criticized in both academic scholarship and opinion pieces for the harm they present to individual autonomy, individual privacy, and individual safety [566, 786, 360, 111, 178, 211]. However, the critical discourse surrounding this technology has grossly omitted engagement with its data subjects. Social media users who generate emotional content on which automated wellbeing interventions are trained upon, and who may be subjects of those technologies, are stakeholders in the development and delivery of emotion AI. This study centers these data subjects, seeking to understand their technology conceptualizations and attitudes as an exploratory step towards understanding the human impact of the surveillance, datafication, and commodification of human emotion—knowledge crucial to any evaluation of the ethical and responsible use of this emerging technology.

We conducted semi-structured interviews with 13 adult social media users in the United States who have experienced both positive and negative meaningful personal experiences in the past year and who reported having shared about them on social media. Overall, we found that participants had negative conceptualizations of and attitudes toward wellbeing-related interventions on social media, but in a minority of cases held more positive attitudes when interventions targeted *other*

individuals, rather than themselves. We first develop an understanding of why people held negative attitudes toward automatic wellbeing interventions. Negative attitudes were rooted in the way people compared traditional delivery methods of wellbeing interventions, by humans, to their conceptualizations of algorithmically-enabled wellbeing interventions, by AI. We identified four attributes that participants remarked were essential to supportive wellbeing interventions: 1) helpfulness and authentic care; 2); personal and professional expertise; 3) morality; and 4) benevolence through shared humanity. Participants felt that these attributes could only be held by humans, expressing doubt that artificial intelligence (AI) could hold them. These comparisons between conceptualizations of automatic wellbeing interventions and human-delivered wellbeing interventions shaped attitudes toward automatic wellbeing interventions in a negative way.

We describe the tension between participants' negative conceptualizations of and attitudes toward automatic wellbeing interventions broadly, and how some participants imagined possible positive social benefits when realizing its potential, assumed impact on *others*. These participants conceived of automatic wellbeing interventions as a potential social good that could support academic research, increase access to wellbeing, and prevent egregious harm. Yet most participants maintained their negative attitudes toward automatic wellbeing interventions when conceptualizing its potential impact on others, revealing worries about the potential harm (e.g., re-traumatization, spread of inaccurate health information, inappropriate surveillance, interventions informed by inaccurate predictions) that automatic wellbeing interventions could cause others. Participants emphasized a requirement for individual and external control to potentially mitigate these harms. These observations highlight the importance of including data subjects in emerging technologies' development rather than conceiving of data subjects as *others* and assuming what their wellbeing entails. Finally, we discuss qualities in either the development or delivery of the data subjects' conceptualized intervention upon which data subjects' attitudes were dependent. These qualities include: 1) accuracy; 2) contextual sensitivity; and 3) positive outcome.

We then situate these findings within the discourse surrounding policy and ethical implications of automatic wellbeing interventions. We argue that in the current practical and regulatory landscape in the US, automatic wellbeing interventions are incompatible with ethical and socially responsible AI applications. Further, we express concern for current state social media intervention processes that include police intervention for mental health crises, especially for racial/ethnic minority populations. We speculate (and critique) that instead of honoring the concerns of data subjects, the entities that employ emotion surveillance technologies have focused on promoting a rhetoric of “safety through surveillance,” and shaping social norms to accept ubiquitous surveillance under the guise of public safety.

The constant monitoring of data subjects' emotions carries tremendous risk in its power to shape human behavior. By developing automated and predictive systems built on normative assumptions

of human emotion, new realities are built around the expectations and anticipations of their outcome while perpetuating stigma around emotion and mental illness [678]. As warned by Couldry and Mejias, “the constant watchability of our every thought and action by external forces changes the field of power in which we exist, transforming a supposed order of individuals into a collection of living entities plugged into an external system” [222]. In other words, even when and if individual people are not subject to the gaze of surveillance capitalism, its practice holds implications for all individuals as part of a larger system [222]. The normative weight of the models used to develop emotion AI technologies, what data is collected to feed those models, and what people (and society) believe about the inferences and predictions of the “interiority, judgments, and potential future actions of human beings” matters if we are to understand the ethical and social implications of emotion AI [802].

3.2 Background and Related Work

3.2.1 Emotion and Risk Profiling through Social Media Traces

Social media in particular has shown promise as platforms from which to harvest emotion-intensive data [178, 37, 249, 252, 483, 561, 715], and infer mental health conditions such as schizophrenia [123, 613], depression [251, 845, 214, 715], post-partum depression [250], or post-traumatic stress disorder [215]. Researchers across disciplines, from medicine to computing, consider the potential benefit of computationally inferring emotions to promote public health [305] by way of early diagnosis of illness [746], sentiment detection and behavior surveillance [395, 627], and real-time intervention [822, 190, 625, 364].

Due to the intimate way in which many people use social media, social media is considered both as a suitable source from which to infer emotions, and as an ideal platform to *intervene* based on those inferences of emotion. Opinion pieces regarding automatic interventions have offered mixed support and criticism, with some questioning its ethical and privacy implications while others laud the interventions’ support of suicide prevention efforts [636, 88, 644, 602]. Perhaps the most prevalent example of an automatic wellbeing intervention is Facebook’s suicide prevention intervention, which uses a combination of n-gram based linear regression and DeepText-based neural network models to flag users at risk of imminent harm, and intervenes by suggesting the contact number of the National Suicide Prevention Lifeline and offering the ability to chat with a crisis worker; the case is also sent for review by a human reviewer, who then decides if the company will involve police for a welfare check [280, 364, 115].

While prior work has examined peoples’ attitudes toward the development, use, and implications of emotion AI or similar approaches on social media generally [311, 313, 180], relatively little work

has explored peoples' attitudes and conceptualizations of the application of emotion recognition to develop, implement, and deliver automatic wellbeing interventions—a key application of emotion inference technologies.

3.2.2 Human Values and Ethical Stakes

The growing interest and development of algorithmically inferred emotions and associated interventions has raised new ethical questions and considerations in the areas of privacy, harm to vulnerable populations, transparency, and fairness. Echoing past work that has shown predominantly negative data subjects' attitudes toward emotion recognition broadly, finding that people perceive algorithmic inferences of emotion as invasive and intrusive [52, 333], scholars from many disciplines including law, computing, philosophy, and psychiatry have sounded the alarm on the potential of targeted wellbeing interventions to infringe on individual privacy and autonomy [111, 178, 211, 566, 786, 360]. Additionally, scholars have warned of emotion AI's potential to increase harm to vulnerable mental health patients through amplification of mental health bias and potential misuse of data [178, 226, 542, 566], express concern about emotion AI's lack of algorithmic transparency [88, 115, 178, 226, 357, 387, 482], and raise doubt in emotion AI's algorithmic fairness [61, 88, 178, 357, 387] and testing practices [178, 177].

There has been little scholarly engagement with the data subjects whose data enable, and who are potentially affected by, automatic wellbeing interventions interventions. Ford et al.'s survey of user perceptions toward Facebook using emotion data to provide targeted mental health advertising found that participants were not comfortable with their Facebook posts being analyzed for targeted advertising by algorithms, and even more uncomfortable with their posts being analyzed by human reviewers [333]. Studying the context of digital phenotyping by technology companies broadly, Costello and Floegel found that individuals with mental illness were wary of automated assessments of mental health and mood-tracking applications. The participants in their study were concerned about the profit motives behind such applications, and were distrustful that their personal data would be used responsibly [221]. These studies highlight an overall public distrust of social media companies collecting, processing, and sharing sensitive information such as emotion data.

3.3 Methods

Recruitment

We conducted in-depth, semi-structured interviews ($n=13$) lasting between 77 to 120 minutes (average=106 minutes) with adult social media users in the US. We recruited participants via

a screening survey and conducted interviews over voice and/or video call. We transcribed the interviews for analysis. We shared calls for participation via personal social media, personal networks, and Craigslist. We chose Detroit and Houston Craigslist pages in an effort to achieve a diverse participation pool, in consideration of these cities' high racial/ethnic minority populations [10]. In three cases, the interview participant was acquainted with the interviewer. To preserve the integrity of the data, another researcher on the team conducted the interviews in those three cases. Participants received a \$30 honorarium. This study was approved by our institution's IRB.

Participation

Out of 100 responses to the screening survey, we invited 20 to participate in the interviews. Survey respondents who did not meet the minimum criteria (based on age, location, and behavior) did not proceed to the next step of the survey. Out of 20 invited to interview, 13 signed a consent form, scheduled, and appeared for the interview. Survey questions included inquiries regarding social media usage, such as whether they had shared positive and negative personal experiences on social media in the past year. Decisions to invite respondents to the interviews were conducted in an iterative manner and partly made based on the identities and experiences represented by the data collected by that point in time. We aimed to interview people who had both positive and negative emotional experiences, and shared about them in some form on social media, due to our study's goal of capturing conceptualizations of and attitudes toward emotion inferences based on real experiences and posting behavior, and our focus on emotions. These real experiences provided a basis for our participants to draw from when probed for scenarios designed to elicit their values and imaginaries regarding automatic wellbeing interventions on social media.

In addition, our goal also included capturing a range of identities (e.g., race/ethnicity, age, gender) and experiences. Examples of positive experiences represented included career accomplishments, educational attainment, and home ownership. Examples of negative experiences represented included job loss, health concerns, and relationship complications. Our study's racial/ethnic makeup included one Indian, two Asian, two Black, and eight white participants. Ages of participants ranged from 22 to 58, with an average age of 32.4. Gender identifications included nine women, one man, one gender-fluid, one agender, and one genderqueer. Education completed included five participants with college degrees, six with graduate degrees, one with some high school, and one with some college. Eleven out of thirteen participants used Facebook regularly. Other social media used included Facebook groups, Instagram, LinkedIn, Twitter, Tumblr, AO3, Reddit, Snapchat, Twitch, YouTube, and Discord.

Study Design

3.3.0.1 Interviews

We followed a semi-structured protocol when conducting interviews to allow for exploration and flexibility. Interviews started by asking participants about their social media use, social media sharing behaviors (particularly in regard to meaningful and emotional experiences), understanding of what happens to such data when shared, and expectations for privacy in those contexts. To facilitate recall, we probed interview participants with what they had shared with us in the survey when needed (e.g., “you had mentioned. . . ”) and encouraged them to refer to their posts as we spoke if they wanted. By eliciting recall of specific experiences in the first phase of the interview, we were able to better understand how participants used social media to share emotional and personal experiences, and positioned participants in a context of emotion-situated experiences when exploring scenarios in the next phase. We did not observe any struggle with participants recalling experiences.

3.3.0.2 Scenarios

The next phase involved using speculative scenarios to elicit values, concerns, and attitudes toward emotion recognition on social media. Participants were allowed flexibility as to which experience they wanted to discuss during the scenarios.

Scenarios have been used in prior HCI and CSCW work. Though a complete review is outside our scope, we emphasize our methodological choice to use speculative scenarios due to its helpful application eliciting values toward technologies (and especially emerging technologies) [49, 152, 172, 411, 895] in cases where people may not be familiar with the technology or topic being examined [318].

To probe for peoples’ values and conceptualizations of automatic wellbeing interventions, we first probed for values related to entities making inferences or predictions based on emotional content shared on social media, and then probed to ask how they might feel if those inferences or predictions were used to offer “wellbeing support or help them feel better.” We kept this question broad, because our goal was to understand what participants would imagine these interventions to be like, or what examples from their experiences they might share with us. Despite critique of using scenarios based on the presumption that what people will do is different than what they say they will do when imagining the scenario, past work has shown that people tend to respond similarly to scenarios as they would in real life in emotional contexts, such as those used in our study [436]. Further, our goal was to surface participants’ concerns and attitudes, for which hypothetical scenarios are useful tools [895].

Our use of scenarios is informed by prior work in algorithmic folk theory and privacy, which

suggests that understanding what people *think* technology (and algorithms) can do or already do, is just as important as understanding how the technology operates in practice [297, 894]. Building on these works, [297, 894, 51], our study centers data subjects’ imaginaries of algorithms and uses hypothetical scenarios to probe for human values including and beyond data privacy. Our focus in this work is data subjects’ conceptualizations of and attitudes toward automatic wellbeing interventions, and the factors we found to shape those attitudes—not all the collected data.

In using scenarios, participants were asked to consider the positive and negative experiences they had posted about in the past year (an inclusion criterion for the study), then imagining how they would feel “if the social media site on which they shared their experience had used computational methods to infer their emotions, either at the time or after their posting.” We then asked participants questions regarding their attitudes and values toward these emotion inferences, based on their social media data. Additionally, we asked participants to consider two specific applications of inferences made from their emotion data: advertisements (not our focus here) and wellbeing-related interventions. Questions asked in these contexts were used to determine factors that shaped participant attitudes.

Scenarios were presented to participants via a link to a Google document. We randomized the order in which scenarios were presented to participants. The document included the following text, once for positive and once for negative emotional experiences, as determined by the participants themselves:

‘I would like you to think about something [positive/negative and personal] that brought out [positive/negative] emotions for you. Maybe the experiences we talked about earlier. Now consider this scenario: You had shared on [insert social media they use most] about that, and had explicitly shared how you felt about it. Everyone reading it would have been able to understand what your experience was and how you felt, there was no ambiguity. Now imagine that [insert social media they posted on] used computational methods to detect what emotions you felt at the time of posting that.’

We began with one experience, and if appropriate and time permitting, we asked “*How about if this was related to your other experience?*” Participants were asked to share which experiences they thought about in relation to the scenarios so as to provide context and establish what emotional connections they made. For all cases, these included the emotional experiences participants shared in the screening survey and in the initial phase of the interview; sometimes participants brought up new topics. Once the emotional context of participants’ imagining of the scenario was established, we probed to elicit their attitudes, concerns, and reactions toward algorithmic inferences of emotion based on social media data. This paper’s focus is not these general emotional inferences, but toward automatic wellbeing interventions specifically, so we do not provide additional detail.

We then proceeded to ask participants questions about prediction, such as “*How do you feel*

about your post being used to predict how you might feel in the future? Tell me more about that. Why do you think companies might do that? How do you feel about that?” Specific to this study, we then asked questions like: “*How do you feel about the platform using this prediction or detection to intervene in some way to support your wellbeing or help you feel better?*” We probed for other application domains of emotion AI using social media data, but detail is not provided here as their scope is beyond the focus of this paper. We were intentionally broad when conducting scenarios regarding automatic wellbeing interventions. As emphasized throughout this paper, we take a broad perspective to automatic wellbeing interventions on social media, describing anything aimed at or framed as aiming to automatically provide support to social media users’ wellbeing.

Our focus was not so much the particularities of the emotional experience, but more so how participants felt about that data being fed into emotion recognition algorithms to be used for wellbeing interventions, and attitudes toward those resulting interventions themselves. Sometimes, if we needed to probe more, we brought up another example that they had mentioned earlier in the interview or survey, and asked the same questions to reveal participants’ attitudes toward automatic wellbeing interventions.

Analysis

Interviews were conducted and transcribed by the second author and another research team member. These researchers analyzed the dataset and published their findings in a separate paper [52].

The first author separately conducted a qualitative analysis of this interview data, scoped to a focus on conceptualizations of and attitudes toward automatic wellbeing interventions on social media. The first author analyzed the data using open coding followed by axial coding [216], engaging in weekly meetings with the second author to discuss observations and patterns, grouping of codes into larger themes, and analytic refinement.

Limitations and Opportunities

Our study’s goal was neither representation nor generalizability [9]. Indeed, the demographics in our study are unique in some regards. We understand that a study about sharing emotional experiences on social media may not have elicited high participation among male-identifying individuals [244], and our sample thus included majority women and other underrepresented genders. Additionally, our study is in alignment with other studies of emerging technology [48, 392] in that most of our participants have attained at least a college degree and therefore may have been more familiar with technology than the general population. Despite these unique demographics and limitations, our work provides unique insights into conversations regarding emerging technologies. Future work on attitudes of automatic wellbeing interventions should include people with lower educational

attainment, older adults, children, diverse races/ethnicities, those with mental illnesses, and people in diverse cultures and geographic locations. Future work can also use methods such as large scale surveys to examine generalizability of our findings.

In-depth interviews with smaller sample sizes allow researchers to make interpretive and generative conclusions rather than conclusions that are definitive and generalizable. Diligent participation selection allows us to explore topics of interest in depth. Our confidence in the validity of reported themes is high, as narratives were similar throughout data collection, confirmed by our analysis.

Furthermore, some participants expressed using privacy settings in general; therefore, it is possible that our work suffered from self-selection bias. Nonetheless, despite these concerns our participants had *still* chosen to share about emotional experiences on social media, which was an inclusion criterion for our study, as it enabled participants to engage in the scenarios around which our study was designed. Our study's goal was primarily to understand and make sense of how people that share emotional content on social media construct meaning from automatic wellbeing interventions, and what values and concerns they hold in this context. Though some participants may have had imperfections in recalling their past experiences sharing on social media, this imperfection would not have interfered with this goal.

Of course, those who do not post emotional content can also be subject to emotion recognition and interventions and engaging with them is important for future work. Yet, as a first step, we wanted to have our understanding grounded in participants' conceptualizations of emotions and emotional experiences, thus our choice of sampling.

For future work, we especially emphasize the importance of examining attitudes toward automatic wellbeing interventions on social media for specific mental health conditions. Individuals living with mental illnesses are not a monolith, and attitudes toward automatic wellbeing interventions may differ across and within subgroups of social media users with mental illness. Prior research has analyzed, for example, how people with eating disorders share supportive and intimate content on social media, and how they might be impacted by the *coded gaze* that makes possible algorithmic inferences of user behavioral state and inferences linked to content moderation [314, 312, 315, 181, 680]. Future work could consider how and if data subjects' use of social media for social support in mental health or other emotion-situated contexts shapes their attitudes toward automatic wellbeing interventions that target their condition specifically.

3.4 Findings

Our findings suggest that social media users have predominantly negative attitudes toward automatic wellbeing interventions on social media. First, we discuss data subjects' negative conceptualizations of automatic wellbeing interventions. Next, we discuss how people imagined the impact of

automatic wellbeing interventions on others. A minority of participants felt tension between their broad negative attitudes toward automatic wellbeing interventions and their conceptualizations of them as a potential social good for others. Most participants however maintained their general negative conceptualizations of wellbeing interventions when thinking of its impact on others. Participants expressed concern for the harm automatic wellbeing interventions may pose, and stressed that individual people should have control over whether they would be subject to them. Lastly, we discuss qualities upon which participants' attitudes depended.

3.4.1 What Only Humans Can Offer

The majority of participants held negative attitudes toward automatic wellbeing interventions. We found that these attitudes stemmed from participants' comparisons between current state wellbeing interventions, delivered by humans, and what they imagined to be future state wellbeing interventions, enabled and delivered by emotion AI technologies. The human versus AI dichotomy was a prevalent theme in participants' conceptualizations, as they considered whether AI could hold certain attributes they considered to be held by humans in wellbeing-supportive roles. These attributes included: 1) helpfulness and authentic care; 2); personal and professional expertise; 3) morality; and 4) benevolence through shared humanity.

3.4.1.1 Helpfulness and Authentic Care

Some participants doubted the helpfulness of automatic wellbeing interventions, and their ability to deliver authentic care. P1, who had personal experience with mental illness, and drew upon that when discussing their attitudes toward automatic wellbeing interventions, reflected on past experiences searching for suicide-related information on Google, saying: "*If you Google like how to kill yourself or whatever, or Google automatically served you just like the 1-800 like suicide hotline number, that as someone who had been suicidal did not strike me as very effective.*" P1 later remarked: "*I don't know that a computer is able to serve the right information to help someone,*" illustrating skepticism about algorithms' ability to provide information that would be helpful to individuals in need of support in moments of distress and vulnerability. Other participants signaled their need for wellbeing support to feel authentic, and felt uncertain that an automatic wellbeing intervention could provide authentic and thus helpful support. On lack of perceived genuine care, P5 said: "*People are people and an algorithm is an algorithm, right? It's not looking to read and ignore like most people. I make a private post on Tumblr, pretty much everybody either just casually hearts it to let you know they're there or ignores it completely because that's uncomfortable. But the algorithm is not there out of any form of interpersonal care, even if it's been put there by a human being. I don't know if I could ever envision a world in which it was put there*

to genuinely help people, which is me being a real cynic but why would they care? I don't know." These examples show how automatic wellbeing interventions can feel impersonal and unhelpful compared to interventions authentically delivered by caring, trained human professionals. Further, the unique insights provided by P1, who disclosed having a mental illness, point to a need to better understand the attitudes of those that live with mental illness toward automatic wellbeing interventions in future work.

3.4.1.2 Personal and Professional Expertise

Participants' remarks reflected a belief that automatic wellbeing interventions lack the expertise that humans have, either due to 1) their professional training or 2) personal experiences. On the first, participants compared automatic supportive interventions to humans trained to provide expert support and interventions to people in their community in times of distress (e.g. mental health professionals), arguing that AI does not, and cannot, compare to expertly trained and trusted professionals due to their expertise (rather than the ability to care as described in 4.1.1). For example, P3 said: "*I don't know, a therapist went to grad school for it. They've studied the thing.*" Echoing this sentiment, P10 said: "*I don't think that's appropriate...because I think it takes a lot of information and often a medical professional to let someone know if they're going through a particular, like a clinical problem, or if they're likely to have a clinical problem in the future.*" Believing that algorithms cannot be as expert as humans would be in providing supportive interventions was a significant factor contributing to negative attitudes toward automatic supportive interventions on social media.

On the latter, participants spoke of the enormous trust people put in friends and other community members to help in times of crisis, and could not imagine themselves able to trust in automatic wellbeing interventions to meet that need. In comparing trained mental health professionals to AI, P3 remarked: "*They also have a certain amount of community attached. I don't feel like an AI could get there.*" P3 continued: "*No, there's a reason why you might sad Tweet about things but you, in the end, will still rather call a friend and talk about it.*" Participants had a difficult time imagining a space where automatic wellbeing interventions would be welcomed, as their need for support was already filled by empathetic and compassionate humans with whom they were personally connected. Ultimately, most participants struggled to imagine how algorithmic interventions could help in the same way humans do—supporting each other within one's personal network and community—and whether there was a need for automatic wellbeing interventions at all. We note that the predominantly college-educated demographics in our sample may have influenced this view, as college-educated persons tend to have more robust networks of support [517].

3.4.1.3 Morality

Participants were skeptical of automatic wellbeing interventions on social media due to their underlying assumptions about and associated attitudes toward social media platforms' financial motivations. Those that traditionally deliver wellbeing interventions such as mental health practitioners are held to certain ethical standards of conduct, establishing an expectation of ethical and moral practice that engenders trust in those receiving wellbeing support or intervention. However, when participants imagined automatic wellbeing interventions on social media, they were cynical that algorithmic interventions—or the platforms that deliver them—could have or prioritize morality or ethics due to the assumed financial incentives of the companies that own the intervention. For example, P5 voiced their cynicism in the intervention's moral and altruistic intentions: "*[I]t could be 100 percent innocent, people who want to make people feel better, but I'm also a bit of a pragmatic, realistic person and I know that there's money in it, and they'll do it for money regardless of where the idea originated or where it came from.*" P2 elaborates further, explicitly questioning the ethical incentives of such interventions: "*I'm just not so convinced that the financial incentives of the companies are such that ... ethical incentives would take priority.*" P9 echoed this sentiment: "*I don't think they could provide support to us to feel better. I think like they just want us to, I think their goal is to earn money.*" P9 reflected that while some kind of support may be nice when one is feeling lonely, the financial motivations of automatic wellbeing interventions could taint how they view, and therefore receive, that intervention: "*I think it could make us feel nice I guess. But if there is a way that they are going to sell us a product, I think that would change how we view, how we see them...At the end of the day it's not our family, our friends, so it's not like genuine care. It's just trying to sell you something.*" These sentiments illustrate the important role of social media users' cynicism toward companies' moral intentions in shaping attitudes toward supportive interactions companies may provide. Moreover, these accounts highlight participants' discomfort of their emotions and vulnerability being commodified and capitalized by platforms delivering automatic wellbeing interventions.

3.4.1.4 Benevolence through Shared Humanity

Participants doubted whether they would welcome automatic wellbeing interventions in the same way as they would from another human. In contrast to the benevolent disclosures and interventions that take place in the context of a trusted relationship with a mental health professional or close friend, participants felt that interventions delivered through algorithmic means were intrusive and creepy, with suspect intentions. As P13 said: "*If it was something about like being sick or something, I don't know. In one case, I feel like maybe it would be good because maybe that will push you to go get it checked out, but at the same time, I'm like, that's kind of...I don't know. Maybe*

going a little too far. Maybe it's a little too intrusive." Similarly, P5 said: "*I do certainly think there is a positive way in which that system could be used. I still think it's kind of creepy but...there isn't an innocence in that sort of concept or an idea.*" Thus, although some participants charitably acknowledged positive possible uses for automatic wellbeing interventions, they resolved that those possibilities did not outweigh their perceptions of the interventions' intrusiveness and creepiness or skepticism of its intentions.

Participants felt that humanity was an essential and primary attribute required of supportive wellbeing interventions, doubting whether non-human, automatic wellbeing interventions could hold secondary attributes that they felt traditionally human-delivered wellbeing interventions held, such as morality, helpfulness, and authentic care. For example, P3 feared that because the algorithms that would deliver interventions (despite the humans that wrote them) do not share humanity with humans, they therefore lack attributes such as morality that otherwise make humans more resistant to manipulation by bad actors: "*But an algorithm is a thing. It's not a person and it doesn't have wants or desires or anything. It isn't similar to you in the way that you both have a shared humanity. It's more of a thing, and that makes me uneasy. Because that means that thing in the wrong hands can do a lot of damage. It's not a person.*" Participants felt supportive wellbeing interventions required an element of humanity that non-human algorithms and social media platforms cannot provide. P1 remarked: "*It feels really impersonal. I don't know. I think it takes more, I think it takes real empathy from a real person as opposed to some generic advice and I don't think just giving someone a 1-800 number or even just talking to a stranger on suicide hotline is really the best intervention long term.*" Echoing this sentiment, P3 said: "*...it's good to be accurate [with recognizing and predicting emotions], but there's no humanity in it, right?*" In these examples, participants felt that because automatic wellbeing interventions are computationally derived, they lacked the essential attributes of humanity and personhood that would help the recipient receive the intervention in a way they only could if delivered by a human. In the end, participants were resistant to an algorithm's ability to embody humanity and were doubtful that automatic wellbeing interventions would be helpful to whom they (are framed to) intend to support.

3.4.2 Emotion AI Uses for Social Good

Some participants imagined latencies of benefit and harm differently when conceptualizing the impact of automatic wellbeing interventions on others compared to their predominantly negative attitudes about the technology in general. A minority of participants, with reservation, imagined automatic wellbeing interventions as a potential social good. Relative to their generally negative attitudes toward automatic wellbeing interventions, these participants responded more positively to the idea of automatic wellbeing interventions on social media when conceptualizing it as a

tool that could benefit *others*, rather than themselves. For these participants, automatic wellbeing interventions held potential as a social good that could benefit others in the following ways: 1) by supporting academic research; 2) by increasing access to wellbeing support; and 3) through egregious harm prevention. It is important to note that we did not ask about these potential positives (or negatives) directly; rather these came up organically as participants discussed their attitudes and were probed to share details. In these examples, the participants maintained their negative attitudes generally, but held slightly more positive attitudes toward the specific use case of promoting social good more broadly. Even in cases of automatic wellbeing interventions for social good, participants expressed caveats that collection of emotion data and subsequent potential interventions should be transparent to individuals, and that they should meaningfully consent to it.

3.4.2.1 Supporting Academic Research

Several participants noted that automatic wellbeing related interventions could be used to support researchers. For example, P7 said “*I would want that to be used in research, and in mental health studies.*” Some felt it was best that the intervention would be developed and implemented by researchers. For example, P3 suggested they “*could trust a brand new person creating this new app with neuroscience and psychiatrist research that has the data to be like, ‘Oh yeah, I think this is going to help change the world.’*” In this example, the participant was generally cautious about social media’s deployment of automatic wellbeing interventions, but felt enthusiastic about the potential of more trusted academic researchers to create them. Other participants who were resistant overall to emotion data collection supported the idea of its use in support of academic research for mental health interventions. For instance, P5 notes they would support “*some sort of study being done to better understand and assist people mentally, especially since the internet seems to be such a hive of toxic interactions, if it’s being used to better understand people’s brains or some sort of medical or academic level, I could see where that would be fine.*” Overall, some participants had positive associations with and attitudes toward the use of emotional data and development of automatic wellbeing interventions by academic and medical researchers, even when they had more negative attitudes about automatic wellbeing interventions broadly and applied to them individually.

Past work has shown that people are generally either comfortable with or ambivalent about the use of social media data to support academic research in specific contexts, though noted that self-selection bias of participants participating in a research study, and younger, some college educated demographics from their sample from MTurk, an online crowdsourcing marketplace, may have played a role in their findings [317]. Likewise, research has suggested that people are supportive of the use of social media data to support researchers’ population-level monitoring of mental health, even when they generally had privacy concerns; however, this study focused on mostly educated people or people working in a white collar capacity [608]. We surmise that the positionality of our

largely college-educated sample, similar to past work with comparable findings [317], may have contributed to these participants' comfort in the use of social media data for research, and may not be generalizable to the broader population.

Participants stressed, however, that automatic wellbeing interventions used to support academic research should only occur if the person has knowledge of the emotion data collection and corresponding intervention. For instance, P5 adds: “*But I wouldn't be okay with that being used without anybody's knowledge because that's just shady. If you're not going to tell people, then why aren't you telling people? I doubt on such a large scale that it would affect their finding.*” P7 echoes, “*I would want that information known to me as the user.*” Thus, while some participants suggested they might feel comfortable with automatic wellbeing interventions if they were used to support research, they felt that research should be used only in a transparent context where the affected individual is informed of the practice and meaningfully consents to it.

3.4.2.2 Increasing Access to Wellbeing Support

Several participants suggested that automatic wellbeing interventions could be a positive benefit for those without a strong support network, thereby increasing access to wellbeing support. While P8 was hesitant about the potential of emotion data's use in other ways such as product advertising or targeting them individually, P8 responded positively to its use toward this type of “good.” When asked about using emotion data to create wellbeing-related interventions, P8 responded: “*That feels more like the social good side of it, like using this for good rather than like here's a weight loss pill...Like that feels less of a social good than like someone is having an acute moment than like this platform can be used to actually provide resources that might help in that moment.*” Here, P8’s attitude toward emotion detection on social media platforms is contextual: while P8 expresses discomfort with its use to promote advertising, they respond positively to its use to promote the social good of mental health support. On social media platforms’ unique position to offer wellbeing interventions, P6 acknowledges “*there's some people that don't have family to intervene and maybe that would be good for a person who does not have anyone and they're using social media as a cry for help.*” P12 echoes “*I think that's good. Again, you see kids on there and they might have a problem. They need help.*” People in distress sometimes rely on social media to seek support, and some participants viewed wellbeing interventions as a tool to provide it where users may not otherwise receive it. This acknowledgement echoes a potential motivation for some of the research behind and application of practical platform-based interventions in place today.

It is important to note that while some participants imagined automatic wellbeing interventions as a potential social good that could increase access to wellbeing support, those same participants remained uncomfortable about its use on *themselves*. Future work could consider the impact of social support networks on data subjects’ attitudes toward automatic wellbeing interventions.

3.4.2.3 Egregious Harm Prevention

Perhaps the most widely acknowledged benefit participants imagined automatic wellbeing interventions could offer was its potential to prevent egregious harm, such as in cases of social media users planning suicide, harm toward others, or domestic terrorism. Some participants considered automatic wellbeing interventions to have the potential to be particularly helpful and effective to prevent harm. P11 saw interventions as helpful “*actually for people who are an immediate threat to themselves or others... or mainly themselves*” and remarked, “*I think it could be a very good thing.*” P8 worried about the example of “*someone...searching about like ways to commit suicide or ways to hurt someone. I think that's when I, I feel like the social good and like someone's bodily safety is at risk,*” and imagined the potential of interventions to mitigate that risk. For these use cases, participants considered the good of preventing suicide and harm toward others to benefit society at large.

Furthermore, P6 discussed automatic wellbeing interventions as a potential way to reduce the number of school shootings, and thought that “*maybe there would actually be less things occurring*” if there were mental health interventions on social media. P6 elaborated that wellbeing interventions might assist public safety officials by understanding content posted by these individuals: “*we need to monitor posts a little more closely to see if there's somebody who's vaguely talking about a school shooting or something, they say, we need to sometimes be responsive, and not just take someone at their word, because someone's word may not express exactly what they're about to walk out the door and go and do.*” This participant’s account can be taken to mean that some individuals feel that surveillance on social media can be justified in efforts to prevent domestic terrorism such as school shootings, but that harm can arise when such surveillance leads to inaccurate and biased identification. Considering the rhetoric and justifications used by participants to imagine wellbeing interventions as a positive force that can prevent egregious harm, we speculate that these attitudes may have been influenced by the positive discourse of some high-profile wellbeing interventions such as Facebook’s Suicide Prevention Program as tools that promote *wellness* and *help users in need* [7]. Likewise, positive positions may have been shaped by rhetoric in the popular press that AI can help *save the planet* [8] and by the government that AI can *empower people* and *improve peoples’ lives* [11].

In summary, some participants conceptualized automatic wellbeing interventions on social media more positively when imagining its greater social impact, going beyond individual concerns. In these instances, participants resolved their tension between their general negative attitudes toward automatic wellbeing interventions and their more positive attitudes when the interventions might benefit others as a possible social good by establishing that automatic wellbeing interventions should be transparent to individuals, and that individuals should meaningfully consent to their use. Future work could delve deeper into such shifts in expectations and attitudes at individual and

collective levels.

3.4.3 Emotion AI's Potential for Harm

When conceptualizing its impact on others, most participants maintained their negative attitudes toward automatic wellness interventions on social media. These participants were primarily concerned that the intervention might commit harm to other people, and emphasized a need for individual and external control to mitigate those harms.

3.4.3.1 Potential Harms

Participants expressed a variety of concerns that automatic wellbeing interventions on social media pose a risk of harm to other users, including risks of: re-traumatization, spread of inaccurate health information, inappropriate surveillance, and interventions informed by inaccurate predictions.

Some participants expressed concern about the potential for re-traumatization caused by interventions. For instance, P10 said: *“Like, it could help people. It could also make people more angry that a machine is telling them, ‘Hey, you sound angry. Please call this number.’ Like, ‘All right, machine. Calm down. Leave me alone.’”* In this example we see that some participants feared the interventions themselves may lead to outcomes of anger and frustration to the individual, leading to re-traumatization of already vulnerable individuals.

Others felt that if medical interventions and diagnoses became commonplace on social media, people may start to believe that whatever information they are given about their wellbeing is accurate and credible. For instance, P7 said: *“It just feels like it’s going to put information into the hands of uneducated people who are then going to assume that Facebook is accurate... I feel like it’s going to lead to people...overreacting.”* In this example, the participant expresses concerns that relying on social media for wellness information can lead to the spread of misinformation, particularly among vulnerable groups such as those with low educational attainment.

Another concern expressed was that the surveillance methods required to enable automatic wellbeing interventions could be applied by individuals in other contexts that could then cause harm through privacy infringement. For example, P3 wondered, *“But again are there parents wanting to use that to monitor their kids? I understand that but I just don’t think it would be good to try to...I just feel you’ll do more harm than good but that’s my fear.”* Participants acknowledged that the data collected from constant monitoring could be used and abused by other entities, and were concerned how that might harm certain groups such as children.

Speaking further to potential harms, P6 said: *“I think that maybe there would actually be less things occurring because people use social media now for everything, as I said before, some of the things people post online, I’m like, I can’t believe you even put that on there. And maybe it would*

be very helpful, but at the same time there could be a fine line because what if you're insinuating something else and you end up investigating someone for something that has nothing to do with what you were thinking they were talking about." Participants worried that automatic wellbeing interventions, especially in cases where the prediction is inaccurate, could harm the intervened subject. Unless the individual had control and agency in the surveillance that facilitates automatic wellbeing interventions, participants felt there would be significant risk that other actors might exploit that surveillance for ethically questionable purposes.

These varied examples show that the harms people imagine automatic wellbeing interventions can commit span a wide range of concerns, and suggest that its potential harm is immense.

3.4.3.2 Individual and External Bounds

Overall, participants were concerned about the expression of power in the user-platform relationship when conceptualizing automatic wellbeing interventions and the potential for harm within that context. As P2 put it, "*Assuming that the intervention was not forced intervention, I think it would be a good thing. If the intervention were forced, then I would tend to say things have gone too far.*" In this example, we see that people are opposed to any intervention they perceive to be unconsented to and forced upon them. Participants stressed that having the choice to control whether they were subjected to these interventions would allow the intervention to reach the people who might need it, while allowing those more reserved about its outcomes a choice in whether they were subject to the intervention. Participants felt they would be more comfortable with the delivery of automatic wellbeing interventions on social media platforms if there were clearly defined boundaries to help those in need of support, and options to enable and maintain user control. P8 noted the need for bounds and control on the deployment of automatic wellbeing interventions: "*I think about it at an individual level. I don't like that idea. But when I think about [the] crisis that we're in and like I think about queer youth or whomever and things that people are posting about and are like crises that people do post to Facebook around in moments of crises. I think if it helps people who are in that acute moment, then maybe I'm okay with it, but I would want there to be like bounds on that.*" P8 was cautious about sanctioning the use of automatic wellbeing interventions for people in crisis, and was sure to underscore the need for measures that would subject the interventions to external regulation and allow for individuals' control before approving of its use.

In these examples we show that participants maintained their negative attitudes toward automatic wellbeing interventions whether they imagined it at a personal or social level. Participants expressed strong preferences for individuals to have control in whether they were subject to interventions, and for interventions to be subject to external regulation, both of which may mitigate some concerns surrounding potential harms.

3.4.4 Conditional Trust and Acceptance

While some participants maintained negative attitudes toward automatic wellbeing interventions at all costs, some imagined particular qualities that, if implemented, might engender some increased degrees of comfort and trust in the intervention. We identified three qualities upon which they felt their level of trust and comfort in this technology depended: 1) accuracy; 2) contextual sensitivity; and 3) positive outcome.

3.4.4.1 Accuracy

Some participants believed that automatic wellbeing interventions for support should be based on highly accurate inferences, and saw potential negative consequences for individual harm should the intervention fail to meet certain expectations of accuracy. For example, P3 said: “*if you see someone caught retweeting about bad shit, and it’s like then clearly you should call him if they say they want to die, they want to die. That’s not always accurate. So I feel like it would make them completely have that option that, if people are at risk or whatever, for them to use that...but again...I just feel that you’ll do more harm than good but that’s my fear.*” Participants understood that while interventions such as suicide predictions could have potential positive benefits, their accuracy would be a determining factor in whether they helped or harmed the individual user. Participants expected highly accurate algorithms that are able to understand nuanced and contextual engagement with the platform before they would consider themselves comfortable with the automatic wellbeing intervention deployed on social media.

Related to accuracy was the quality of relevance. Some participants expressed a requirement that they perceive automatic wellbeing interventions as relevant to their condition. These participants might feel more comfortable with the idea of automatic wellbeing interventions, so long as those interventions were accurate enough to offer relevant support to them. For instance, P12 said: “[*Y*ou might be able to learn something about yourself and about the condition too. I think it’s great, it’s free help, you know? As long as it’s a credible source...you can learn a lot about new treatments, and therapy, and that type of thing. It might even help you because maybe you’ve tried all these different medicines and remedies, and you’re not getting anywhere. Now they have a new breakthrough, wow look at this. I’m always researching, and always looking into new things. I would like that. It might be really good, it might help me.” Participants imagined that interventions that were accurate enough to have specific relevance to their individual conditions could then be helpful, through the advancement of individual knowledge about the relevant condition and its treatment options.

These examples highlight the importance of accurate automatic wellbeing interventions, yet suggest that they should be optional for the data subjects (not all desire precise accuracy, and some

just desire relevance), provide customized support, and be relevant to their condition.

3.4.4.2 Contextual Sensitivity

The specific wellbeing context in which interventions were provided mattered to some participants. For example, an imagined intervention suggesting resources in one's geographic area about a physical illness was seen to be less intrusive than resources regarding mental illness. To this point, P7 said: "*Let's say I have some rare medical condition and it shows me an ad for a clinical trial in my area, that could save my life. But yeah, I don't know why, maybe it's such a stigma, but for some reason if it's a mental health thing, that seems more slimy to me that they're advertising towards that, that they're taking advantage of me. But if it's like any other health issue it doesn't seem as slimy.*" Here we see that some participants felt that interventions for physical health conditions might be helpful, but felt that mental health related interventions were too intrusive and exploitative. P8 commented, as discussed in 4.2.3, that only in specific contexts, such as preventing harm toward others or themselves, that automatic wellbeing interventions could be a positive benefit to the community. P8 explains: "*If someone were searching about like ways to commit suicide or ways to hurt someone. I think that's when I, I feel like the social good and like someone's bodily safety is at risk, you know, theirs or someone else's. It feels like that's a time when the fact that this is all one soup, that should be used, but I think that would probably be the line for me.*" In this example, we see that people with overall negative attitudes toward automatic wellbeing interventions might temper their objection in contexts where the technology's potential for what they deemed as social good outweighs their own reservations.

Our results indicate that participants' comfort level with the deployment of automatic wellbeing interventions was highly dependent on the context in which the intervention would be used. For what types of interventions people would welcome automatic wellbeing interventions is an area for future research, but is certainly not a trivial question. What is more, our findings show that assuming that all automatic wellbeing-related interventions would be welcomed by data subjects is inappropriate.

3.4.4.3 Positive Outcome

Some participants' attitudes were dependent upon tangible impacts the intervention may have on them. In 4.3.1 we describe how anticipating harm was a reason for negatively held attitudes toward automatic wellbeing interventions. Here we describe how if the intervention were proven helpful, participants might be more comfortable with it; if the outcome were not helpful, they would not welcome the intervention. For instance, P7 said: "*Because if it's successful and I feel better, then I feel like I can't be upset about it.*" P10 echoed similar attitudes, and additionally suggested that

a layer of assurance such as a certification process would increase their confidence in the positive outcome: “*I think that I would feel okay with that, as long as that support is I guess somehow certified or goes through a process of guaranteeing that it’s not shitty so I feel worse. I think I could support that use of data.*” Participants felt that if the outcome of the intervention were successful, then they could embrace its use. Our findings show that data subjects have strong preferences for automatic wellbeing interventions to assure *positive* outcomes on them.

In summary, these insights into development and delivery qualities of automatic wellbeing interventions upon which data subjects’ attitudes depend suggest some individuals may welcome accurate, contextually sensitive wellbeing interventions with guaranteed positive outcomes. Who, to what extent, and in what contexts, would welcome interventions developed and implemented with such preferences is an area for future work.

3.5 Discussion

Our study examined data subjects’ conceptualizations of and attitudes toward automatic wellbeing interventions on social media. At a high level, we contribute to discourse around the development of socially aware, trustworthy, and ethically responsible AI advancements, with a focus on emotion-sensitive technologies.

Our findings suggest that data subjects’ negative conceptualizations of automatic wellbeing interventions are shaped by a human versus AI dichotomy and beliefs that automatic wellbeing interventions could not hold attributes of supportive wellbeing interventions traditionally delivered by humans: helpfulness and authentic care; personal and professional expertise; morality; and benevolence through shared humanity.

Relatively positive conceptualizations show their presence when imagining the social impact of automatic wellbeing interventions on *others*, however. Some imagined the tool as a potential social good that could benefit others by supporting academic research, increasing access to wellbeing support, and through egregious harm prevention. These positive attitudes are complicated by participants’ concerns of potential harms that automatic wellbeing interventions could present to others (e.g., re-traumatization, spread of inaccurate health information, inappropriate surveillance, and interventions informed by inaccurate predictions). Even when imagining interventions as a social good, participants expressed requirements that automatic wellbeing interventions are transparent to individuals, that individuals meaningfully consent to them, that individuals have control over their use, and that interventions are subject to external regulation. Lastly, we contribute a characterization of what makes (and does not make) for an ethical and trustworthy automatic wellbeing intervention on social media: accuracy, contextual sensitivity, and positive outcome.

While we found that data subjects’ attitudes track well to similar themes of harm and privacy

concerns found in the literature critical of emotion AI [111, 178, 211, 566, 786, 360], our study empirically centers the voices and concerns of the humans that make the technology possible to begin with—and those subject to its consequences—as relevant social groups [768] with a stake in defining the requirements and considerations for ethical and trustworthy emotion AI applications. Our goal is not to make normative statements about whether automatic mental health interventions *should* exist, but rather to contribute the voices and concerns of the humans most impacted by the technology to existing discourse.

3.5.1 What Makes an Ethical and Trustworthy System?

Participants in our study were overall consistent and clear in their rejection of automatic wellbeing interventions on social media: they neither wanted nor needed it, including those who spoke from personal mental health experiences. Participants did not trust automatic wellbeing interventions on social media to deliver support in the way humans can, and were concerned about the potential harm interventions could cause others, including re-traumatization, spread of inaccurate health information, inappropriate surveillance, and interventions informed by inaccurate predictions. Compared to human support, they deemed automatic wellbeing interventions as unhelpful, immoral, incompetent, and ineffectual. Even for those few participants that held slightly more positive attitudes regarding automatic wellbeing interventions when conceptualizing its impact on others, the benefits they imagined were counterbalanced by concerns including potential harm to individuals. These insights reflect past findings by commercial research and advisory firm Gartner, that showed out of 4,000 US and UK respondents, more than 52 percent did not want their faces to be subject to affect recognition [589]. While participants expressed requirements and qualities that might improve their attitudes toward or trust in automatic wellbeing interventions, these requirements and qualities are incompatible with current social media practices, and might be challenging to deliver (i.e., a guaranteed positive outcome).

Our findings complement work that centers human perspectives in understandings of wellbeing interventions in other contexts. For example, past work has shown that older adults express a willingness to use smart home technologies in support of self-management of wellbeing [272]; however, has not focused specifically on emotions. Yet emotions are a sensitive and unique kind of data, different from other types of data that people may deem private [52]. Future work is needed to identify the contexts in which people may welcome automatic wellbeing interventions with more nuance. For example, it may seek to understand how older adults perceive of the use of automatic wellbeing interventions using voice assistants, rather than wellbeing interventions delivered via smart home technologies broadly. Our work provides preliminary insights that data subjects are hesitant to receive automatic wellbeing interventions on social media. In addition, our paper’s

findings identifying negative conceptualizations of and attitudes toward the impact of automatic wellbeing interventions for others resonate with past work on human-AI collaboration, showing that “trust is the most correlated with human preferences of optimal human-machine delegation” [543] and that without trust, humans are not likely to feel comfortable with the delegation of traditionally human tasks to AI [444] (and as we find, especially not those as intimate as wellbeing).

If data subjects neither want nor need automatic wellbeing interventions on social media, socially aware and ethically responsible design must listen. People should not be subject to such an opaque and invasive technology through which social media companies capitalize human emotion, and consequently present harm to its data subjects. More work is needed to identify in what contexts people might welcome automatic wellbeing interventions, such as in a non-commercial medical context under the supervision of medical providers—and by proxy, medical data privacy protection and regulations. Our work has shown that data subjects have overall negative attitudes toward automatic wellbeing interventions in the context of social media, and have clear and specific requirements for accuracy, contextual sensitivity, and positive outcome before they could welcome such interventions on social media. Based on our findings, we urge social media platforms that have deployed (or are considering deploying) automatic wellbeing interventions to align their applications with data subjects’ requirements for trustworthy delivery of automatic wellbeing interventions on social media.

3.5.2 What if Individuals Consent?

We acknowledge that despite data subjects voicing alternative preferences, automatic wellbeing interventions and emotion AI more broadly will continue to expand. Emotion AI is projected to be a twenty-five billion dollar market by 2023 [13], and has current applications in industries that impact the lives of the population at large, including law enforcement [547], recruitment [174], financial services [321], medicine [400], education [2], and advertising [448]. In practice, many people are subject to emotion recognition without either their knowledge or consent, and emotion AI’s commercial viability and growth suggests that this trend will continue. For example, the Chromebooks used by children in over ten thousand schools across North America are subject to an educational management and monitoring system, GoGuardian. Its Beacon module, an automatic suicide prevention and early detection tool, is offered to all of GoGuardian’s admin customers at no additional cost [14]. Beacon algorithmically monitors “web searches, social media, chat, forums, email, and online collaboration tools” to detect students’ mental state and predict violence and safety threats, under a veneer of *safety through surveillance* [6]. Children and their parents have little to no option to opt-out, as GoGuardian “obtains school-based consent under the Children’s Online Privacy and Protection Act (COPPA)” [12].

In another far-reaching example, Facebook’s suicide intervention program scans all posts on the social media site for risk of imminent harm, with no option for individuals to opt out. In response to a journalist’s inquiry, a Facebook representative explained: “By using Facebook, you are opting into having your posts, comments, and videos (including FB live) scanned for possible suicide risk” [106]. Facebook’s suggestion that its users consent to all of their data practices by using Facebook is rooted in the “notice and choice” framework the Federal Trade Commission (FTC) uses to safeguard data privacy. Under this model, online information providers (and collectors) are required to disclose to consumers their data practices, and then the consumer decides whether or not to continue with the service [641].

While social media plays a crucial role in humans’ social capital, information access, and wellbeing, platforms themselves rely upon the commodification of the personal data people produce to sustain their business model [425]. Despite public calls for greater individual control and agency over the use and sharing of personal data on social media [428]—calls echoed by the participants in our study—platforms flex their strong position in the power asymmetry between social media platform and data subject by failing to implement tailorable and context-sensitive privacy controls [633, 86]. Instead, they offer only the binary option to accept their terms of service entirely or opt out of their service entirely. For those that try to read them [358], privacy notices are written in often obtuse, hard to understand language heavily slanted toward the interests of the service provider, with little regard for consumer interests [342, 316, 539]. Opting out of such sites as social media presents an enormous social and personal cost to individuals. To the already marginalized people that rely upon social media for crucial information and support, forcing a choice between information access and community, or privacy, autonomy, and control, only further disadvantages them while sustaining the power imbalance between data subjects and the corporations that collect and commodify their data, livelihoods, and experiences.

Thus, platforms that fall back on the traditional “notice and choice” argument in data collection (including automatic wellbeing interventions) and fail to take these criticisms into account when employing invasive, controversial technology are at odds with advances to promote ethical and socially responsible AI technologies. In current practice, inferences about mental health data are made on unwitting individuals with little to no regulatory oversight over the collection, protection, and dissemination of those inferences, under the pretense of protecting a small fraction of individuals.

More work is needed to explore alternatives to the “notice and choice” framework, and how people might actually welcome and benefit from AI-driven interventions, and not simply get accustomed to them as Zuboff warns [928], particularly in the context of platforms that employ automatic wellbeing interventions. Our findings indicating that some data subjects acknowledge automatic wellbeing interventions on social media as a potential social good, yet are 1) concerned

about its potential harms, 2) desire individual and external controls in its application, and 3) qualify that such interventions should be accurate, contextually-sensitive, and guarantee positive outcomes to the data subject, provide fertile groundwork for this important future work.

We draw attention to our finding that some people felt more positively about automatic wellbeing interventions if they were developed in concert with academic researchers. As Google ethicist Alex Hanna and AI Now Institute co-founder Meredith Whittaker have recently pointed out, the corporate gatekeepers of AI enjoy a close relationship with academic researchers by providing significant funding to top computer science departments, offering concurrent positions to researchers who hold appointments at universities, and publishing papers together. “This blurs the boundary between academic and corporate research and obscures the [economic] incentives underwriting such work” [394]. Highlighting the case of Google’s recent act of firing Timnit Gebru—co-lead of Google’s ethical AI team who researches racial and gender bias in AI systems and was let go after Google demanded she rescind a paper under peer review that exposed bias in (highly profitable) large language models—Hanna and Whittaker warn that “powerful companies like Google have the ability to co-opt, minimize, or silence criticisms of their own large-scale AI systems—systems that are at the core of their profit motives” [394]. We caution that collaborations between social media platforms and academic researchers developing automatic wellbeing interventions on their platforms might obviate data subjects’ requirements for its development and delivery qualities of accuracy, contextual sensitivity, and positive outcome, by manipulating peoples’ trust in academic institutions to silence criticism.

3.5.3 Harm to Vulnerable Populations

Emotion data should be considered sensitive in research and practice [52]. While automatic wellbeing interventions can target any individual whose emotions can be inferred or predicted from their online behavior, harms from emotion inference systems might be most acutely felt by certain vulnerable populations. In a healthcare context, vulnerable populations are defined as those “at greater risk for poor health status and healthcare access” and include the economically disadvantaged, racial and ethnic minorities, and those with chronic health conditions including mental illness, with vulnerability increasing with factors such as “race, ethnicity, age, sex, and factors such as income, insurance coverage...and absence of a usual source of care” [4]. Mental health patients are an exceptionally vulnerable population in the unregulated space of emotion AI and wellbeing interventions: they are subject to involuntary, coerced care more than any other population [820], potentially exacerbated by unregulated intervention programs. Recent work exploring mental health related apps and digital phenotyping involving technology companies broadly has shown that individuals with mental illness are wary of algorithmic inferences made

of health status and associated advertising from their use, and echo many of the concerns with mental health applications and mental health condition inferences that our study's participants had regarding corporate profit motives, distrust, and calls for controls such as external regulation [221]. As our findings show, data subjects are also concerned that automatic wellbeing interventions on social media carry significant risk of harms such as re-traumatization. For those living with mental illnesses that seek support on social media, their use of the platform might result in unwanted (and unwarranted) traumatic experiences.

Inferences made regarding mental health states can hold grave consequences, especially for racial and ethnic minorities that have been shown to be more likely to be admitted involuntarily to mental health institutions [595]. Further, the interventions that rely on those inferences, such as Facebook's suicide intervention which surrenders an individual's personal information to police, who then respond with a 'welfare check,' when it infers an individual is in need of crisis, may subject certain communities to adverse harm. Police encounters between people in behavioral crisis and police often end in unwarranted brutality [696], an outcome already disproportionately affecting Black, Brown, and Indigenous people in the US [455, 758]. When the police are called to respond to the mental health crisis of a person of color, it is an all too recurrent outcome that the individual in crisis will not only receive inadequate care, but will be subjected to police violence instead [133, 832]. In addition to a concern for harms such as re-traumatization, participants in our study expressed a concern for harm to data subjects from interventions based upon inaccurate emotion inferences. The algorithm's false positives could present harm from law enforcement involvement when a person was never in crisis in the first place, leading to uncalled for and unjustified risk. Future work should examine the impact of automatic wellbeing interventions that include protocols to involve police for mental health calls on individuals with experience being targeted by them.

Recent research has shown the feasibility of detecting emotion and "violence estimation" from social media data [892], work in which the US government has shown interest in deploying [119]. In light of the civil unrest and cultural reckoning the US has experienced with the revival of the Black Lives Matter movement in 2020, these predictions of protest activity—and their co-predictions of violent risk—from social media raise questions about the role of data harvesters and their responsibilities to the individuals that enable their technology. The dissemination of social media data that can be and has been used to target a population already disproportionately criminalized [455, 758] might produce chilling effects in civil rights protest activity, reifying and perpetuating dominant power structures. Social media companies could consider ways to prevent inappropriate uses of social media data, such as a screening measure when sharing data with third parties [439].

3.5.4 Trading Privacy for Safety

Empirical work has suggested that the apparent contradiction between individual actions in loosely sharing and disclosing information online and strong individual preferences for privacy can be resolved when understanding the nuanced contextual variables in which people disseminate information [577]. For example, sharing sensitive information such as health data within commercial flows (i.e., with a health insurance agency, or at a doctor’s office) generally meets privacy expectations within that particular, appropriate context, but the subsequent sharing of that same information in another context—say, to one’s employer or made available to public record—generally does not meet peoples’ privacy expectations [577]. Our study found that while people held generally negative attitudes toward automatic wellbeing interventions, some participants adopted a positive attitude when imagining its use in limited use cases, such as to prevent egregious harm. However, the methods required to employ an intervention tool that prevents harm necessarily means that individuals cannot be granted their preferences for privacy of emotional inferences or to share that information in contextual, nuanced settings: the algorithms must scan most or all content to be effective, thus violating the contextual integrity of the disclosed information [769, 646]. The participants in our study stressed a preference for individual autonomy and control over being subject to automatic wellbeing interventions on social media, a design option that would enable individuals to control the sharing of information they disclose online.

We argue that rather than designing privacy controls that respect individual preferences to control sensitive information sharing, which would restrict social media platforms’ commodification of valuable user data, platforms instead have focused discursive efforts to influence social norms such as those viewing interventions as a tool that promote public safety. For example, Facebook has framed their Suicide Prevention algorithm as an *AI-fueled detection effort* that provides *timely help to people in need* [7]. GoGuardian, which contracts with school districts to monitor student devices, has promoted its AI-enabled behavioral risk detection as a tool that promotes student *safety* and identifies students *in need* of a *psychological intervention* [581]. We speculate, based on our findings, that discursive strategies to frame automatic wellbeing interventions as a way to promote public safety have likely worked: participants in our study who reported to generally feel negatively toward automatic wellbeing interventions targeting themselves, somewhat contradictorily, felt they might positively impact society by preventing egregious harm. We suggest this tension might be explained by the influence of public relations efforts pushed by companies that have deployed wellbeing interventions to frame them as a positive social good, and discourse in general by the popular press and government framing AI as a human savior [8, 11]. These efforts, we suggest, gently shift social norms of mass surveillance toward acceptance [342].

As Shoshana Zuboff has argued, surveillance capitalists (as well as governments) have a vested interest in nudging people to abandon privacy and accept data collection, a practice from which

surveillance capitalists financially and strategically benefit [928]. Indeed, Facebook’s CEO Mark Zuckerberg has famously and controversially said that “privacy is no longer a social norm” [450]. This suggestion aligns with past work that has argued that people “naively or unwittingly trust their personal information to corporate platforms” and extend that trust to data-sharing with external parties such as law enforcement [266]. We surmise that the discourse by powerful actors painting AI as a tool for human salvation, along with the trust people generally place in corporate platforms, has contributed to the approval of some and apathetic acceptance of others to automatic wellbeing technologies. We urge caution of these corporate strategies to promote unfounded acceptance of and trust in mass monitoring, especially of emotions, masquerading as a public good.

3.6 Conclusion

Through centering data subjects’ conceptualizations of and attitudes toward automatic wellbeing interventions on social media, we contribute to discourse around the development of socially aware, trustworthy, and ethically responsible AI advancements. We found that people have predominantly negative attitudes toward automatic wellbeing interventions, and conceptualize harmful consequences including re-traumatization, spread of inaccurate health information, inappropriate surveillance, and inaccurate predictions. We find that data subjects’ attitudes toward automatic wellbeing interventions were rooted in their conceptualizations of the human versus AI dichotomy, and human attributes they doubted wellbeing interventions could hold. We also found that people conceptualize different concerns when thinking of the impact of automatic wellbeing interventions for others, rather than at a general or personal level. We identified qualities in either the development or delivery of the intervention upon which attitudes depended. We argue that technology companies that deliver or consider delivering automatic wellbeing interventions ought to consider the attitudes and concerns of the data subjects that enable their technology—and those vulnerable to its potential harms—in alignment with proposed industry goals to promote ethical and socially aware AI applications. Participants in our study (including those with real mental health-related experiences) did not want to be subjected to automatic wellbeing interventions and had difficulty imagining a need for them. Imposing people to such exploitative technology when they neither want nor need it—and when they do not have explicit knowledge about it—is nontransparent and ethically questionable. We argue that to increase the trustworthiness of automatic wellbeing interventions on social media, companies that deploy them would need to *at least* fulfill requirements that preemptively protect individuals from the vast harms it presents, take measures to attenuate harms, and align with data subjects’ development and design requirements. These requirements include high computational accuracy, contextual sensitivity, positive outcome guarantees, individual controls, external regulation, and meaningful consent over being subject to automatic wellbeing

interventions. We conclude with a message of caution and restraint about the use of automatic wellbeing interventions on social media in the US, based on its current regulatory landscape and social context.

CHAPTER 4

Emotion AI in Hiring: Values Underpinning Technosolutions to Labor Problems¹

4.1 Introduction

Increasingly, “artificially intelligent” hiring services claiming to infer a candidate’s emotions and other affective phenomena are entering the commercial marketplace [139], promising organizations the ability to better predict and control employment outcomes. The emergence of emotion AI hiring services comes at a social and historical moment where the ethics of using emotion AI in high stakes contexts like hiring are contested [802, 885, 365], while emotion AI’s promises to organizations are profoundly alluring [668]. While more broadly AI development and design continues to grapple with how to best approach its negative societal impacts [325], commercial adoption of emotion AI hiring services continually increases [668, 156]. Consequently, job candidates are unwittingly [600] assessed by emotion AI despite its potential legal, ethical, and privacy consequences [765, 92, 812], the lack of available guidance to mitigate those consequences in ethical, responsible ways [281], and the contested validity of the technology itself [92].

Emotions experienced in work and job seeking contexts mediate peoples’ perceptions in ways that influence future decisions and the pursuit of labor [410, 635, 725], rendering the examination of emotion AI a question of social impact [735]. While examining the role and social implications of AI in hiring is a growing and important area of scholarship [30, 31, 35, 32, 789, 529], the implications for *emotion* AI in hiring remain relatively unknown thus far. Recent requests by the United States Office of Science and Technology Policy (OSTP) for information about the social implications of technologies that infer “attributes including individual mental and emotional states” [16] further highlight the importance of examining emotion AI’s implications.

¹This chapter is based on: Kat Roemmich, Tillie Rosenberg, Serena Fan, and Nazanin Andalibi. 2023. Values in Emotion Artificial Intelligence Hiring Services: Technosolutions to Organizational Problems. Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 109 (April 2023), 28 pages. <https://doi.org/10.1145/3579543>. This material is based upon work supported by the National Science Foundation under Grant No. 2020872.

What are the implications of emotion AI in hiring for our socio-technical futures? Technology's societal implications are deeply entangled with humans' moral and political values [891], and are operationalized to make normative claims of what *should be* rather than just what *is* [779]. By eliciting how human values are negotiated and materialized in technology, we can reveal how technology is used by and affects society [518, 780]. Hiring services are a key stakeholder in the development, design, and adoption of emotion AI applications. To identify the values that underpin the desired uses of emotion AI in hiring these key stakeholders promote, we applied a values lens [518, 780] to a content analysis of the promotional claims made by 229 emotion AI hiring services on their public-facing websites to ask: *For what organizational problems do emotion AI hiring services promote their technology as a solution? By what mechanisms do emotion AI hiring services claim to solve these problems? What core values underpin these desired uses of emotion AI promoted by emotion AI hiring services?*

Our analysis contributes several key insights:

1. We find that emotion AI hiring services promote emotion AI as a technosolution to the purported organizational hiring problems of hiring (in)accuracy, hiring (mis)fit, and hiring (in)authenticity.
2. We unpack each problem to surface how emotion AI hiring services legitimize their technosolutions under corporate ideals.
3. We identify the mechanisms by which emotion AI hiring services claim to solve those problems, specifically: the automatic creation and extraction of a candidate's *affective value*, in turn facilitating the algorithmic *affective commodification* of human labor; and the automatic exclusion of candidates on the basis of informational asymmetries and inferred psycho-biological traits.
4. Lastly, we reveal the core values that underpin the desired uses of emotion AI promoted by emotion AI hiring services as techno-omnipresence, techno-omnipotence, and techno-omniscience, showing how emotion AI hiring services position use of their technology as a moral imperative by characterizing emotion AI as the one true entity capable of solving organizational problems in hiring.

4.2 Background and Related Work

We discuss emotions' relevance to hiring, followed by a review of literature on AI in organizations, workplace management technologies, and critiques of AI use in hiring.

4.2.1 Emotions in Hiring

Hiring is conventionally emotional and interpersonal [725]. To assess an applicant's candidacy, employers use signals to make estimates of a candidates' human capital, social capital, and demographic characteristics [368, 674], which in turn influence their perception of candidates' interior traits [308, 725]. Such perceptions "may stem from implicit or explicit stereotypes, perceptions of average group ability, or personal experience" [725]. Attending to the variance in scholarship studying effects of subjective hiring decisions, Rivera introduces the theoretical framework of "emotional energy development" to describe how the emotional energy [206] interviewers feel toward candidates modulates hiring outcomes. Research suggests that employers describe the interviewers' emotional experience as the most important factor when evaluating candidates [725]. That is, interviewers tend to seek new hires who are not only competent but also excite them, and with whom they anticipate developing intimate personal and professional relationships [725].

The interaction of emotions with the hiring process is dynamic and interpersonal, also affording candidates important information to determine employment outcomes [725]. Employers may penalize applicants, for example, when interactions with candidates during interviews elicit negative feelings (e.g., anger, boredom) in the interviewer or if candidates fail to elicit positive emotions (e.g., excitement). In turn, candidates process their own perceptions of the interviewers' emotional reactions to inform how they proceed with the interview process. For example, candidates might "cash in on [interviewers'] emotional responses for jobs" by leveraging that information to effectively negotiate higher salaries [725]. Thus, the emotional experience candidates and interviewers engage in to assess candidacy and negotiate employment outcomes is one that both parties can, to varying degrees, use to their advantage. The automatic, one-sided way in which emotion AI hiring services augment or replace conventional, human-based employment decisions potentially disrupts the roles emotions play in hiring processes and outcomes.

4.2.2 AI in Datafied Organizations

Organizations are increasingly implementing emotion AI systems as part of human capital and talent management strategies to automate or augment hiring decisions [668]. Emotion AI-enabled enterprise systems claim to generate inferences about internal employees' and external applicants' emotions and other affective phenomena [695, 694, 594, 164], which can then be used to assess an individual's candidacy and drive personnel decisions with data [668, 827, 826]. Such systems can be used in all stages of the recruitment process, including algorithmic candidate sourcing and matching [307], automated candidate assessments and screening [32], and fully automated hiring platforms [35].

How data-driven hiring systems justify their "ideological grounds of datafication" has important

implications for workplaces, by invoking normative expectations about which types of work and workers should be assessed and allocated “around a vision of the common good” [256]. Dencik et al. argue that studying the implications of AI in hiring from the perspective of technology providers is a critical component to understanding the technology’s broader societal implications, as it “compels us to consider how data-led processes spread and how data-informed knowledge is sought to be legitimated” [256]. Ajunwa et al. [35] reveal how automated hiring platforms provide affordances to managers that together generate a “managerial frame” that enables the “fungible” allocation of workers, whereby workers are “available on demand and easily ported between job tasks and organizations” [35]. These examples highlight how using algorithmic tools and their inferences about candidates (including of emotion and/or affect) to inform hiring decisions is closely tied to strategic efforts to manage organizational workforces in data-driven ways.

Digital surveillance and datafication in the workplace is disparately applied to and perceived by different demographic groups in ways that reproduce social inequality, often along racialized and gendered lines [788, 804, 336]. Extraction and capitalization of data has political and moral dimensions, whereby people are classified and categorized against standards of desirable behavior defined by powerful actors [336]. We build on past work articulating the implications of AI in datafied organizations to examine the the moral and political implications of emotion AI in hiring.

4.2.3 Workplace Talent Management

Using emotion AI to infer the emotions and other affective phenomena of employees and job candidates is part of a longer history of academic and organizational interest in collecting information about workers’ interior states. The industrial and organizational (I/O) psychology of personnel selection grew largely out of the work of Scott and Münsterberg, deeply influenced by Darwin’s concept of “survival of the fittest” [489].

U.S. organizations began partnering with industrial and organizational researchers to attain information about workers’ interior states in the early twentieth century to gain insight into employee loyalty, work conditions, and relations with other employees [443]. One popular psychological test, the Minnesota Multiphasic Personality Inventory (MMPI), assessed a candidate’s fit for a job or promotion by screening existing and potential employees for personality traits (e.g., neuroticism) as well as inferences of health status and conformity to sex-typed norms [378, 438].

I/O psychology of personnel selection has persisted over the decades, despite concerns of its fairness, validity, and potential for discrimination [805, 639]. Today, the application of emotion AI to these processes to collect information about workers’ and job candidates’ emotions, affect, and other interior states and traits to inform personnel and workplace decisions is part of a larger trend of workplace digital transformation [66].

4.2.4 Criticisms of Algorithmic Hiring

Scholars, activists, and industry practitioners have raised concerns around AI use in hiring regarding ethics, privacy, technical accuracy, bias, and legality. For example, a Harvard Business Review report warns that AI may be able to infer information about a candidate’s physical or mental disability in discriminatory ways, questioning the accuracy and scientific validity of AI hiring systems and AI’s ability to effectively control for adverse impact on protected groups [242]. The authors emphasize the lack of “convincing hypotheses or defensible conclusions” regarding whether and how such tools that generate inferences about an applicant based upon their physiological attributes are ethical [242], and raise questions about the effectiveness of existing legislation in the US to stem the potential discriminatory effects of AI in hiring.

Public deliberation about AI in hiring has largely focused on demographic bias concerns. In 2018, Reuters reported that Amazon’s gender-biased AI/ML talent management systems favored male applicants over other genders [241]. The controversy reinvigorated debate about how AI/ML systems reflect and perpetuate bias in hiring. Consequently, there has been increased attention aimed at mitigating algorithmic bias in hiring. Particularly, the development and application of technical solutions to this problem is an active research area [926, 340], aiming to de-bias machine learning models generally either through increasing the diversity of training datasets or technical de-biasing methods [501, 325]. However, some scholars remain skeptical about the effectiveness of technical efforts alone to mitigate algorithmic bias, including in hiring [893, 708].

On bias, Lee contends that *explicit* racial biases in algorithms can be reduced through existing policy and regulations, but that *implicit* racial biases are more difficult to detect and mitigate. Consequently, candidates adversely affected by implicit biases in hiring algorithms may have limited access to redress until larger structural changes are instituted, such as increasing diversity in workplaces and public policy [522]. Nakamura posits that implicit AI biases may privilege able-bodied candidates and reinforce discrimination against disabled people, as implicit bias can only be detected when tested on external datasets [632]—unlikely for organizations using AI developed and trained on internal data. Organizations consider this “as a feature rather than a bug—there is absolute deniability of any hiring bias against protected categories” [632].

The general commercial response to algorithmic bias concerns has involved companies that offer AI-enabled applications, claiming that they mitigate bias and discrimination in their algorithms. Some companies have provided publicly available documentation about how they mitigate bias, which researchers have analyzed. Closest to our work is a study by Raghavan et al. investigating the technical and legal implications of what automated pre-employment assessments vendors disclose about how they detect and mitigate bias [708], finding that their generally vague claims are unclear about how their datasets are selected, whether and how their models are validated, and how inferences generated are used to recommend candidates [708]. Moreover, few vendors explicitly

discuss issues of compliance and adverse impact. Those who offer more details about how they detect and de-bias their systems claim that they test their models for bias, and de-bias with technical approaches such as “removing features correlated with protected attributes when adverse impact is detected” [708]. Raghavan et. al discuss limitations of outcome-based debiasing, showing how the principles and guidelines that govern anti-discrimination law have methodological requirements (i.e., representative samples) that are not addressed by the vendors, leaving open the question of the sufficiency of self-regulatory approaches to detect and mitigate bias in hiring algorithms. They argue for policy-based approaches to better understand and address bias in algorithmic hiring practices [708]. Sanchez-Monedero et al. also analyzed publicly available content from AI hiring vendors that address bias and discrimination and situate them in the social and legal context of the UK [752]. They show how industry practices of AI hiring services, especially those developed in the US, may not meet the standards of EU law, and argue that the UK’s data protection laws and regulatory approaches to hiring anti-discrimination offer a model to countries like the US to address concerns about AI hiring vendors’ transparency and their effect of obscuring, rather than improving, “systemic discrimination in the workplace” [752].

While scholarship on the social and ethical implications of AI in hiring has increased in recent years, most have focused on either technical or legal “solutions” to address concerns of accuracy and bias in AI. There has been limited attention aimed at exploring the social and ethical implications of *emotion* AI in hiring particularly. This is important as emotions are central to our social and private lives. Additionally, scholars have linked emotion AI use to the practices of physiognomy and phrenology [803, 809, 749], such as in a law article where Stark and Hutson describe emotion AI as “Physiognomic” AI that reanimates “the pseudoscience of physiognomy and phrenology at scale” [803].

While this past work is insightful, there remains a need to *systematically* and *empirically* investigate what exactly emotion AI vendors (such as in hiring) claim and what values they embody regarding their desired uses in hiring—a gap we address by adopting methods similar to [708, 256, 752] and building on critical AI studies.

4.3 Methods

Values in technology shape our socio-technical presents and futures: with great normative weight, the values laden in technology assert how things ought to be [779], shaping the values of those affected by technologies. We turned to the claims made by emotion AI hiring services on their websites to identify the values that emerge in the desired uses of emotion AI they promote. Recent work has shown that, despite the practical challenges associated with the study of AI in opaque organizational settings, researchers can learn a lot about industry practices from the public claims

that AI service providers make about their technology [708, 752, 749]. Therefore, we conducted an in-depth content analysis of the claims made on the websites of 229 emotion AI hiring services to address our RQs.

Some methods are better suited to locating values and their position on the values dimension spectrum than others [781]. Content analysis, as Shilton notes, is an appropriate method for revealing core values [781]. Our analysis aims to describe and interrogate the values that emerge in the desired uses of emotion AI in hiring, informed particularly by Shilton’s emphasis on the importance of identifying the location of values [779]) while maintaining a broad, non-prescriptive framing [487, 780] to our identification and analysis of values in emotion AI hiring services.

Data Collection

Data collection involved three stages: 1) identify commercially available emotion AI hiring services and their websites; 2) review websites to determine eligibility for study inclusion, 3) collect website content for analysis.

Identification of emotion AI hiring services

We consulted four websites: Crunchbase (a directory of start-up vendors used to identify AI services [708]), and three crowd-sourced enterprise software review sites: G2, TrustRadius, and Capterra. We first searched Crunchbase using the following terms: *emotion recognition, affect recognition, emotion AI, emotional AI, emotion AI, emotional artificial intelligence, sentiment analysis, emotion detection, affect detection, and emotion analytics*. This returned a limited number of results, and did not successfully identify emotion AI hiring services we knew existed (e.g., through existing market reports, news articles). We then queried Crunchbase for a small number of randomly selected names of these already-identified emotion AI hiring services that were not included in Crunchbase search results, and found that their Crunchbase profile did not specify use of *emotion AI* or related terms. For example, we expected to see Retorio in our query, identified in a biometric technology policy publication as a recruitment service that generates inferences about an individual’s affective states [421]. Yet Retorio’s profile on Crunchbase was labeled with general tags such as “Artificial Intelligence” and “Machine Learning” rather than tags specifying its use of emotion AI.

The research team then conducted a superficial review of websites for already-identified emotion AI hiring services, and found that claims made on the service’s websites were especially ambiguous in their technical descriptions of their product’s underlying technology. Rather than explicitly describing their services as enabled by emotion AI, our review suggested that emotion AI hiring services generally described themselves more broadly as “AI” applications, employing vague descriptors to make claims about their product’s generation of inferences about a candidate’s

interior state. For example, our review of Retorio’s website found they referred to their technology as “behavioral analytics AI” that “revealed” candidates’ “soft skills” based on “psychological science,” rather than explicitly describing their tool as enabled by emotion AI (or other related terms). Review of the tags, classifications, and websites of the already-identified emotion AI hiring services that did not appear in emotion AI-related search results revealed that emotion AI hiring services generally employ a broad variety of non-standard and non-technical terms to refer to their technology, and do so in ways that obfuscate their identification as an emotion AI hiring service. As a result, identifying emotion AI hiring service vendors using emotion AI-related search terms posed a unique challenge to our data collection efforts.

Application of inclusion criteria

Consequently, we pivoted our data collection to first identify all commercially available Human Resources software vendors and their associated websites, and then manually reviewed each website for the following inclusion criteria: 1) if claims on the website marketed its technology to hiring organizations to inform hiring decisions, and 2) if claims on the website referenced generating inferences about a candidate’s emotions or other affective phenomena. In addition to Crunchbase, we searched industry-oriented crowd-source software review websites G2, TrustRadius, and Capterra to identify emotion AI hiring service websites to ensure we identified services that are commercially available and likely to be in current use. For each website, we collected the names and website information for all organizational software tagged under Human Resources related terms (i.e., HR Analytics, Workforce Analytics, Employee Engagement, Employee Recognition, Performance Management, Recruiting Software, Talent Management, and Talent Intelligence). This effort resulted in an initial dataset of 3195 unique commercially available Human Resource vendors.

Dataset compilation

We then divided this dataset among four researchers. Each researcher manually reviewed a website to determine each vendor’s identification as an emotion AI hiring service according to the inclusion criteria defined above. We excluded non-English websites given our research team’s lack of fluency in other languages. Between May 2021 and July 2021, this effort resulted in a dataset of 229 emotion AI hiring services and their websites. We used a browser extension that captured these 229 websites as PDF files and imported those files into a qualitative coding tool.

Data Analysis

We divided the dataset among three team members to analyze the website content for each of the 229 emotion AI hiring services, attending especially to their claims.

Values in technology can emerge in the definition of a problem and the ways in which designers develop technological solutions to solve them, and may be influenced by the assumed values of the various stakeholders with whom the technology interacts and for whom the technology is built [780, 327]. Though values manifest in locations at all points in the technological development and design process [779], we can reveal technologies' core values by identifying the *practical end* for which they are desired to be *used* [414, 861] as a means to achieve. Values cannot be directly observed, but they can be inferred from language and behavior [486]. Our analysis of emotion AI hiring service claims to elicit their promoted values thus necessarily focused on language choices. The interpretation of language is multi-dimensional, context-dependent, and individually-situated [591]; as such, we employed interpretivist [195] approaches in analysis. Given our interpretivist methods and their epistemological roots in attending to power and discourse [591, 85], we deemed quantitative approaches to qualitative data, such as quantitative measures of reliability like Inter-Rater Reliability (IRR), as inappropriate here. Our goal was not merely to report emotion AI hiring service claims as the final outcome, but to *interpret* claims as part of the methodological process. As McDonald et al. argue, when "codes are the process not the product," *non-use* of IRR is methodologically best practice. Further, it is important to note that the emotion AI hiring service claims we analyzed reflect the discourses of powerful actors: technology companies that have effectively shaped the hiring process in questionable ways, whose claims are oriented toward (and presumably influenced by) the hiring organizations for whom they design their service [327]. As a result, coding emotion AI hiring service claims *without* interpretation would have risked this study reproducing the power imbalance and inequality entrenched in emotion AI hiring service claims [137, 591].

Nevertheless, our analysis approach preserved our findings' reliability. First, the first author conducted close, open coding of a random subset of data to develop an initial codebook that organized codes into distinct units of analysis according to the type of claim made. The research team then collectively reviewed the initial codebook to develop a common understanding of how to organize open codes. Next, the remaining content was divided among three research team members. All coders used close, line by line coding by using *in vivo* codes that mirrored the language choice in emotion AI hiring service claims. This choice functioned to 1) preserve the meaning present in emotion AI hiring service claims, and 2) mitigate potential disagreement regarding coding interpretation [196]. The team met weekly to discuss and document themes that surfaced during open coding.

Once emergent themes took shape, the team collectively refined the codebook's organization

to include observed themes, enabling axial coding. The research team collaboratively developed thematic codes and grouped existing open codes by theme. Similarities and differences between researchers' codes and code groupings were iteratively identified, discussed, and resolved. Once agreement was established, the research team continued coding the remaining data with a combined open coding and axial coding approach. They continued to meet weekly to collectively discuss and refine emergent, recurrent, and divergent themes. Once axial coding was complete, the team's weekly meetings turned to collectively interpreting relationships between codes, enabling theory construction [183]. Finally, the first author employed selective coding to delimit codes [183] around the core notion of desired and promoted use of emotion AI hiring services.

Limitations

As detailed in 2.1.1, the disparate and vague ways in which emotion AI hiring services refer to their technology posed a challenge to identifying commercially available emotion AI hiring services. As our methods to identify emotion AI hiring services required subjective interpretation of their claims, we cannot say with certainty that all of the emotion AI hiring services included use emotion AI. It is possible that our interpretation of website claims resulted in the mis-classification of a service as using emotion AI. Still, only services that claimed to measure and/or infer emotion and related affective phenomena were included. Further, our methods of identifying emotion AI hiring services by narrowing down lists of pre-categorized vendors may have missed some commercially available emotion AI hiring services not listed on these sites.

We are hopeful that our comprehensive dataset ($n = 229$) of commercially available emotion AI hiring service claims, and our systematic process to identify and analyze them, mitigates the impact of these possibilities on the reliability of our analysis. Further, it is important to emphasize that our analysis is not intended as merely an identification of commercially available emotion AI hiring services and the claims they make, but to reveal the values and ideologies present in the larger, collective discourse of emotion AI hiring service claims.

Importantly, said claims made by emotion AI hiring services should be interpreted as desired uses of emotion AI promoted by emotion AI hiring services, rather than values that necessarily emerge from emotion AI use in practice. Certainly, the values that emerge in emotion AI hiring services' promotional content may be influenced by the assumed values of the groups for whom they design their technology [327], and as our findings show, these claims are legitimized by corporate ideals presumably held by hiring organizations. Nonetheless, our findings locate values that emerge in the *desired* use of emotion AI as promoted by emotion AI hiring services, revealing the core values that emerge in the proposed uses for emotion AI hiring services as a means to achieve technosolutions to organizational hiring problems [414, 861]. As such, our findings should

not be conflated with values that necessarily emerge with emotion AI use in practice – an area for future work (e.g., through interviews with hiring organizations).

Lastly, future work could build on this study to examine how emotion AI vendors whose publicly available artifacts we analyzed in this study may react to these observations and the potential implications of these services, for example, through acknowledging their services' limits.

4.4 Findings

Values in technology emerge in the practical ends the technologies are designed to achieve: the problems they purport to solve and the culmination of those means toward a greater end [414]. Our analysis shows how emotion AI hiring services promote desired uses of emotion AI as means to achieve technosolutions to three purported organizational problems. For each purported problem and its associated technosolution, we first unpack what the claimed hiring problem is, why it is a problem, and for whom. Then, we interrogate those claims to: 1) identify the corporate ideals that legitimize the purported emotion AI hiring service technosolutions as a means to achieve those ideals; 2) the mechanisms by which emotion AI hiring services claim emotion AI solves the purported organizational problem; and 3) the core values that emerge in the desired uses promoted by emotion AI hiring services to solve each problem as a means to achieve corporate ideals.

We find that emotion AI hiring services promote emotion AI as technosolutions to the purported problems of hiring (in)accuracy, hiring (mis)fit, and (in)authenticity through the creation and extraction of a candidate's *affective value*. In turn, this process enables the *affective commodification* of candidates along affective and emotional dimensions, and the exclusion of candidates on the basis of psycho-biological information generated about them by emotion AI hiring services that is asymmetrically visible to and wielded by hiring organizations. The desired uses of emotion AI that emotion AI hiring services promote to solve these hiring problems are legitimized by their claims that emotion AI use is a means to achieve corporate ideals including data-driven decision making, continuous improvement, precision, loyalty, and stability. Taken together, we locate the core values emerging in the desired uses of emotion AI promoted by emotion AI hiring services: techno-omnipresence, techno-omnipotence, and techno-omniscience showing how emotion AI hiring services position emotion AI as the *one, true entity* capable of solving hiring problems and achieving corporate ideals, organized below by the three aforementioned hiring problems.

4.4.1 Hiring (In)accuracy

The most salient claim made by emotion AI hiring services is that the adoption of their technology will improve hiring organizations' *accuracy* in hiring. As ZappyHire claims, features like their

“AI-enabled video interview” platform will “*Improve Your [Organization’s] Hiring Accuracy by 72%*” by analyzing candidates’ “personal traits,” promising organizations the ability to “*make the right hiring decision with the right data points.*” In other words, emotion AI hiring services market their product as a technological solution to the problem of hiring inaccuracies.

What is inaccuracy in hiring, and why is it a problem? According to our analysis, emotion AI hiring services claim that accuracy in hiring is achieved when the hiring decision is made in an 1) objective, 2) unbiased, and 3) intelligent way.

Objective Hiring emotion AI hiring services claim that their technology enables objective hiring by standardizing the candidate evaluation process. Notably, the operationalization of objectivity as an attribute of emotion AI in emotion AI hiring service claims is not a claim about the objectivity of emotion AI inferences, but rather, of the technology computationally applying an automatic, algorithmic process to assess all candidates with the same, purportedly objective, parameters.

Take for example HiredScore, a “human resource intelligence” provider that recently partnered with emotion AI hiring service pymetrics to infer candidates’ soft skills. HiredScore claims to “*enable a future where hiring is efficient and fair*” by ensuring “*all people are evaluated the same for the same jobs*” with their “*highly-accurate candidate scoring (A, B, C, or D).*” By applying the same parameters to all candidates, emotion AI hiring services like HiredScore deem their technology an objective way to achieve hiring accuracy.

Moreover, emotion AI hiring services claim that the defined parameters by which their service assesses all candidates are themselves objective. FaceCode, an intelligent technical interview platform that automatically analyzes candidate responses to interview questions to infer their level of engagement and other unclear “AI-based behavioral insights” claims that it “*combines objective, standardized evaluation parameters with AI-based insights for the most accurate and effortless coding interview reports ever. All to help you make the right decisions.*” At the same time, these services undermine the service’s objectivity claims by promising employers the ability to subjectively customize parameters to suit the organization’s needs. For example, the “talent intelligent platform” Eightfold.ai claims that its service can “*optimize every configuration and product feature to meet customer requirements.*” Parameters for emotion AI-based candidate selection that are designed to best suit the organization—or allow the hiring organization to customize the parameters—are not objective, but subjective, with moral and political consequences [137]. We posit that applying subjective parameters to evaluate all candidates in the same way does not sufficiently enable objective hiring. In contrast, it enables subjective hiring at scale.

Unbiased Hiring emotion AI hiring services claim that emotion AI enables unbiased, and therefore more accurate, hiring. These claims are not directly related to emotion AI’s underlying

technical capabilities (i.e., through technical debiasing methods; see Raghavan et al. [708]) but rather, as a result of emotion AI hiring services displacing human laborers and their purported biases in the hiring process.

Vendors reinforce their implication that conventional, human-based hiring decisions are a problem for organizations because they preclude organizations from achieving hiring accuracy by discrediting the role of people in the hiring process. emotion AI hiring services refer to human-based employment decisions as mere human guesses riddled with bias and subjectivity, suggesting that replacement of people in hiring decisions is necessary to achieve unbiased and accurate hiring.

For example, Elevatus, a service that analyzes video interviews, claims that “*by using our Advanced Analytics, A.I. and videos, [organizations] can start making decisions based on reliable data, rather than guesswork.*” Echoing this claim, employee engagement platform Bob, which analyzes and profiles employees according to their predicted risk of “burnout” and “taking off,” claims that their service enables hiring organizations to “*base management decisions on evidence, not assumptions.*” Similarly, iMocha, a pre-employment assessment provider that analyzes candidates’ face and voice to identify “suspicious activity” and infer their “emotional intelligence,” claims that their automated scoring “*eliminates human error in grading [applicants]...ensur[ing] that the skill evaluation process is free of human error, and it is more valid and reliable.*” Such claims demonstrate emotion AI hiring service suggestions that human-based employment decisions are at the heart of the purported problem of hiring (in)accuracy—a problem that emotion AI claims to solve.

By discrediting the human labor that conventionally makes employment decisions as mere assumptions, hunches, and guesswork that are inherently biased, emotion AI hiring services simultaneously position human hiring processes as the obstacle preventing organizations from making accurate hiring decisions and their technology as the solution to overcome it.

Intelligent Hiring Vendors underscore their claims that emotion AI in hiring is objective and unbiased with assertions of emotion AI’s superior intelligence. Reejig, a talent management software that generates inferences of candidates’ soft skills to profile and shortlist candidates, claims that organizations can “*use the infinite intel from the Reejig mastermind to map out succession plans right across your business, without bias.*” emotion AI hiring services like Reejig market emotion AI as a “mastermind” with superior “intel,” and posit that organizations that harness emotion AI’s intelligence will improve the accuracy of their workforce planning efforts without bias.

Intelligence, here, is the technology’s algorithmic ability to computationally analyze and interpret large amounts of data from multiple data sources. For example, talent platform retrain.ai claims its “accurate matching algorithms” improve hiring accuracy by “*leverag[ing] the power of artificial intelligence.*” retrain.ai claims that “*by connecting three robust datasets about people, jobs and*

training programs, we generate useful, validated, unbiased and actionable workforce intelligence,” demonstrating how emotion AI hiring services claim that through computational ability to derive insights from disparate datasets, emotion AI offers organizations superior intelligence to solve the problem of hiring (in)accuracy.

Thus, to solve the business problem of hiring (in)accuracy, presumably a dilemma for organizations that make employment decisions that rely solely on human-based employment decisions, emotion AI hiring services offer their products as a technosolution that promises objective, unbiased, and intelligent hiring. Now that we have unpacked the purported problem of hiring (in)accuracy, the following sections interrogate emotion AI hiring service claims that reveal 1) the *corporate ideals* that legitimize emotion AI hiring services as a desirable technosolution to achieve hiring accuracy, 2) the *mechanism* by which emotion AI hiring services claim to do so, and 3) the *core values* that underpin the desired and promoted use of emotion AI hiring services to solve organizational hiring (in)accuracy.

4.4.1.1 Corporate Ideals: Data-driven Decision Making and Continuous Improvement

In promoting the desired use of emotion AI to achieve hiring accuracy, emotion AI hiring services appeal to corporate ideals of data-driven decision making and continuous improvement. By promising organizations these qualities, emotion AI hiring services not only legitimize their technosolution to the purported organizational problem of hiring (in)accuracy, but also position emotion AI as a moral imperative required to achieve the organization’s greater ideals.

Jive, a people analytics and productivity management software that uses features such as continuous sentiment analysis to keep an “*ongoing, real-time read on employee morale and engagement,*” claims to empathize with organizations’ experiences of hiring (in)accuracy: “*You’ve had to rely on hunches, vague statistics and hindsight. But...what if you had accurate, data-driven insight to guide your tactics, make timely corrections and better target your efforts for maximum impact?*” In positioning their technology as a solution to hiring (in)accuracy, emotion AI hiring services like Jive appeal to organizational ideals: organizations are promised the ability to continuously improve their workforce with data-driven decisions and insight. Achieving these ideals is not simply an added bonus, but a moral imperative. As demonstrated by iMocha, emotion AI hiring services claim that their technology “*should be arranged for objectivity of scoring, and the elimination of personal judgment concerning correct answers,*” underscoring how emotion AI hiring services normatively claim their services “should” be used to enable objective and accurate assessments of job candidates, rather than subjective, inaccurate human-based assessments. These claims illustrate how emotion AI hiring services are positioned not only as a solution to a particular organizational problem, but as an *imperative* for organizations to adopt to achieve corporate ideals of data-driven employment decisions and continuous organizational improvement.

To make data-driven decisions about human capital, thus, organizations must first quantify and measure those features. Lattice, a people analytics provider that applies sentiment analysis to employee-generated enterprise data, claims that its software can motivate team members by offering organizations the ability to “*measure the health of [their] organization and take data-driven action to increase productivity, decrease employee turnover, and build a winning culture.*” To achieve organizational labor objectives that continuously improve human capital investments (e.g., increased employee productivity and retention, “winning” organizational cultures), these services suggest that features contributing to human capital can and should be measured. Such measurements purportedly enable organizations to make better, more data-driven decisions.

JourneyFront, the self-proclaimed “World’s Most Accurate Hiring Software,” infers candidates’ personality, values, satisfaction, and other internal states and traits from pre-employment assessments and job interviews. Offering an example of how data-driven decision-making imperatives are underlain by an ideal of continuous improvement, JourneyFront claims “*continuous improvement...if you can’t measure it, you can’t improve it.*” If quantified, usable data points to use as measures are necessary for improving accuracy in hiring, organizations must use that data to achieve ideals of continuously improving organizational processes. As Journeyfront claims, “*our process constantly tests, tracks, and makes changes that continuously improve your hiring process.*” Calling attention to the organizational imperative to achieve corporate ideals such as continuous improvement, Jive reiterates: “*After all, if what you’re doing isn’t improving your results, why do it?*”

Legitimized by corporate ideals of data-driven decision making and continuous improvement, emotion AI hiring services position their technology as an imperative to achieve these goals.

4.4.1.2 Mechanism: Creating Affective Value and Affective Commodification

In order for candidates to advance in hiring processes using emotion AI, candidates must have what we refer to as *affective value*: the emotional data generated about job candidates as a measure of the candidates’ value to the emotion AI hiring service and hiring organization.

Through automatically generating inferences about one’s candidacy by analyzing their affective expressions, and advancing those candidates through the hiring funnel who fall within parameters of the emotion AI hiring service’s desired *affective value*—and excluding those who don’t—candidates then are rewarded if they satisfy the encoded expectations of *affective value*.

For example, JourneyFront claims that its “auto-score” feature scores and ranks candidates after generating inferences of their emotional and affective traits, allowing organizations to “*automatically filter qualified candidates*” in order to “*save time and know where to focus [their hiring] efforts.*” Similarly, Jabri, a talent acquisition and video interview provider that measures candidates’ “emotional and social aptitudes like interpersonal skills, communication skills, and

personality traits” claims that hiring organizations can use Jabri’s “game-changing analytics” to “uncover crucial insights” about candidates by using “*the power of Jabri’s digital video interview to discover their character*,” purportedly enabling organizations to “*review all critical personality skills important to [the] organization*.” Here, emotion AI hiring services like Journeyfront and Jabri suggest that by “*measur[ing] what matters*”—the candidate’s affective value—hiring organizations can automatically make accurate hires by excluding those candidates whose affective valuation is deemed unworthy of the hiring organizations’ time and efforts, while advancing and rewarding those that do.

The algorithms assess and rank candidates with encoded rules that assign value to candidates’ emotional and affective expression—their affective value. By creating a determination of the desired affective value of a candidate, and making employment decisions in part based on whether emotion AI hiring services’ valuation of that candidate meets encoded expectations of affective value, emotion AI hiring services introduce invisible “rules” to the hiring process to which candidates don’t have access. Candidates’ affective value is then commodified through the process of candidate selection and subsequent employment decisions that purchase the labor of those candidates who meet the emotion AI hiring service and hiring organizations’ expectations of affective value.

4.4.1.3 Core Value: Techno-omnipresence

The technosolution of emotion AI hiring services as a means to solve the purported problem of hiring (in)accuracy rests on its ability to be present everywhere: to generate inferences that reach into all places, even the once private domain of an individual’s inner emotions, through their purported ability to access and process large amounts of data about a candidate in an objective, intelligent, and unbiased way. By positioning emotion AI hiring services as the only means to achieve hiring accuracy, emotion AI hiring services supplant human labor that conventionally makes hiring decisions and their alleged subjective and biased limitations as inherently incapable of solving the problem of hiring (in)accuracy due to their limited capacity to *be* everywhere.

Legitimized and enabled by corporate ideals of data-driven decision making and continuous improvement, the moral imperative in emotion AI hiring service claims that emotion AI “should” be implemented to correct the subjective and biased features of human-based employment decisions, reflects a belief in what we term *techno-omnipresence*; that emotion AI can and should be everywhere—in places previously inaccessible (i.e., a candidate’s internal state) and by replacing the presence of the human labor that conventionally makes hiring decisions. Emotion AI hiring services’ embodiment of techno-omnipresence as illustrated here thus appeals to beliefs in emotion AI’s divine superiority over humans.

QPage, an “AI Mock Interview Machine,” demonstrates these beliefs in techno-omnipresent values, predicated on beliefs of emotion AI’s superiority over humans: “*Picking out the right*

talent by conducting an interview seems like a job for everyone, or at least, that's what we all tell ourselves. In reality, however, choosing the right talent is well beyond ordinary comprehension, and it should be left to professional software." The belief that emotion AI is superior to humans, with its omnipresent abilities to make employment decisions "beyond ordinary comprehension," thus underpins the technosolution of emotion AI hiring services to solve hiring (in)accuracy. By situating emotion AI hiring services as the sole means to achieve hiring accuracy, beliefs in emotion AI's techno-omnipresence justify the displacement of human labor required to solve hiring (in)accuracy and the mechanisms by which emotion AI hiring services claim to do so: creating affective value and commodifying affect accordingly.

4.4.2 Hiring (Mis)fit

What is a hiring (mis)fit and why is it a problem? emotion AI hiring services claim that "fit" is an organizational imperative that occurs when there is alignment between the job candidate and the hiring organization along the axes of values, beliefs, character, and culture. Conversely, a "misfit" is a candidate who does not fit those attributes. Misfits are an alleged problem because hiring (mis)fits impair corporate efficiency.

As HRPuls, a pre-employment assessment provider of automated "psychometry" that claims to recognize candidates' "conscious and unconscious motives" puts it, by "*identifying motivation and values through cultural fit analysis*" organizations can make employment decisions where "*talent matches the company's values.*" To achieve hiring fit, organizations must first algorithmically measure candidates' values, beliefs, and character to assess whether they fit the organization's culture. emotion AI hiring services position their technology as the only means to do so, with its unique capability to generate inferences about candidates' internal states as proxies for these attributes. As an example, Equalture, a "neuroscientific gamification" vendor that measures and auto-scores interior traits of current employees and external job candidates "*to hire the best-fits without bias,*" explains to hiring organizations that "*AI can help hire for cultural fit*" by first "validating" company culture. To do so, Equalture subjects hiring organizations' current workforce to their emotion AI technology to "*objectively...assess whether candidates are aligned to this culture,*" claiming that self-assessments for fit by either the job applicant or the hiring organization "*will never be objective.*" Moreover, emotion AI hiring services frame hiring for "fit" in normative terms. For example, Equalture, claims that "*the principle of hiring for cultural fit is the #1 rising star in recruitment,*" demonstrating how emotion AI hiring services posit the achievement of hiring for fit as a hiring "principle."

Services legitimize this principle by positioning hiring (mis)fits as a problem that impairs corporate efficiency. For example, Ducknowl, a service that measures candidates' "soft skills" in

video interviews, claims to improve hiring “efficiency” with their technology’s purported ability to “*identify candidates with strong resumes but who won’t fit well in an organization... lead[ing] to quick and hassle-free hiring results.*” Here, we see how services like Duckknowl characterize hiring (mis)fits as a “hassle.” Likewise, HireOnboard, a software that automatically infers candidates’ cognitive and personality traits to assess “culture fit,” claims that hiring organizations that select the “right” talent will preserve organizational resources by “eliminat[ing] applicant mischiefs.” Such claims suggest that hiring (mis)fits are a waste of organizational resources, and that this problem can be “eliminated” by hiring for fit aided by emotion AI.

emotion AI hiring services promise organizations that adopting their technology offers the ability to hire only “fits” and exclude hiring (mis)fits, efficiently and at scale. As illustrated by Humantic, an “AI with Emotional Intelligence” that claims its automated pre-employment assessments and analyses of bot-based candidate communication and interview records will “convert 30% more top candidates by using custom personalization assistance,” whose technology promises to preserve organizational efficiency through emotion AI’s alleged ability to “judge [candidate] culture fit without taking a test” and develop “*data-driven candidate shortlists that take zero effort.*” Similarly, services like Logi-Serve promise efficient hiring for “fit” with “interactive job simulations” that automatically infer an applicant’s personality and other interior traits to “*identify top performers*”, “*determine a candidate’s job fit and aptitude to perform*” and “*instantly predict future performance.*”

Altogether, emotion AI hiring services position their technology as the only means to solve the problem of hiring (mis)fits and achieve the organizational imperative to hire for “fit,” through emotion AI’s purported ability to efficiently measure job candidates’ and hiring organizations’ values, beliefs, characters, and cultures. The automatic nature of the service promises organizations the ability to automatically determine the “right” hiring fit and exclude hiring (mis)fits.

4.4.2.1 Corporate Ideals: Precision and Loyalty

emotion AI hiring services claim that organizations can solve hiring (mis)fit problems by adopting their technology’s alleged ability to *precisely* measure the emotions and related affective phenomena of job candidates and workers to produce precise hiring fits. For example, HireOnboard claims that its AI-enabled “culture fit” assessments that measure interior traits like cognitive ability and personality will “*find the perfect fit for the job,*” illustrating how emotion AI hiring services promise to identify “perfect” hiring fits with absolute precision. While the emotion AI hiring services’ determination of hiring fit—a candidate that precisely aligns with organizational values, beliefs, character, and culture—*prima facie* appears value-neutral and “objective,” claims that hiring fits promise organizational loyalty demonstrate how such assessments are subjectively oriented toward organizational desires.

As one example, HRPuls claims its “psychometric” and “cultural fit” assessments select “talents that really fit” by identifying candidates whose “motivation and values” match “*the company’s values...[and] achieve higher productivity, satisfaction, and loyalty to the company.*” Here, we see how emotion AI hiring service determination of hiring fits are those whose human values like loyalty are oriented toward organizational ideals. As illustrated by Journeyfront, “*When a person is working on things they are interested in they are more engaged, work hard, and stay at jobs longer. Measuring a person’s interests is a must, when considering accurate job fit.*” By precisely measuring for job “fit,” emotion AI hiring services suggest that the traits they measure for alignment with organizational values, beliefs, character and culture are traits of *loyal* employees that will “work hard” and “stay at jobs longer,” maximizing benefit to hiring organizations.

The subjective determination of hiring “fits” whose emotional and affective traits meet organizational desires is further illustrated by people analytics provider KQ analytics, claiming that hiring organizations that adopt their technology can stay “*focused on building a high-performance organization that lives and breathes [the organization’s] values.*” The precise measurement of a hiring “fit” then, on the basis of a candidate’s internal states and traits, promises precisely “perfect” hiring fits that loyally live and breathe their commitment to the organization.

4.4.2.2 Mechanism: Information and Psycho-biological Exclusion

For emotion AI hiring services to solve the problem of hiring (mis)fits, they not only identify “perfect” hiring fits but exclude hiring (mis)fits. As illustrated by RecruitPack, a “predictive hiring” software that claims to read candidates’ psychometric attributes, then automatically ranks and scores them to “pick ‘A-player’ candidates,” emotion AI hiring services hire for fit by promising organizations the ability to move “*forward only those with desired attitudes and culture fit,*” by “*identifying misfits in attitudes and values at the time of application.*” According to emotion AI hiring service claims, excluding hiring (mis)fits is not only necessary to hire job fits, but a desirable outcome that avoids wasting organizational resources. As RecruitPack promises: “*you can eliminate [misfits] early and focus on the best candidates.*” Thus, the mechanism by which emotion AI hiring services solve the problem of hiring (mis)fits is by *exclusion*: 1) by emotion AI generating inferences that hiring organizations can use to make employment decisions while not making visible such information and/or its existence to job candidates, emotion AI hiring services enhance information asymmetry between candidates and hiring organizations; and 2) by emotion AI hiring services generating inferences about a candidate’s emotional and affective states they claim are psycho-biological, immutable attributes inherent to their personhood, they purport to solve the problem of hiring (mis)fits by excluding candidates whom the emotion AI deems to lack those attributes.

Information Exclusion Conventional, human-based assessments of “fit” between hiring organizations and job candidates involve dynamic, human interactions that inform a mutual determination of “fit” by both parties (i.e. during live, onsite interviews). In these processes, hiring organizations and candidates provide each other information about their respective values, characters, beliefs, and culture that each can use to determine “fit” for the job. emotion AI hiring services replace the conventionally mutual, two-way evaluation of “fit” with an automated, one-sided process that *excludes* candidates from participating in a determination of hiring fit by design.

The automatic, one-sided process deprives candidates of the opportunity to gain information they need to assess mutual fit, while generating information for hiring organizations to use to their advantage: to assess “fit” on the basis of whether candidates’ internal attributes align with organizations’ desires by using inferences emotion AI generates about candidates but generally does not make visible to them.² The exclusion of candidates’ participation in assessing job “fit” is thus a *feature* of emotion AI hiring services. As Zappyhire, a software that features “robotic video interviews,” “AI assessments,” and other “predictive hiring” tools depicts, emotion AI hiring services promise hiring organizations that their determination of hiring “fits” will enable hiring organizations to spend their time and resources only on those candidates who “matter,” by assessing candidates “*even before [hiring organizations] speak to them.*”

Altogether, we see how emotion AI hiring services frame the automatic exclusion of job candidates as a benefit to hiring organizations, to avoid wasting resources on mis(fit) candidates by automatically generating inferences about candidates’ internal traits beyond what they explicitly choose to disclose. In this process, emotion AI hiring services generate an information asymmetry between jobseekers and employers that reinforces the power employers already wield over job candidates, and further disadvantages jobseekers by excluding candidates from the participation of determining hiring fit with information (and/or the existence of it) that is invisible to candidates.

Psycho-biological Exclusion In addition to information exclusion, we identify a second mechanism by which emotion AI hiring services claim to solve the purported problem of hiring (mis)fits: through the *exclusion* of (mis)fits based upon presumed immutable, psycho-biological attributes.

emotion AI hiring service claims suggest their inferences of a candidates’ emotional and affective states identify psycho-biological attributes about a candidate’s personhood, and assume those attributes are immutable. As evident in their service’s name, HireMojo’s “Job Genome Project” assumes there is a biological, genomic component to an individual candidate’s fitness for a job. As HireMojo asserts, “*historical, analytic and prescriptive analytics combined with machine learning*

²Notably, the validity and accuracy of emotion AI is highly contested [225], with biases that reflect and perpetuate discrimination against minority groups [429, 472, 369, 719, 885]. As such, the information emotion AI generates about candidates’ emotional and affective states to assess hiring “fit” may be inaccurate, but candidates cannot correct this information if inferences are not visible to them.

and big data is yielding not only answers to the problem that many have never considered, but new questions that redefine relationships." Here, we see how the "Job Genome Project" assumes there is a scientific basis to the determination of a person's fit for a job, based on biological markers that can be quantified to determine fit. Similarly, Jive claims that its sentiment tracking will improve "*culture fit while employees thrive naturally*," suggesting that there is a biological, "natural" component to emotion AI's automatic determination of job "fit." Exemplifying how emotion AI hiring services presume a biological and immutable basis to the attributes they claim to identify, HRPuls claims its "psychometric pre-employment assessments "select talents that really fit" by "*determin[ing] values and corporate cultural competence by means of complex algorithms, evolutionary processes and computer linguistics*." These examples illustrate how emotion AI hiring services justify their inferences about a candidate's interiority and assumptions that those characteristics are immutable by suggesting that, as an "evolutionary process," candidate selection through psycho-biological exclusion is "natural."

To further ground their assumptions and legitimize their claims, emotion AI hiring services invoke scientific validity. For example, Good&Co claims its software can determine "cultural fit" through its "Proprietary Psychometric Algorithm (PPA)" that "*Explore[s] Candidates Beyond Their Resume*." Good&Co claims its "bespoke measurement tool" is "*steeped in decades of research into career and individual differences literature, [and] is based on psycho-biological frameworks of personality, rooted in neuroscience and behavioral genetics*." Here, we see how emotion AI hiring services justify excluding hiring (mis)fits based on assumptions that hiring for fit involves the identification of psycho-biological characteristics assumed to be immutable. Services legitimize the selection (and exclusion) of candidates based upon those characteristics by invoking "scientific" authority that assumes a genetic, evolutionary component to hiring for fit. It is worth noting that this controversial field of psychometrics has been used to legitimize racist and misogynist beliefs and practices [548, 745].

4.4.2.3 Core Value: Techno-omnipotence

The value emerging in adopting emotion AI hiring services as a technosolution to the purported problem of hiring (mis)fit is a belief in *techno-omnipotence* —that emotion AI technology can and should have the power to determine hiring "fits" and exclude hiring (mis)fits. emotion AI hiring services remove the power candidates have to determine whether a job is a "fit" for themselves with automated assessments that, as Good&Co puts it, use "*intelligent, scientifically derived, and probability-driven algorithms [to] match jobseekers with the culture that's right for them*." In their exclusion of (mis)fit candidates predicated upon racist and misogynist scientific assumptions, emotion AI hiring services exercise a *techno-omnipotent* power over both job candidates and hiring organizations to determine hiring (mis)fit "*for them*."

Belief in emotion AI's techno-omnipotence is reinforced by emotion AI hiring service claims reflecting a belief in emotion AI's divine power. emotion AI hiring services commonly describe their technology as powerful (i.e. "AI-powered") and appeal to beliefs in emotion AI's superior capabilities to justify transfers of power to emotion AI hiring services. For example, Jabri invites organizations to "*use the power of Jabri's digital video interview to discover their character.*" As illustrated here, Services like Jabri claim to solve the problem of hiring (mis)fit by relying upon an uncontested belief in emotion AI's superior power to "discover" a candidate's character and determine a candidate's fitness for the job.

Belief in emotion AI's techno-omnipotence reflects hiring organizations' attraction to power and dominion, promising organizations these qualities by first transferring control to emotion AI hiring services to determine hiring fits, and then using that bestowed power to inform workforce strategies. For example, Eightfold, an "AI-powered talent intelligence" platform, promises adoptive hiring organizations to share in the power of their "*deep-learning AI to help each person understand their career potential, and each enterprise understand the potential of their workforce.*" Similarly, RecruitPack promises hiring organizations that adopt its "unique blend of psychometric tools" the ability to "*quickly identify those with the can-do skills, will-do attitudes, and the fit-to characteristics for your role and your organisation*" and "*consistently shortlist the best applicants and secure them before your competitors do,*" illustrating how emotion AI hiring services promise hiring organizations enhanced abilities to maintain control over employees by using the "power" of emotion AI to "secure" hires that best "fit-to" the organization. These examples demonstrate how emotion AI hiring services promise organizations benefits that strengthen organizational control and domination over the workforce, by surrendering to the purportedly superior power of emotion AI to assess a candidate's potential and determine hiring fit.

4.4.3 Hiring (In)Authenticity

What is truth in hiring and why is it a problem for organizations?

According to emotion AI hiring service claims, hiring organizations achieve truth in hiring when they have insights into a candidate's interiority to fully and deeply *authenticate* a candidate. QPage, an "Autonomous Hiring Software" that offers automated psychometric assessments, claims to "get a deeper insight" about candidates to "verify" the truth about them, enabling organizations to "*decide on the next action by having a full flow of information from candidates' detailed analysis.*" QPage claims that its "scientifically based" assessments combine "*measurement of cognitive skills and personality traits that will result in the best candidates match[ed to] the right job.*" Similarly, Reejig claims its technology allows organizations to "*gain full skills visibility so that you can have informed and accurate data to power your talent ...planning.*" These examples

illustrate how emotion AI hiring services claim to verify the “full” truth about who candidates are by extracting “deeper” insights, and combining those inferences with other information to enable hiring organizations to make hiring decisions with “full visibility” into who a candidate is. The assumption that underlies this claim is that by emotion AI hiring services generating inferences about an applicant’s interior states and traits, they extract “deep” knowledge about their interiority to reveal the “full” and complete truth about a person – their authentic self.

To establish hiring truth as a problem, emotion AI hiring services position candidates as untrustworthy and inauthentic, and rely upon an assumption that there is an objective truth that can be revealed about presumably distrustful candidates beyond what they choose to disclose. For example, Equalture, a provider of “neuro-scientific gamification” pre-employment assessments, asserts that *“of course intelligence isn’t something you can fake; personality, however, is one of the easiest things to fake.”* Offering an explanation as to why a candidate might “fake” their personality, Equalture says, *“No, it’s indeed not smart to do, but you just want that job.”* Here, we see how emotion AI hiring services like Equalture refer to candidates and how they present themselves to hiring organizations as “fake,” justifying the use of emotion AI hiring services to extract the truth about candidates’ personality and other inner states beyond the “fake” and inauthentic persona they are deemed to display.

By adopting emotion AI hiring services, hiring organizations are promised a way to avoid making employment decisions based upon untrustworthy and inauthentic candidates and their “fake” displays of personality, by truly—“fully” and “deeply”—knowing a candidate. emotion AI hiring services like Ducknowl claim that their technology allows organizations to *“find the genuine candidate”* and *“avoid bait-and-switch situations.”* And they promise to do so quickly, with emotion AI hiring services like Idwall’s “face match technology” that promises to automatically uncover the truth about candidates with “automated solutions” that claim to read candidate emotions to ensure candidates are *“really who they say they are,”* to allow organizations to hire “quicker.”

4.4.3.1 Corporate Ideals: Stability

emotion AI hiring services appeal to the corporate ideal of stability to legitimize use of emotion AI to solve the purported problem of hiring (in)authenticity. emotion AI hiring services frame their technology’s “insights” into a candidate’s interiority as a technosolution that mitigates uncertainty associated with “fake”, inauthentic, and untrustworthy candidates, enabling organizations to make stable hiring decisions with information that purports to reveal the “truth” about candidates.

FaceCode, the self-described “most intelligent coding interview platform,” measures candidates’ interior attributes like engagement and “high-level thinking” during video interviews. Demonstrating how emotion AI hiring services promise hiring organizations the ability to make “truly informed” hiring decisions by generating inferences about a candidates’ internal states and traits,

FaceCode claims that adoptive hiring organizations can make “*truly informed hiring decisions thanks to automated interview summaries with AI-based behavioral insights.*” Likewise, Eightfold, a provider that aggregates data about candidates from multiple sources (i.e., HR data and public web data) to create “rich profiles with deep insights” that provide “contextual intelligence” about candidates, promises that its “*deep-learning AI not only delivers a comprehensive understanding of workforce capabilities, but also understands each individual’s capabilities, skills adjacencies, and demonstrated learnability to provide a concrete, future orientation to talent strategy.*” Services like FaceCode and Eightfold illustrate how emotion AI hiring services appeal to the organizational ideal of stability to legitimize emotion AI as the technosolution to the alleged problem of hiring (in)authenticity: by using emotion AI to truly “understand” candidates, emotion AI hiring services promise hiring organizations a more “stable” “future” with “truly informed” talent strategies.

Under the assumption that emotion AI’s inferences reveal the whole truth about a candidate, emotion AI hiring services like Duckknowl promise to “mitigate the risk” and associated instability from making uncertain, “bait-and-switch” hiring decisions with “predictive” hiring. Retorio, a video interview platform that claims that its AI technology will “reveal hidden soft skills and traits,” allegedly “*measures personality and predicts future potential.*” emotion AI hiring services like Retorio appeal to the corporate ideal of stability by offering their services as a way to “predict” talent outcomes by “revealing” “hidden” information about candidates. Further, QPage, an AI “Mock Interview Machine” provider, claims that conventional interviews are “*rarely predictive of success on the job.*” By positioning emotion AI hiring services like QPage’s “AI-led,” “automated interactive interviews” as a technosolution to hiring (in)authenticity that enables hiring organizations to better predict talent success, emotion AI hiring services promise less risk and more stability to the hiring organization.

These examples illustrate the promise that emotion AI use leads to certainty, predictability, and stability for organizations. By avoiding hiring decisions made without the whole truth about who a candidate is authentically—“*those ‘bad hires’ who look good at interviews but under-perform on the job,*” as RecruitPack describes it—emotion AI hiring services promise a hiring solution that mitigates the “risk” associated with “those bad hires” and in turn, a more stable, predictable organizational “future.”

4.4.3.2 Mechanism: Extraction of Affective Value

The mechanism by which emotion AI hiring services purport to solve the hiring (in)authenticity problem is extraction: using emotion AI to extract information about candidates’ interiority beyond what they choose to share about themselves to apply an affective valuation that assesses whether candidates “really are” of value to the organization.

The information extracted about a candidate’s interiority carries a particular value to the organi-

zation: a candidate's affective value. emotion AI hiring services like Reejig promise that by using emotion AI to “*extract meaning*” about a candidate, organizations can “*create their workforce of the future*,” illustrating how the affective value emotion AI hiring services obtain about candidates is a valuation oriented toward organizational goals. eLamp, a service that claims to assess candidates’ “critical” skills, including “soft skills” from “any document,” echoes this assertion. eLamp posits that “*getting to know one’s employees better enables [organizations] to make decisions that are anticipated and better adapted to operational demand*,” demonstrating how emotion AI hiring services presume that by extracting information to truly know a candidate and assess whether they have affective value to their company, hiring organizations can make better “anticipated” and stable hiring decisions that suit the organization’s needs.

HireVue is a known emotion AI hiring service that recently discontinued facial recognition-based emotion AI after high profile criticism [175]; however, according to its website at the time of data collection, continues to generate inferences about candidates’ internal states with speech and text inputs. HireVue echoes this promise with claims that their technology enables organizations to “*Go Beyond Resumes*” to reveal “*what really matters*” about candidates. emotion AI hiring services like HireVue promise that using their technology to extract information about a candidate’s interiority ensures that organizations “*engag[e] with the highest quality candidates first*.” Thus, the technosolution of emotion AI to solve the problem of hiring (in)authenticity by extracting “what really matters”— candidates with high affective value to the organization - purportedly enables hiring organizations to hire the “highest quality” candidates.

4.4.3.3 Core Value: Techno-omniscience

The core value that emerges in emotion AI hiring services ’ technological solution to the purported problem of hiring (in)authenticity is a belief in *techno-omniscience*, the idea that emotion AI embodies all-knowing “intelligence,” and its supreme ability to completely know who a person truly is ought to be used to attain authenticity in hiring.

The assumption that information about one’s interior states and traits constitutes one’s authentic, complete, true self, and that hiring organizations have a legitimate interest in knowing a candidate’s authentic self to determine one’s candidacy, form the foundation to the purported problem of hiring (in)authenticity. emotion AI hiring services claim that by emotion AI transgressing “beyond” what a candidate willingly and intentionally shares, emotion AI has the sole, supreme power to truly know a candidate. As exemplified by Adoreboard, whose “Emotics” text analysis platform classifies “*over 24 emotions from any text*,” emotion AI hiring services claim to solve the purported problem of hiring (in)authenticity “*by revealing the ‘Unknown Unknowns’ of...Employee Emotions*” to deliver “business answers” only made solvable with emotion AI’s purportedly superior knowledge about who a candidate truly is. Thus, emotion AI hiring services’ technosolution to this problem requires

a belief in emotion AI’s techno-omniscience to solve it.

4.5 Discussion

Ethics is not missing in technology. Our ethics and values are always present in the creation and use of technology. The technology society creates and chooses not to create is a window into the ethics and values of the powerful” [607].

Birhane argues that one reason why AI/ML practitioners have limited their engagement with the ethical and social implications of their field is related to the dominance of a rationalist “God’s eye view” paradigm: the view that data science practices have a uniquely superior ability to construct objective and absolute knowledge with advanced computational methods that overcome historical challenges to attaining such knowledge performed by neutral machines that claim to isolate objective reason from human complexity, interdependency, and emotion [121].

Our findings show how emotion AI hiring services position emotion AI as a technology with a “God’s eye view” derived from its alleged divine attributes of omnipresence, omnipotence, omniscience. In framing humans’ internal states and traits as isolable and immutable, emotion AI practitioners perpetuate a rationalist view of epistemology [359] that, as we show, claims emotion AI’s construction of knowledge about humans is universal, static, and objective. We argue that the perceived superior rationality and rightness of emotion AI’s neutral God’s eye “view from nowhere” [121] perpetuated by these claims rationalizes its “harmful artificial intelligence outcomes” [607] that operate by exclusion and dehumanization as rational [607, 643] and even “productive” [105] effects.

Yet, in a divergence from the rationalist Cartesian tradition of eliminating the taint of human feeling and emotion in the search for an absolute and objective truth about humans [257], our findings show how emotion AI practitioners explicitly seek to uncover human emotion as an essential finding necessary to objectively and absolutely “know” humans authentically. emotion AI hiring services position human emotion and affect as the elusive missing knowledge to the puzzle of objective and absolute truth about humans (in hiring), alleging that the conditions that had so far precluded possibilities to uncover this knowledge are now possible with emotion AI’s divine capabilities. Whether this departure is simply an attempt to reduce the complexity of human emotion to fit into the rationalist paradigm, or marks a turn in rationalist assumptions that renegotiate human emotion as the key to attaining absolute and objective knowledge, is an area for further inquiry.

Extending Birhane’s observations that the “God’s eye view” paradigm that dominates computational fields serves to excuse its practitioners from meaningful engagement with the technology’s ethicality behind a shield of value-neutrality, our findings show how emotion AI vendors suggest

this convenient effect transfers to the organizations that adopt the “God’s eye view” of emotion AI hiring services as technosolutions to the purported problems of hiring (in)accuracy, hiring (mis)fit, and hiring (in)authenticity. The privileged interests and concerns of emotion AI hiring services and the organizations for whom they build their product render these “knowers” [121] as ill-equipped to detect oppression and injustice associated with their technology under what D’Ignazio and Klein call the “privilege hazard” [265], in positions that stand to benefit from this ignorance.

Next, in interpreting our findings, we center jobseekers adversely affected by emotion AI to discuss the implications of our findings for design and policy.

4.5.1 Designing for Perceptible Fairness

While we do not advocate for emotion AI use in hiring and fully support regulation to limit its development and use, we recognize that emotion AI is already pervasive and deeply hidden in hiring services (which our challenging data collection process affirms). At the very least, we advocate for more transparent information sharing in emotion AI use, such as aligning with the Organisation for Economic Co-operation and Development’s (invaluable, albeit inadequate) [399] Fair Information Practice Principles (FIPP). As our findings show, information asymmetry operates as a mechanism by which services purport to solve the problem of hiring (mis)fits. We argue that emotion AI hiring services should reduce emotion AI-induced information asymmetry by: 1) designing for candidates’ transparent access to information generated about them, and 2) offering candidates the opportunity to correct inferences they believe to be inaccurate, challenging emotion AI services’ desire to discover “the truth” about candidates. Such approaches would facilitate candidates’ more meaningful participation in the hiring decision-making process, allowing them to reflect upon how their candidacy is evaluated based upon their affective expression in the job interview, and assess for themselves whether the job is a fit *for them*.

This suggestion does not address important implications of emotion AI hiring services’ use to shape social norms and exclude those that do not meet normative expectations of affective expression. However, introducing such a process would improve transparency and accountability of the entire ecosystem, offering visibility into the inferences generated about individuals, ways to assess its accuracy, and facilitate contestability [623] and reform.

4.5.2 Enforcing Fairness

More transparent processes may also be facilitated by existing regulatory frameworks, including Section 5(a) of the Federal Trade Commission Act (FTC Act) (15 USC §45) that “prohibits unfair or deceptive acts or practices in or affecting commerce” [717]. Indeed, the FTC recently released a memo of their new enforcement priorities, which commits to addressing concerns of deceptive

practices by nascent technologies that reinforce power asymmetries and the marginalization of communities by instituting timely interventions before deceptive practices lead to widespread harm [553].

Extending Stark and Hutson’s argument that “Physiognomic AI” is unfair and deceptive [803], our analysis shows how the claims emotion AI hiring services make in their technosolutions to the purported problems in hiring operate under unfair and deceptive methods that amplify power imbalances and perpetuate hiring harms, falling under the purview of the FTC’s priority goals. Moreover, Consumer Reports, in responding to OSTP’s Request for Information [16] on private and public sector use of biometric technologies (including those inferring emotional and cognitive states), recommends increased funding for the FTC to “go after and identify companies that are engaging with biometric-related pseudoscientific claims in the AI space” [749]. Our work provides much-needed empirical evidence for these policy suggestions which are clearly important to the OSTP. We advocate for enforcement action by the FTC as one possible avenue to stem the concerns of unfairness, power imbalance, and inequity that accompany the use of emotion AI hiring services in hiring. Below, we explicate the 1) unfair and 2) deceptive acts and practices identified in our analysis of emotion AI hiring service claims.

4.5.2.1 Unfairness in the Mechanisms of Emotion AI Hiring Services’ Technosolutions

As we have shown, emotion AI hiring services claim to solve hiring problems through the mechanisms of 1) informational and psycho-biological exclusion; and 2) creating, extracting, and commodifying affective value. These mechanisms perpetuate unfair organizational practices that may unethically enhance power asymmetries and promote exploitation.

Exclusion Hiring has emotional dimensions [725], despite what employers may wish to convey. Past work examining emotions’ roles in employment decisions in traditional settings shows that the candidates’ elicitation of positive feelings (i.e., excitement) in the interviewer “is a form of emotional capital that has economic conversion value” [725, 136]. Indeed, extant social psychology and organizational behavior studies have established that candidates’ emotions and emotional expressions not only shape their own behaviors during interviews, but also the interviewers’ and their subsequent evaluation of the candidate’ suitability for the job [807]. Positive tones and self-promotion tend to lead to favorable outcomes [806, 807] while candidates presenting as anxious or introverted do not tend to receive favorable outcomes [213]. It is not hard to imagine why these processes might be inequitable in traditional settings, especially for candidates who do not fit some normative expectation of affective presentation and expression.

We argue that through its mechanisms of informational and psycho-biological exclusion to purportedly make “accurate” and “true” hiring “fits,” the use of emotion AI hiring services unfairly

impedes the candidates' ability to negotiate "emotional capital" in the hiring process, exploiting workers with processes that simultaneously advantage organizations while disadvantaging workers. While this unfairness extends to all job candidates subject to emotion AI hiring services, it is disproportionately unfair to those whom emotion AI excludes for their perceived lack of "affective value." For those candidates that are selected by emotion AI hiring services, we argue these methods perpetuate the exploitation of human labor through the psycho-biological exclusion of hiring (mis)fits and normativizing the production of *loyal* hiring fits that "live and breathe" company values.

Affective Valuation Ahmed's concept of "disciplinary technologies" describes how powerful institutions use digital technologies to enforce the moral imperative of human "usefulness" by positioning the people subject to them as "potential" bodies and re-orienting them toward "useful" ends. Building on this work, Lin and Lindtner explore how the dominant "Techniques of Use" value system in HCI masks its associated harms, showing how the uncontested ideal of "usefulness" silences critical approaches in ways that reinforce and perpetuate injustice, exploitation, and exhaustion [534] in computing systems.

Applying these insights, we argue that under the ideal of "usefulness," emotion AI hiring services operate as a disciplinary technology through their creation, extraction and commodification of a candidate's *affective value*. Affective value, defined in 3.1.2, ranks and scores candidates based upon extraction of their affective expressions, according to measures of affective value developed by emotion AI hiring services and hiring organizations. *Affective commodification* turns affective value into a commodity. Processes of creating, extracting, and commodifying affective value ascribe a candidate's worth to emotion AI hiring services and hiring organizations, and reward those that meet the emotion AI hiring service and hiring organizations' expectations of affect. However, candidates are often unaware of how³ or whether they are subjected to emotion AI due to a lack of transparent application (as challenges in our data collection also illustrate).

Through opaque affective valuation processes that assess a candidate's desirability and usefulness to hiring organizations along emotional and affective dimensions, we argue that emotion AI hiring services unfairly and unethically promote the exploitation of human emotion. As a disciplinary technology, emotion AI hiring services orient not just human bodies, but human affect and emotion toward usefulness to the hiring organization. Moreover, we suggest that the ideal of usefulness that underlies the desired and promoted uses of emotion AI hiring services may obscure the harms associated with an uncontested belief in emotion AI's techno-supremacist capabilities to *usefully* solve organizational hiring problems by hiring only those deemed *useful* to the organization, and

³While candidates may be aware that their "soft skills" are measured by a hiring service, they may not be aware that inferences of their emotions, affect, and internal states are generated by the emotion AI hiring service to make such measurements.

hope to encourage more critical scholarship in this area.

4.5.2.2 Deception: Pseudoscientific Approaches Obscure and Perpetuate Hiring Harms, Not Mitigate

Emotion AI hiring services claim that their technologies 1) resolve bias in hiring, and 2) exclude hiring (mis)fits. We argue these claims deceptively obscure and perpetuate—rather than mitigate—hiring harms facilitated by pseudoscientific, physiognomic emotion AI.

Bias Emotion AI algorithms may be biased along racial, gender, and ability lines [719, 654]. For example, facial emotion recognition performs poorly on individuals with facial disfigurement, paralysis, or Down syndrome; blind or low vision individuals who may not make eye contact with the camera; and hard of hearing or deaf individuals who may struggle to hear questions [654]. As such, biased emotion AI algorithms may lead to discriminatory outcomes for minority groups.

As we have shown, emotion AI hiring services invoke larger discourses surrounding concerns of bias and discrimination in hiring [241, 351, 501, 325, 522, 632] to claim that emotion AI hiring services improve “accuracy” in hiring without bias. We note a surprising misalignment between established algorithmic bias concerns and scaled discrimination prevailing in emotion AI hiring services’ description of how their technology mitigates bias in hiring. emotion AI hiring services generally do not make claims of their technology’s technical capabilities to mitigate bias, but rather, claim that emotion AI inherently lacks bias by virtue of a machine, rather than a human laborer, assessing the candidate.

We argue that by emotion AI hiring services invoking these discourses with claims that their technology is unbiased, emotion AI hiring services deceptively suggest that their technology has resolved concerns associated with bias in AI-enabled hiring (i.e., through technical de-biasing methods [708]). The insufficient and shallow explanation that emotion AI hiring services provide as cover to claim that their technology is “unbiased” suggests that their emotion AI’s algorithmic bias remains unaddressed.

Moreover, we raise concerns about the reliance that hiring organizations may place in emotion AI hiring services as technical authorities, amplifying the deception in emotion AI hiring service claims that their technology is unbiased. Under the guise of adopting “unbiased” emotion AI hiring services, hiring organizations may divert their corporate attention away from bias in hiring by displacing corporate accountability over fair hiring practices to emotion AI hiring services [572] that in reality don’t solve for fairness in hiring, but instead obscure bias and perpetuate unfairness. In effect, this displacement of corporate responsibility for fair and unbiased hiring onto emotion AI hiring services may serve as an “excuse for why [hiring organizations] need not act” [675] or respond to harms associated with (and obscured by) emotion AI use in hiring.

Pseudoscientific, Physiognomic AI Emotion AI hiring services purport to solve problems in hiring with unfounded, pseudoscientific approaches [789]. Moreover, the ways in which emotion AI hiring services engage with discourses of discrimination is deeply concerning. As our findings reveal, emotion AI hiring services promise to solve the purported problem of hiring (mis)fits by adopting eugenic rhetoric and invoking racist, misogynist histories [548, 745] to *exclude* presumed hiring “(mis)fits” on the basis of their psycho-biological characteristics. By uncovering the problems emotion AI hiring services purport to solve and what values underpin their solutions, our findings provide much-needed *empirical, systematic* evidence for suggestions that 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 characteristic” such as that in emotion AI hiring services should be “declare[d] unfair and deceptive” [803].

4.6 Conclusion

This study reveals the ramifications of taking claims of privileged knowers at face value. By unpacking and interrogating the claims emotion AI hiring services ($n=229$) make in promoting their technology, we reveal how 1) the desired uses of emotion AI promoted by emotion AI hiring services are legitimized by their alignment with corporate ideals; 2) the mechanisms by which emotion AI hiring services claim to solve those problems unfairly exclude and exploit job candidates through the creation, extraction, and *affective commodification* of a candidates’ *affective value*; and 3) emotion AI hiring services promote beliefs in technology’s ultimate displacement and control of human labor by appealing to core values that emotion AI’s alleged supreme omnipresent, omnipotent, and omniscient attributes can and should be leveraged to the benefit of corporations. This work, we hope, helps enable the creation of more equitable and just futures of work by encouraging and facilitating discussion regarding the use of emotion AI hiring services within and beyond CSCW, better equipping us to consider human values in the continued deliberation about emotion AI’s ethical and responsible role in hiring, and make choices about the human values we want for our socio-technical futures.

CHAPTER 5

Emotion AI at Work: Emotional Privacy, Surveillance, and Autonomy¹

5.1 Introduction

Workplace surveillance is expanding to include automatic monitoring of worker emotion, mood, affect, and related constructs [925]. Emotion AI promises organizations the ability to better know, manage and monitor employees' interior states and traits in ways that support organizational goals, including improved productivity, mitigated security and safety risks, increased customer loyalty and sales, and improved corporate wellness [827, 826, 146, 862, 683, 408, 454, 813, 354]. By one industry estimate, 50% of US employers will use emotion AI to monitor their employees' mental wellbeing by 2024 [827].

Commercially available emotion AI-enabled enterprise systems feature diverse capabilities [139]. Some are fully extractive, whereby employees are surreptitiously subject to emotion monitoring as part of larger workforce analytics programs that collect, aggregate and process data from a variety of enterprise sources (i.e., digital communications, IT security infrastructure, wearable sensors, eye trackers, external social media, and geolocation data), and mined for insights into workers' interiority, including energy levels, wellbeing, sentiment, personal preference, opinion, and emotions [694]. Systems may be designed to make data accessible to organizational leadership (i.e., supervisors, department heads), while others may be more limited in scope and access. For example, IT security programs may use emotion inferences to screen for insider threats to workplace safety and security, with access to that data under tighter access controls [170]. More obtrusive forms of emotion monitoring include wearables that use bio-sensors and physiological

¹This chapter is based on: Kat Roemmich, Florian Schaub, and Nazanin Andalibi. 2023. Emotion AI at Work: Implications for Workplace Surveillance, Emotional Labor, and Emotional Privacy. In CHI '23: ACM Conference on Human Factors in Computing Systems, April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 20 pages. <https://doi.org/10.1145/3544548.3580950>. This material is based upon work supported by the University of Michigan's Rackham Graduate Student Research Grant and partially supported by the National Science Foundation under Grant No. 2020872.

signals that aim to infer employees' affective and emotional states in real-time, which may be implemented to influence worker behavior [637, 710]. Despite the increasing commercial availability and adoption of emotion AI in the workplace [694, 497, 827, 637], claims that emotion AI improves organizational outcomes [827, 826] are not scientifically well-established [734].

Privacy risks in applications of emotion AI may be particularly prevalent in the U.S. workplace, where employer surveillance practices perpetuate and reify social inequality [79, 788, 804], and workers' exposure to and interaction with emotion AI-enabled workplace monitoring may occur regularly [925]. Industry guidance suggests that organizations implementing emotion AI address their potential to internally exploit these flaws by adopting policies that reflect the “especially sensitive nature of this data and individuals' right to be free from emotional manipulation” and prohibit uses of emotion data that might induce “disadvantageous outcomes for workers” [826]. However, as both emotion AI applications and employer surveillance practices remain shielded behind the opacity of organizational operations, we lack 1) an empirical understanding of the implications of emotion AI-enabled workplace surveillance that foregrounds *workers'* perspectives—an important party directly impacted by emotion AI use in the workplace; and 2) legal protections or regulatory safeguards that enforce, recognize, or even define an individual right to privacy over emotions—ends to which our work contributes.

Acknowledging the inherent power asymmetry between organizations and workers, the perspectives of workers who are or may be subject to and impacted by emotion AI in their everyday work interactions is key to developing an understanding of the ethical and social implications of emotion AI in the workplace. Workers' location on the weaker end of the power spectrum render them best suited to identify its harms and injustices [669, 265, 121]. To this end, we conducted semi-structured interviews with US adult workers ($n=15$) to address the following questions: What are workers' general perceptions of emotion AI in the workplace (RQ1)? In what ways do workers experience or anticipate behavioral adaptations in response to emotion AI in the workplace (RQ2)? What consequences do workers experience or anticipate associated with emotion AI in the workplace (RQ3)?

We contribute four novel insights:

1. We contribute an understanding of workers' perceptions that emotion AI violates workers' *emotional privacy*, a term we introduce to describe privacy over one's emotions.
2. Drawing on the sociological theory of emotional labor [417], emotion AI in the workplace may function both as a tool to surveil employees' emotions and enforce workers' compliance with perceived expectations of emotional labor.
3. Workers may perform emotional labor as a way to preserve their emotional privacy.

4. Emotion AI-enabled workplace surveillance can expose workers to a range of harms including privacy and emotional labor-induced harms.

These findings demonstrate the need for critical attention to emotion AI's social and ethical implications, in and beyond the workplace, including research and policy on how to define, preserve, and protect emotional privacy. While we discuss implications for policy and design, we note that because many of emotion AI-enabled harms that we identify in this work cannot be mitigated through either technical or policy solutions, we advocate for approaches such as critical refusal [347] in the first place.

5.2 Background and Related Work

In this section, we review prior work on workplace surveillance, emotion AI at work, and emotion AI's adverse consequences that inform our study.

5.2.1 Surveilling Workers' Interiority

Surveillance of employees' interior states precedes today's digitally mediated surveillance practices. In the 1920s, employers started using surveys, interviews, and other methods to penetrate workers' "conscious barriers and [bring] out latent or unconscious sentiments," gaining insight into employees' thoughts, feelings, and emotions under the guise of improving the workplace [443]. As scientific and technological advancements grew, so too did employers' surveillance practices to probe employees' interiority. By the 1960s, employer use of psychological and personality tests to traverse borders between the worker as presented and the worker's psyche to reveal the "otherwise invisible inner man" was commonplace [378, 438], and by the late 1970s, employer use of lie detector tests (i.e., voice stress analyzers, psychological stress evaluators, and polygraphs) to identify employee deception was widespread [438].

Though the emergence of these increasingly invasive surveillance practices were met with public concern over employee privacy [228, 612, 798, 346, 75], US employers have by and large continued to expand their surveillance practices unrestrained, with electronic performance monitoring via methods including key stroke logging, computer screen capture, network logs, phone monitoring, and video surveillance emerging in the 1980s and continuing through today [79]. One notable exception constraining workplace surveillance is the passage of the Employee Polygraph Protection Act (EPPA) of 1988 that banned lie detector use by most private employers [75].

Advances around emotion recognition technologies have spurred employer desires to monitor and manage workers' interiority. Increasingly, employee monitoring practices have converged with aims to promote worker wellbeing. Intelligent systems promise to analyze enterprise data for

inferences of worker emotions and related affective constructs in efforts to advance goals including work-life balance and worker happiness [465, 163, 889]. However, worker emotions are influenced by the workplace context [381], and emotion recognition technologies fail to adequately account for the contingency of emotions to the workplace context [592, 464]. For example, a recent study using automatic emotion recognition methods to infer worker emotions suggests that leading emotion metrics (i.e., detecting dominant emotions) fail to consider the nuances of emotional expression at work and to accurately detect emotions in line with subjects' self-reports [464].

Given the expansion of workplace monitoring practices to include the detection and monitoring of workers' affective phenomena (i.e., emotion, mood, and core affect [282]), as well as the contextual sensitivity of these constructs to the workplace [375], our study investigates workers' general perceptions of emotion AI (RQ1).

5.2.2 Workplace Applications of Emotion AI

The purpose of workplace monitoring is not simply to monitor employees' behavior and activities, but to also *shape* them [29]. Through its alleged capabilities to automatically infer, analyze, and/or respond to workers' affective phenomena at scale, emotion AI-enabled workplace technologies promise to support organizations in better managing organizational outcomes by influencing employee emotion and related constructs [827, 826].

Workers' affective phenomena and organizational outcomes are mutually constitutive. As drivers of human behavior and decision-making [541, 909, 236], workers' emotions, moods, and affects influence organizational outcomes and events including sales [26], productivity [727, 103, 267], workplace violence [87], and insider threats [423]. Employer interest in shaping workers' affective phenomena to support organizational goals is underscored by the wide range of organizational purposes for which emotion AI in the workplace is adopted, including monitoring and managing workers' emotion, mood, and affect to detect and mitigate safety and compliance risks; monitor and improve employee wellness, productivity, and engagement levels; analyze and predict employee behavior; and automatically deliver real-time support, management, and coaching to employees [827, 826, 146, 862, 683, 408, 454, 813, 354].

Alongside this organizational interest, HCI researchers have designed context-specific applications of emotion AI for the workplace. For example, recent work has leveraged emotion AI to promote happiness and productivity in the workplace by mediating breaks [465], enhance communication with audiences during online presentations [624], and develop post-meeting feedback systems to improve meeting effectiveness and inclusivity [751].

Regardless of the use case, emotion AI in the workplace generates *emotion data* about workers—*inferences of workers' emotions, moods, affects, and other interior states and traits*—providing em-

ployers with information that they may leverage to inform organizational strategy, drive workforce decisions, and manage employees more precisely. Yet, little is known about how the collection and sharing of workers' inferred emotional information may impact worker behavior beyond shaping it to support organizational outcomes. This work contributes to addressing this gap by investigating workers' perceived behavioral changes in response to emotion AI in the workplace (RQ2).

5.2.3 Risks to Workers

While the potential benefits of emotion AI to organizations are well-established [827, 826], its effects on workers and associated social and privacy implications remain relatively unknown, though there is some indication that people have negative attitudes toward emotion AI. The importance of examining worker attitudes and perceptions is aligned with recent work suggesting that AI in the workplace can have negative effects on workers. For example, AI implementation can displace organizational responsibilities onto workers [619], and expose workers to unwanted monitoring and productivity management practices that risk employee privacy [343]. Notably, in a scenario-based survey regarding peoples' privacy attitudes toward video analytics technologies (one common source of emotion AI input data), Zhang et al. found that people were more uncomfortable and less willing to consent to video analytics that detect employee mood to predict productivity than their aggregated preferences across all surveyed scenarios [921]. Mantello et al. similarly found that job seekers have negative attitudes toward emotion AI in the workplace; their findings indicate that cultural background shapes attitudes toward emotion AI, suggesting that emotion AI may disproportionately induce stress and anxiety among workers of disadvantaged ethnicities, gender, and income classes [563].

To ensure fair treatment of workers, emotion AI technologies should be used fairly and ethically [826]. Fair and ethical use of emotion AI may include commitments by actors deploying emotion AI systems that it is meaningfully consented to [160]; that its (potentially biased, unreliable, and inaccurate [92, 226, 719, 654]) information is transparent and contestable [376, 826, 185]; and that its use does not widen power asymmetries, such as those already present between workers and their employers [35, 185]. Yet, so far, the use of emotion AI in workplaces remains largely unconstrained and unregulated, and in the modern US workplace, the growing adoption of emotion AI-enabled workplace surveillance is predicted to become the new norm [925]. In generating deeply private and sensitive emotion data that is prone to manipulation and misuse, emotion AI threatens the autonomy of its data subjects. Like with emotion AI's predecessors for workplace surveillance (Section 5.2.1), workplace conditions of weak worker power within the US [925] place workers in a position whereby they may be unable to meaningfully consent to—or protest—the inference and collection of their emotion data in the workplace [736, 322], even if they are aware of the practice

[79]. Indeed, US workers are provided with insufficient privacy protection in the workplace [736], and are thus particularly vulnerable to privacy harms posed by ubiquitous workplace monitoring that has expanded to surveillance of workers' emotion and affect [925]. Yet, we lack an understanding of the privacy implications of emotion AI in the workplace that is grounded in the experiences and perceptions of *workers*.

To understand emotion AI's privacy implications, we must first understand privacy theories. Altman's privacy regulation theory regards privacy as a temporal and dynamic process of regulating interpersonal boundaries with others to achieve one's (or one's social group's) desired privacy levels [46], compared against their actual privacy levels. Further refining Altman's theory, Petronio's communication privacy management theory (CPM) posits that such processes are underpinned by the belief that individuals own, and thus have a right to control flows over, their private information [689]. According to CPM, ownership of private information is shared with those whom information is shared, and privacy violations occur when rules regarding the management of that information are perceived to be broken [689]. Conversely, Nissenbaum's theory of contextual integrity (CI) posits that privacy is afforded with appropriate information flows, dictated by contextually specific norms; under CI, privacy violations occur when such norms are not followed [646]. Considering these privacy theories, emotion AI may implicate workers' privacy if the boundaries, rules, and norms around emotional information sharing in the workplace are ruptured, which may then expose workers to harms. Following Citron and Solove's privacy harms taxonomy (not specifically developed in the context of AI) [194], such harms may include physical, economic, reputational, psychological, autonomy, discrimination, and relationship harms.

Motivated by these gaps in knowledge about the privacy implications of emotion AI in the workplace, and informed by these privacy theories (as further described in Section 5.3), our study investigates emotion AI's risks of adverse consequences as perceived by the workers subject to and affected by emotion AI (RQ3).

5.3 Methods

We conducted semi-structured interviews ($n=15$) with adult workers in the US both with ($n=6$) and without ($n=9$) cognizant experience subject to emotion AI in their workplace.

Interview Protocol

We designed a semi-structured interview protocol with four phases. In phase 1, we established an understanding of the respondent's workplace and the monitoring tools in place. In phases 2 and 3, we covered respondents' anticipated or experienced responses to emotion AI in their workplace,

individually and to workers/the workplace as a whole. In phase 4, we asked privacy-related questions if the respondent had not yet mentioned privacy-related concerns. We designed the protocol to begin with general topics and questions, and then lead to more specific and sensitive topics to avoid influencing the participant's answers and to establish rapport to help facilitate disclosure. Interviews lasted approximately 90 minutes and participants received a \$35 honorarium for their participation.

Our sample included workers with and without cognizant experience subject to emotion AI, as emotion AI might be used without workers' explicit knowledge, a key challenge in studying this technology (i.e., existing data streams may feed into emotion AI without that being disclosed to the employee). Between the two groups, our protocol differed only in that for those without cognizance of emotion AI in their workplace, we used scenario-based interviewing grounded in an understanding of workers' general experience with employer monitoring established in phase 1. For example, if a participant without cognizance of emotion AI in their workplace indicated that their employer used video surveillance cameras, we asked the participant to imagine that those video cameras were equipped with computational capabilities to automatically detect, predict, and/or respond to their emotions, feelings, moods, and/or other internal states or traits. To avoid potential bias associated with terms such as "AI," "surveillance," or "mental health," these terms were not introduced to the respondent unless they used these terms first.

We conducted interviews between June and October 2021 and recorded them via Zoom video conferencing software. Participants who were uncomfortable with being recorded on video conducted an audio only interview. We used Zoom's live transcription feature to automatically transcribe interviews, then manually revised transcripts for accuracy before data analysis.

As an interview study on the sensitive and charged topic of emotions, AI surveillance, and the workplace, it was important to take steps to acknowledge and mitigate potential researcher bias and social desirability bias throughout the interview process. To avoid leading participants to respond in a negative way, we took particular care to ensure questions were asked in neutral ways, consciously avoiding prescribing meaning or assumptions upon the respondent by adopting participants' language (i.e. vocabulary choices) in follow-up questions, and when appropriate, repeating back our understanding to the respondent to confirm their agreement with our understanding of their responses [847]. In addition, we encouraged participants to respond with in-depth, narrative style responses, remaining flexible with the order of interview questions to follow the respondent's lead, a technique to reduce researcher bias [847, 345, 882]. By reserving potentially priming questions (i.e., privacy-related questions) for the end of the interview, and only asking those questions if the participants had not brought up those topics earlier in the interview first, we were able to stem researcher bias. A copy of our interview protocol is provided in Appendix B.2.

Sampling and Recruitment

To capture a wide range of worker perspectives about the emergent use of emotion AI in the workplace, we sought to gather participant experiences and backgrounds along dimensions of gender, race/ethnicity, age, industry/occupation, and cognizance of emotion AI in their workplace. Participation included workers both with ($n=6$) and without ($n=9$) cognizant experience subject to emotion AI in their workplace, denoted with alphanumeric codes of P_c and P_n, respectively. Table 5.1 includes participants' demographic information. Of note, occupations of participants with cognizance of emotion AI were predominantly public-facing roles (i.e., customer service representatives), suggesting these occupations may either be more likely to be subject to emotion AI, or simply more likely to be aware of it given the emotional demands of their occupation. Toward the end of data analysis, we identified no new themes and did not need to refine constructed theories, at which point we ended recruitment.

For sample diversity, we recruited participants from three sources: 1) occupation-related subreddits (i.e. r/supplychain), after gaining moderator approval; 2) the Prolific recruitment service; and 3) Facebook Ads. We solicited participants via an online recruitment message, which directed interested participants to a pre-screening survey to establish eligibility for the interview, determine cognizance of emotion AI at work, and gather demographic information to facilitate diverse participant selection.

We included a link to the pre-screening survey in our recruitment messages. The pre-screening survey collected information from interested respondents, including their cognizance of being subjected to emotion AI in the workplace, their demographic information (using best practices, i.e., [799]), various types of information collected and/or processed about them at work (i.e., information about what they look like, how they feel, their mental health state), the source of that data (i.e., phone, email, CCTV, microphones), and how that data was collected and/or processed (i.e., digitally recorded by the respondent in a self-report, automatically analyzed by a technological tool or device). To mitigate potential self-selection bias of those respondents highly concerned with workplace emotion AI, we recruited respondents aware of employer monitoring in general, rather than emotion AI specifically. We determined that those who indicated their employers inferred information about their internal states and/or traits automatically through a technological tool or device that inferred that information had cognizant experience with emotion AI. A copy of our pre-screening survey is provided in Appendix B.1.

We reached out to eligible respondents via email, which contained detailed information about our study's protocol and data management practices, and included a copy of our consent document. We asked eligible respondents to review the information provided and, if they wished to proceed, respond to schedule an interview. We obtained additional verbal consent from each participant at the beginning of each interview session and answered any questions they had.

Our institution's IRB determined this study exempt from oversight. Given the higher risk to which participants may have been exposed from participating in a study about their employer's practices [74, 814], we received IRB approval to classify our study under a higher tier to waive individual documentation requirements that otherwise would have provided our institution with information that could link participants' identities with participation in our study.

Participant	Gender	Age	Race/Ethnicity	Industry/Occupation
Pc1	woman	45-54	white	K-12 teacher
Pc2	man	35-44	Black/Latine	customer service representative
Pc3	man	25-34	Latine/white	customer service representative
Pc4	woman	18-24	Asian	research and development associate
Pn5	woman	45-54	Black	manufacturing team lead
Pc6	woman	35-44	Black	customer service representative
Pc7	woman	35-44	Black	healthcare aide
Pn8	woman	55-64	Black	K-12 teacher
Pn9	man	25-34	white	custodian
Pn10	man	35-44	white	insurance claims adjuster
Pn11	woman	35-44	white	social worker
Pn12	man	25-34	Latine	media services associate
Pn13	woman	25-34	white	audit manager
Pn14	man	45-54	white	immigration officer
Pn15	woman	25-34	white	K-12 staff

Table 5.1: Participant demographic table.

*Pc = *with cognizance of emotion AI*; Pn = *not cognizant of emotion AI*

Data Analysis

We imported de-identified interview transcripts and analytical memos written after each interview into NVivo, a qualitative data analysis software. Drawing upon grounded theory, the first author inductively analyzed interview data using interpretivist approaches to allow themes and patterns to emerge from the data rather than “imposing them prior to data collection and analysis” [681, 217], and met with the full research team weekly during the analysis for regular discussion and refinement of identified themes.

We initially open coded the data, ensuring developed codes remained close to the data and reflected participants' language and meaning [836, 182]. The first author took a line-by-line approach when open coding to help ensure a critical and focused analytic process and to identify actions, processes, gaps, and leads in the data to pursue [182]. The first author paid special attention

to respondents' language to create *in vivo* codes, thus grounding the analysis in participants' worlds and ensuring the analysis aligned with participants' meanings [182].

Following open coding of the first few interview transcripts, we began to identify themes. The first author triangulated the themes that emerged from interview transcripts with those noted in interview memos to create thematic codes according to the identified themes, then grouped existing open codes under the newly developed thematic codes. This exercise resulted in a hierarchically structured codebook with open codes organized by theme, which was then used to code the remaining data using a combined open coding and thematic coding approach. As data analysis continued, we scrutinized and refined emergent theories by constantly comparing newly analyzed data against thematic codes [811]. This method ensured open codes reflected member meaning, and could be regrouped as patterns and themes emerged, diverged, and were refined throughout the analysis.

Finally, the first author employed selective coding to organize thematic codes around a core concept of privacy perceptions, impacts, and harms and connected them to related concepts and theories [424]. General perceptions were strongly connected to privacy theories of contextual integrity [646], privacy regulation [46] and communication privacy management [689]. Perceived impacts codes were related to the sociological concept of emotional labor [417]. Perceived consequences codes closely resembled the typology of privacy harms recently introduced by Citron and Solove [194]; to facilitate scholarship clarity and consistency when identifying privacy harms, we chose to adopt the privacy harms typology and mapped harms codes accordingly where relevant. We did not set out to use these theories in our analysis to begin with, rather we observed that our initial analysis pointed to parallels in our analysis and these theories. Our findings' connection to these theories are summarized in each findings section.

Limitations

As an interview study, the standard limitations of self-report data apply. Additionally, many participants did not know whether they were subject to emotion AI at their workplace ($n=9$); we conducted scenario-based interviews with this group. Scenario-based (i.e., speculative) methods are sometimes criticized for their findings' construct validity and generalizability to real-life experience. As described in Section 5.3, we ensured that scenarios were grounded in participants' actual experiences with workplace monitoring and followed best practices. As our analysis revealed consistent thematic overlap between the two groups, our confidence in the validity of our findings remains high.

While this study does not aim for generalizability, the small sample size ($n=15$) and representation of job types is a limitation and as such our results may not generalize to workers broadly. Indeed, the

impact of emotion AI on some occupations, such as those not conventionally subject to management of their emotions, may be different from impacts identified in this work. Nonetheless, the fine-grained and in-depth nature of our interviews and subsequent analytic process allowed us to, rather than gaining validity through enumeration [233], provide generative insights regarding emotion AI's privacy implications in the workplace that are grounded in the experiences and perceptions of those who are or may be targeted and most impacted by this emerging technology, despite our study's small sample size. Future work could draw on these insights to examine workers' perspectives on emotion AI with larger sample sizes and other methods such as surveys, for example, to assess attitudes across identity lines and occupations.

5.4 Findings

We first describe the general perceptions of emotion AI in the workplace held by participants in our study, finding that (1) participants experienced and anticipated emotion AI in the workplace as a deep privacy intrusion that inappropriately probes private and sensitive information about their emotions, suggesting that emotion AI in the workplace breaches the contextual norms that govern the appropriate flow of emotional information in the workplace [647]. In describing participants' boundary management processes [689, 46] around whether and to what extent their emotional information is inferred and shared in their workplace, we (2) show how participants perceived emotion AI to violate these boundaries.

Second, our findings integrate the sociological concept of emotional labor [417] to show that (3) emotion AI-enabled workplace surveillance may function to enforce workers' compliance with emotional labor expectations and that (4) workers may engage in emotional labor as a mechanism to preserve privacy over their emotions, as indicated by participants.

Lastly, our analysis draws on participants' perceptions of and experiences with emotion AI in the workplace to (5) reveal how emotion AI-enabled workplace surveillance can expose workers to a wide range of harms on account of its emotional surveillance and enforcement of emotional labor.

5.4.1 Crossing Emotional Lines

The main theme across participants' perceptions regarding emotion AI encompassed privacy concerns. Our findings suggest that workers may reasonably expect that they have privacy to their emotions in the workplace, and establish how participants perceived emotion AI in the workplace to violate their privacy over their emotions.

5.4.1.1 Emotional Inferences are Inappropriate and Irrelevant to Employers

The predominant concern underlying participants' perceptions of emotion AI was the perceived *inappropriateness* of their employer digitally monitoring and algorithmically inferring workers' emotions and related affective constructs. Participants understood employers' attention to their outward expression as it relates to professionalism, but described how the use of emotion AI to monitor their outward expressions in order to infer their interior emotions was irrelevant and inappropriate.

For example, Pn12 did not want employers to infer his emotions and noted how what should matter to employers is job performance, not employees' emotions: "*Don't worry about how I feel, just let me do my job...if you're getting the output that you need, if I'm performing the way you need me to whether I [actually] feel bad, sad, good or happy, it shouldn't really make a difference.*" Pn12 emphasized that workers' inner emotions should not be of concern to employers, and questioned why the company even "*cares how I feel about XYZ as long as I'm working, I'm doing my job.*" Here, Pn12 establishes the perceived irrelevancy and inappropriateness of worker emotions to appropriate employer concerns. Echoing this point, Pn8 noted that detection of workers' emotions inappropriately exceeds the scope of the transactive relationship between workers and employers: "*because you pay me to work, you don't pay me to have conversations about how I'm feeling.*"

These perceptions of emotion AI's irrelevance and inappropriateness in the workplace suggest that emotion AI in the workplace may breach contextual norms regarding appropriate information sharing in the workplace—a violation of contextual integrity [646].

5.4.1.2 Emotion Data Sensitivity

Participants described how their emotions are not only private, but a particularly sensitive type of private information. Participants noted that the decision whether and to what extent to share their inner emotions should be an individual decision, and likened their emotions to components of their individual health and body.

As such, participants compared the emotion data generated by emotion AI to other sensitive information types, such as biometric and health data. Workers like Pn9 described how they view records of their emotions "*just like your medical information*" and that "*it should be kept private*" as such, while others like Pn11 suggested that emotion data "*should be regarded as like mental health information.*" Pn11 questioned the distinction between emotion data and mental health information, asking "*whether it be depression and anxiety, you know, so why is [emotion data] any different than those?*"

Given the perceived sensitivity of emotion data, participants perceived emotion AI's inference of emotions as an especially flagrant type of privacy intrusion. As Pn9 described it, use of emotion

AI to infer worker emotions is not simply a general violation of privacy, but “*a total invasion of your privacy, like in an acute way.*” These findings indicate that workers may perceive the emotion data that emotion AI generates as particularly private and sensitive, and expect that emotion data is handled in accordance with its heightened sensitivity.

5.4.1.3 Emotion AI Violates Boundaries Over Emotional Information

Participants described how conventional disclosure practices regarding how they felt at work were a personal choice that allowed them to control boundaries around whether and to what extent they shared how they felt with employers. Participants perceived emotion AI to traverse those boundaries and erode workers’ ability to manage their privacy over their emotional information.

For example, Pn11 compared emotion AI to employee feedback surveys that asked employees to share with their employers how they felt. Pn11 described how such self-reports were acceptable ways for employers to obtain this information as they preserved employee control over what and to what extent they shared their emotional information, but that using emotion AI to automatically infer what workers feel violates this personal boundary: “*If you want to ask me a question, and I choose to answer it, that’s fine. But to...basically put me under a microscope and see how I’m writing things, or how my body’s responding to different things [to infer that information]...I don’t like.*” Here, Pn11 highlights participant concerns around the automatic and continuous nature of emotion AI-enabled workplace surveillance.

Yet, participants’ concerns were not only how and to what degree they were monitored, but *what* was monitored—their emotions. Demarcating clearly between expressed and felt emotion, participants described how emotion AI inferring their emotions beyond whether and to what extent they choose to express them transgresses those boundaries. As Pn9 put it, emotion AI inferring their “*deeper*” felt emotions is akin to “*spying*” that crosses “*a huge privacy boundary.*” By traversing boundaries between expressed and felt emotion and bypassing workers’ ability to manage those boundaries, participants perceived emotion AI’s inferences as an intrusion of their interiority that extracts emotional information they perceived as inherently their own; as Pn11 put it, “*That’s mine. I don’t need someone monitoring that. It’s my information. It’s my emotions.*” Indeed, participants emphasized that the core issue at stake in inferring their emotions was not simply disclosing emotions they otherwise wanted to conceal, as if there were something to hide [794], but that it was problematic because it eroded workers’ autonomy to manage privacy over their emotional information. As explained by Pn8, even emotion AI’s inferences of a worker’s positive emotions can be troublesome: “*it could show that I’m really happy, that I enjoy what I’m doing. And I don’t know that anybody needs to know that either.*”

Participants’ perceptions indicate that the automatic, continuous, and intrusive nature of emotion AI-enabled workplace surveillance inferring information about workers’ interior emotions and

affect may be profoundly unsettling to workers. All together, they illustrate how workers' boundary management over the disclosure of their emotional information [46, 689] is circumvented by emotion AI's automatic inferences, and how those inferences may violate workers' desired privacy over their emotions by providing workers with an *actual* level of privacy over their emotions that is less than *desired* (see [45] for further detail about Altman's concept of actual and desired privacy).

5.4.2 Emotional Labor, Coerced and Claimed

Integrating the sociological concept of emotional labor—inducing and suppressing feelings to convey a particular emotion as required by their job [417], our findings of participants' anticipated and experienced behavioral responses to emotion AI suggest that it may operate as a surveillance tool that enforces workers' compliance with workplace expectations around workers' emotion management. In addition, our analysis of participants' perceptions and experiences finds that workers may engage in emotional labor [417] not only to comply with perceived expectations of their emotional expression, but also as an impression management strategy [363] that influences what others perceive them to feel while managing and preserving privacy over what is known about their emotions. As such, our findings suggest that workers may engage in emotional labor to preserve their privacy over their emotions, to the extent that the performance of emotional labor can afford.

5.4.2.1 Emotional Surveillance Enforces Emotional Labor Expectations

Participants with cognizant experience of emotion AI in their workplace characterized it as an emotional surveillance tool that enforced their compliance with workplace expectations of their emotional labor [417]. Offering an illustrative example, Pc6, a customer service representative, shared that if the emotion AI that monitored customer calls inferred that "*you're not perky enough,*" it would intervene by nudging the employee to induce more positive emotion: "*you get a whisper, 'Hey, we need you to smile more, you got this!'*"

Aware of the continuous monitoring of their emotions and enforcement of emotional labor expectations, but without visibility to what information is generated or how it is used, participants described how this information asymmetry enforced a constant expectation that workers convey a positive affect out of fear of how the emotion AI would detect their non-compliance with emotional labor expectations and, consequently, how its inferences could be used against them by their employers. As described by Pc7, emotion AI acts as an "*authority*" that holds workers "*liable*" to "*do [their] best*" and "*discipline*" them to "*obey the rules*"—including rules around emotion management.

Participant descriptions of the use of emotion AI to systematically monitor worker emotions

and enforce expectations of emotional labor provide support for an understanding of emotion AI as a tool that enables emotional surveillance [550]. These findings indicate that under emotion AI-enabled workplace surveillance and the information asymmetry it generates, workers may assume the need to constantly practice the emotional labor they perceive is expected of them.

5.4.2.2 Emotional Labor as Privacy Practice

Building on our findings established in Section 5.4.1.3 that emotion AI violates workers' privacy over their emotions, we find that workers may engage in emotional labor as a way to *preserve* privacy over their emotional information in response to emotion AI. Participants described how the emotional labor of inducing and suppressing their emotions at work protected them by allowing them to manage what and to what extent their employers knew about how they felt. Participants experienced and anticipated how emotion AI further erodes the privacy afforded by emotional labor through automatic inferences of their emotions. Thus, emotion AI not only enforces adherence to emotional labor expectations but simultaneously also penetrates workers' ability to use emotional labor to protect their interior emotions.

Participants with cognizant experience subject to emotion AI in their workplace described how they modified their emotional expressions in response to emotion AI-enabled workplace surveillance in order to convey a particular emotion readable to the machine. These participants shared how this practice was not simply to comply with perceived emotional labor expectations, but also to manage what information was inferred by the emotion AI and subsequently shared with their employers. P_{c1}, a teacher whose tone of voice and facial expressions during remote instruction were analyzed for emotion inferences as part of performance metrics, shared how emotion AI would reveal information to her employer that she did not want to share, such as disagreement with an automated lesson plan, as her expressions "*sometimes will say*" how she feels even if she chose not to explicitly express it. Consequently, P_{c1} shared how she had "*to really be in control of [her] facial expressions*" and vocal tone to avoid the emotion AI from inferring emotions such as stress or being upset (i.e., "*modify*" and "*lower*" her vocal tone). Experiences like P₁'s suggest that workers may manage their emotional expressions not simply to comply with workplace expectations of emotional labor, but also as a privacy behavior that utilizes the boundary between expressed and felt emotion to manage what is known about how they feel to their employers.

As such, participants anticipated how emotion AI's inferences would disrupt the preservation of privacy over their emotions afforded by emotional labor. For example, P_{n14}, an immigration officer for the federal government, described the "*mentally distressing*" emotional labor expectations of his job that required officers to "*grind it and just keep going*" when confronted with administrative demands that conflicted with their personal values. P_{n14} described how it was unsafe for officers to voice how they felt, and feared that if emotion AI were used in his workplace, it could expose him

and his fellow officers as employees that did not support the organizational changes (i.e., detecting officers that did not “*like the way it was being presented, or what was being laid down to us,*”) which in turn could jeopardize their employment.

These findings suggest that workers may engage in emotional labor practices of inducing and suppressing emotions not solely as a requirement of their occupation, *but also* as a mechanism to manage and maintain privacy over their emotions in order to maintain stability and security in their jobs. Through automatic and continuous monitoring practices that bypass the affordances of emotional labor for protecting privacy, emotion AI then can disrupt workers’ practices for managing their privacy over their emotions.

5.4.3 Beyond the Usual Harms

Participants experienced and anticipated how emotion AI in the workplace and its inferences of worker emotions exposes employees to a multitude of harms. Mapping our analysis to Citron and Solove’s general taxonomy of privacy harms [194], which was not developed specifically in the context of AI, we identify both parallels with this typology as well as emotional labor-induced harms expressed by our participants that the typology does not quite capture: amplification of emotional labor’s negative effects, disparate effects of emotional labor amplification, and chilling effects to workers’ own, felt emotions.

5.4.3.1 Privacy Harms

We first discuss how emotion AI implicates established privacy harms, in alignment with Citron and Solove’s privacy harm taxonomy [194].

Psychological Harm. Psychological harms refer to negative mental responses experienced as a result of privacy violations [194]. Participants shared how the practice of emotion AI-enabled surveillance can induce emotional disturbance and distress, harming workers’ psychological well-being with negative effects including worry, stress, and paranoia.

Pc3, whose call center analyzed recordings from employees’ web cameras to monitor their emotions, shared how he maintained “*a sense of...worrying*” throughout his experiences being subject to emotion AI. Pn15, who did not have cognizant experience with emotion AI in particular but did have experience with her employer maintaining digital records of observed employee emotions, described how if she was aware that she was subject to emotion AI, it would be “*very stressful, and it would make it so that the only place I could really relax is outside of work...and I would have felt very unhappy at the workplace.*” Similarly, Pn10 anticipated that “*if [he] knew it was happening, [he] would be a bit paranoid*” and Pn11 noted that she “*would feel like [she’s]*

under a microscope, like people are watching" which would "*put [her] back on guard.*" These examples illustrate how emotion AI's surveillance itself can result in direct harms to workers' psychological wellbeing.

Autonomy Harm. Autonomy harms involve constraints on people's freedom to make choices [194]. In line with findings from Section 5.4.1.3 that emotion AI violates workers' privacy over their emotional information, participants emphasized how being subjected to emotion AI would acutely harm their autonomy by automatically extracting and sharing inherently personal information about their emotions, which could expose them to emotional manipulation by their employers. Moreover, participants shared how they perceived employer efforts to obtain consent to emotion AI as coercive, suggesting that standard employer monitoring consent practices (i.e., asking an employee to sign a notice consenting to emotion AI) may be perceived as coercive, and should not be viewed as worker consent to the privacy violations imposed by emotion AI in the workplace.

For example, Pn9 described their emotional information as deeply personal, and believed that individuals alone should have the ability to exercise choice in sharing it. Pn9 stated that "*I think it should be up to your own person to decide what information...about your health and body*" is shared, and that the decision to share that information should be decided "*not [by] your employer...or anyone else.*" Pn9 exemplifies participant perceptions that in eroding workers' privacy over their emotional information, emotion AI can harm workers' autonomy over when and how they share their emotions.

While obtaining consent for emotion AI to collect or infer workers' emotional information may arguably mitigate its autonomy harms, our findings suggest that this may be insufficient as it may be perceived as coercive rather than freely given consent. Of note, Pc3 was the only participant with cognizance of emotion AI in their workplace who noted their employer sought their consent, specifically to use "*camera tracking*" to monitor call center workers' emotions. Pc3 found this to be coercive, as employees felt obligated to sign the consent document because their job was on the line. Pc3 explained that "*everyone just felt obliged because it was an all-in-or-nothing sort of situation...everyone, if they wanted to keep their employment, they had to sign that document.*" Underscoring the coercive nature of seeking consent to emotion AI-enabled workplace surveillance, Pc3 shared that a coworker had to leave the organization because "*they didn't sign the document on their own accord.*"

Our findings suggest that the dissemination of workers' emotional information may leave workers vulnerable to emotional manipulation by their employers. For example, Pn12 anticipated how the use of emotion AI would indirectly manipulate workers to "*think a lot more...company-oriented things*" once awareness of the emotion monitoring grew. Yet, employers may use this information to directly influence workers' emotions as well. Pc1 reported that her employer used emotion infer-

ences and metrics to “coach” teachers by informing them that they weren’t expressing themselves “*the right way*” and warn that they “*might not get rehired*” if teachers did not embody the emotional expectations their employer demanded. Demonstrating how workers’ emotional information can expose workers to emotional manipulation, Pc1 reported that their employer would use emotion data to influence teachers to feel how the district wanted them to feel: “*That’s not how you should be feeling about [your lesson plans]. This is the way you should be approaching this. This is the way you should think.*”

By denying workers the ability to control what is known about their felt emotions and in a context where workers do not have a free choice to consent to the practice, emotion AI-enabled workplace surveillance harms workers’ autonomy by coercing workers to relinquish control over their private emotions to their employer. In addition, it poses a risk of future harm to workers’ autonomy by revealing emotional information that employers can then use to manipulate workers into aligning their feelings with the interests of the organization. Importantly, these effects of introducing emotion AI are happening regardless of emotion AI’s precision in recognizing emotions, a point we discuss further in Section 5.5.1.

Physical Harm. Physical harms characterize privacy violations that injure one’s body [194]. Participants described how the stresses and psychological harms of emotion AI collecting and sharing information about workers’ emotions can manifest physically, injuring workers’ physical wellbeing.

For example, participants with cognizant experience with emotion AI described how it can deplete workers of physical energy and vitality. As illustrated by Pc6, being subject to emotion AI “*drains the snot out of [her]*.” Likewise, Pc3 explained that “*it takes away from people’s energy that could be used towards more productive things for both themselves and the company while working.*” These examples illustrate how emotion AI can physically harm workers by stripping them of physical energy. What’s more, this effect may impair worker productivity, which may pose an economic risk of harm to employees as well as employers.

Noting the close relationship between emotional and physical health, Pn8 anticipated how being required to use emotion AI at her workplace would just make her angry, which could in turn impair her physical wellness: “*You have a piece of equipment on me, that can tell people that I’m angry about something, annoyed about something, probably more anger, because my blood pressure will probably go up.*” Pn8’s observation highlights how the physiological responses to emotion monitoring can adversely impact one’s physical wellness. Even if those changes are temporary (i.e., temporary blood pressure spikes), they can lead to longer term consequences (i.e., organ damage [560, 380]).

Economic Harm. Economic harms are the result of privacy violations that lead to monetary loss [194]. Participants described experiences and concerns related to economic harms resulting from the processing of their emotional information, as the revealed information may hinder future job opportunity or result in job loss. Particularly, participants were concerned that emotional information inferred by emotion AI could be used to make employment decisions or to justify performance evaluation decisions—upon which raises, promotions, and bonuses often depend.

Illustrating how using emotion data in performance evaluations can economically harm workers, Pc3 described how a colleague’s performance review, which included metrics aggregated from video-based emotion tracking along with other data sources to infer employee satisfaction and engagement, suggested that the employee was not satisfied with their job. As a result, Pc3 explained that management then began to doubt whether the employee was “*up to the role*,”. Pc3 expressed disdain for his employer “*questioning a person’s ability to continue [the job] based on...minimal information*” derived from emotion AI inferences, threatening workers’ job security. In addition, workers shared concern that use of emotion AI’s inferences in performance reviews could result in the loss of economic opportunity, such as denying a promotion or raise. For example, Pn11 worried their emotion data would lead to a poor performance review and pass them over for a potential promotion, on the grounds that “*I wasn’t necessarily happy or something like that*.” Participants’ shared experiences and concerns suggest that certain uses of emotion AI (i.e., in performance evaluations) can expose workers to economic harm.

Reputational Harm. Reputational harms involve injuries to one’s reputation or standing [194]. Touching on concerns about emotion AI’s reliability and validity, participants reported that inferences of felt emotion are invalid and unreliable to assess how employees feel due to the high variation of emotions experienced in the workplace, the indistinguishability of emotions felt about work from other contexts, and technical inaccuracy. Participants expressed concern about consequences to their reputation as a result of misleading or inaccurate emotion AI inferences.

Pn10 described a recent example where his felt emotions varied significantly throughout the week, “*feeli[ing] very angry and concerned and just paranoid*” at the beginning of the week due to a higher than usual workload, but felt “*very happy*” by the end of the week as he “*got everything caught up*,” ending the week feeling accomplished. Pn10 highlights here how workers can experience felt emotions more deeply and extreme than they express them, an emotional phenomena that can be attributed to one’s care for consequences [341]. By conflating workers’ felt emotion with its modulated emotional expression, Pn10 worried that the “*extremes that you would get*” could confer a misleading impression of one’s overall emotional wellness to their employer.

What’s more, Pn10 worried that the blurred boundaries between the personal and the professional would render emotion AI’s inferences about workers’ emotional lives at work indistinguishable from

their personal ones [375]. Pn10 emphasized that emotions felt while at work are often related to private life events rather than work concerns, such as recent “*bad news about a family member*” or upset at something relatively “*dumb*” like the cancellation of a favorite TV character, raising concern that the inferred emotional information may give his employer the wrong impression of how he feels as only “*some of [his] emotional responses are going to be work related.*”

In addition, participants shared concerns that emotion AI’s technical inaccuracies may create a false impression about workers. As a supervisor at a production facility with workplace hazards (i.e., pneumatic air and dangerous machinery), Pn5 acknowledged how emotion AI could improve workplace safety (i.e., detecting fatigue to reduce workplace accidents), yet remained concerned about emotion AI’s potential to injure an employee’s reputation as a result of potentially inaccurate inferences. Referring to her personal concerns, Pn5 reported that she doesn’t “*have the most friendliest face,*” describing that she could feel “*happy as I don’t know what,*” yet others may misread her face as “*stoic...or upset.*” Given her experience with others misreading her emotions from her facial expressions, Pn5 was concerned the emotion AI would as well: “*I wouldn’t want it misreading....if the human can do it, then I know a piece of technology could do it, so that’s not cool in my opinion.*” Consequently, Pn5 was concerned of what “*everybody would think of [her]*” if the emotion AI continued to misread her emotions negatively. Marking the significant difference between felt emotion and expressed emotion, Pn5 also shared concerns that detecting felt emotion would lead to unreliable and invalid predictions about workers: “*I’m so mad I want to shoot someone. So that don’t mean I’m gonna go ahead and do it.*” Describing the effects inaccurate inferences would have on workers as “*probably [her] biggest fear,*” Pn5 expressed concern that emotion AI’s inaccuracy could unfairly harm workers’ reputation in the workplace, and worried about what other potential consequences this might entail for workers: “*will it spill over? ...what’s the consequence behind how you feeling?*”

In addition to reputational harms, Pn5’s concerns raise important implications for employer liability, as employers may be compelled to act on certain inferences (i.e., anger) so they are not held liable for negligence in case that person threatens workplace safety and/or security (i.e., inflicts violence). As the algorithmic detection of anger has been shown to exhibit racialized bias [719, 429], employer interventions could involve unjust actions taken against workers of color erroneously detected as angry that not only harm a workers’ reputation, but as recent scholarship has observed, potentially expose them to dangerous interactions with law enforcement as well [731].

Participants’ insights illustrate how emotion inferences are likely a poor construct to assess employee wellness, which can mislead others to have a false impression of workers and unfairly harm workers’ reputation. In addition, they suggest that the detection of some affective phenomena (i.e., fatigue) carry different risk profiles than others (i.e., anger), which may expose workers to additional harms (i.e., discrimination and economic harms).

Relationship Harm. Relationship harms concern injury to personal and professional relationships [194]. Participants shared experiences and concerns with how emotion AI in the workplace can damage trust and amplify tension between employers and employees and limit the capacity for workers to engage with and support each other, injuring professional relationships between and amongst workers and their employer.

Participants reported how they perceived the organizational decision to implement emotion AI in the workplace as a suggestion that their employer does not trust them. Pc3 described how the implementation of emotion AI in their workplace fostered “*a sense of distrust*” and “*disconnect between [workers] and [their employer]*.” Similarly, Pc7 shared that after emotion AI was introduced, she and her colleagues immediately wondered, “*why is the organization not trusting us?*” As a consequence, participants shared that this sense of distrust would damage the professional relationship between workers and employers. For example, Pn12 shared that they “*would probably feel disregarded*” by their employer if they were to implement emotion AI in their workplace, and anticipated how “*a lot of people...would probably be really put off by the fact that a company is willing to roll something out...that kind of privacy violation tool.*”

In addition, participants indicated that the decision to adopt emotion AI could amplify pre-existing tensions between workers and employers. For example, Pn11 perceived emotion AI in the workplace as an inauthentic way to promote wellness that, in effect, shifted the employers’ responsibility to manage a workplace environment that is conducive to worker wellbeing onto individual workers. Likening emotion AI to employee wellness initiatives (i.e., encouraging workers to practice self-care), Pn11 underscored the hypocrisy of employers that “*drive [workers] for profits*” using emotion AI to promote an “*individual responsibility to take care of yourself*” instead of addressing underlying workplace conditions that can impair workers’ wellbeing as a “*whole disconnect...that doesn’t really line up for [her]*.” Pn11’s observations suggest that worker responses to the implementation of emotion AI—even when presented positively as a way to promote wellness—can exacerbate already present tensions in the employer-employee relationship regarding employee wellness.

Moreover, participants shared how emotion AI could constrain relationships between workers. As Pc3 described, “*everyone always complains about it...how ridiculous it is,*” but that they had to do so carefully. Pc3 explained that workers were careful to only bring up concerns with each other in-person “*when just having conversation*” so that their concerns were not digitally recorded or inferred by the organization. Moreover, Pc3 described how his boss would sometimes hear their conversations, but would “*remain neutral*” as their boss was not in a position to advocate employees’ concerns. Pc3’s experience suggests that emotion AI-enabled workplace surveillance may damage the professional relationship among workers as well, by limiting workers’ capacity to support and engage with each other, and potentially suppress dissent among them.

Discrimination Harm. Discrimination harms perpetuate social inequalities of disadvantaged groupin ways that leave “a searing wound of stigma, shame, and loss of esteem...knowing that one is viewed as less than human, as not worthy of respect” [194]. Participants described experiences and perceptions of how emotion AI-enabled workplace surveillance can perpetuate and obscure gender-based discrimination in the workplace.

For instance, Pc7 described how her colleague experienced negative emotions related to her pregnancy, explaining how “*pregnancy comes with...so many things going on around the body*” that can negatively affect how one feels while at work. Pc7’s colleague had not yet disclosed her pregnancy to their employer, so when their employer expressed concern about her negative emotions and the “*mistakes*” she made by failing to engage with patients warmly enough, the colleague felt “*forced*” to disclose her pregnancy to explain away the emotion AI’s inferences about her negative emotional state. The unwanted disclosure of pregnancy to their employer that Pc7’s colleague felt forced to reveal as a consequence of emotion AI-enabled workplace surveillance ultimately gave their employer a way to evade anti-discrimination requirements. Instead of modifying their expectations to accommodate the employee’s pregnancy, their employer tied emotional expression to work performance (i.e., compliance with emotional labor expectations) and eventually gave the colleague a choice to either “*quit their job, or improve*” the negative emotions they experienced as part of their pregnancy that manifested in their interactions with patients. Describing the difficulty her colleague experienced in attempting to manage her pregnancy-related negative emotions how their employer expected, particularly when subject to emotion AI-enabled workplace surveillance, Pc7 explained that “*once she realized that [emotion monitoring] was going on...it kind of like changed her attitude in a way, because now you are acting under force, and pressure.*” Though Pc7 indicated that her colleague “*really tried her best*” to improve, the colleague ultimately had to leave the organization. This example suggests that emotion AI can harm workers by inducing disclosure about private matters (e.g., pregnancy) that may then be used by employers to justify discriminatory practices.

Underscoring the concerning potential for emotion AI-enabled workplace surveillance to perpetuate and obscure discrimination, Pn13, a manager, anticipated how emotion AI could be beneficial to her organization by affording managers information about employees that could be used to justify employment decisions that otherwise lacked documented support. For example, Pn13 described “*a situation a couple of years ago where we had to terminate a [female] employee, and it was without cause,*” noting that emotion AI could be useful to employers in similar situations. Pn13 shared that it would be useful for “*IT management use it on an as-needed basis*” because it would offer employers “*concrete data*” to “*build a case*” against a worker they wished to terminate (who otherwise would have been fired without cause). Explaining further, Pn13, a woman herself, shared that “*females are stereotyped to have more emotion*” and that women “*need to, you know, keep*

your emotions out of the workplace.” Pn13 described her “*negative experiences*” as a manager working with womens’ emotions in the workplace, such as “*disagreeing with a manager, and not wanting to do what they ask, resulting in storming off.*” Pn13 thought emotion AI-enabled workplace surveillance could be particularly beneficial to the organization if it could detect “*emotions in the workplace from females that were extreme, and over the top and inappropriate.*” P13’s remarks here demonstrate the stigma surrounding womens’ emotionality in the workplace, and the eagerness employers may have in adopting emotion AI-enabled surveillance systems that afford employers information they can wield to legitimize otherwise risky employment decisions (i.e., firing a woman without cause) and potentially shield them from discrimination claims.

5.4.3.2 Emotional Labor Harms

While many of the harms experienced and anticipated by participants align with Citron and Solove’s privacy harms taxonomy as discussed in Section 5.4.3.1, emotion AI and its interaction with emotional labor also surfaces harms that exhibit nuanced qualities that do not neatly align with the taxonomy. We identify three harmful aspects to emotion AI as a surveillance mechanism to enforce emotional labor: (1) enhanced enforcement of compliance with emotional labor amplifies emotional labor’s negative effects; (2) negative effects of emotional labor disproportionately endured by workers of marginalized identities and backgrounds (i.e., Black women as presented in our sample); (3) chilling effects to workers’ own, felt emotions.

Emotion AI Amplifies Emotional Labor’s Negative Effects on the Worker Participants described how the automatic, continuous emotion monitoring provided by emotion AI worsened, or could worsen, the adverse impact to their wellbeing they already experienced from the emotional labor they performed at work through constant discipline and enforcement of emotion rules, in effect amplifying these known negative effects of emotional labor [417] that are only partially recognized by the privacy harms taxonomy [194].

For example, Pn11 anticipated how emotion AI’s emotional surveillance would heighten the emotional labor they already practiced as a mental healthcare provider. Pn11 noted how difficult it would be to continue to express care and concern for her clients under emotion AI-enabled workplace surveillance: “*rather than being present with my clients, so I wouldn’t not only have to watch my emotions and my reactions, and also still be present for the clients, but then I would have to also be on guard to whatever this technology is trying to infer about me.*” Here, Pn11 highlights how both emotion AI’s enforcement of emotional labor expectations and emotion AI’s surveillance of worker emotions can amplify the already difficult performance of emotional labor and associated negative effects, in effect harming workers’ wellbeing, but also divorcing workers from their own emotional experience.

For participants, the negative psychological effects of continuously complying with emotional labor expectations under emotion AI-enabled surveillance carried a deeper quality than psychological disturbance and distress, leading to a sense of alienation that can estrange workers from their own selves and those around them [417, 580]. For example, Pc6 shared how the distress of being subject to emotion AI's constant emotional surveillance and emotional labor enforcement inducing feelings like hopelessness and fear reduced her sense of purpose to datified performance indicators: "*I'm like getting nowhere, that all of this stuff is counted against my metrics.*" Likewise, Pn15 worried about the self-estrangement that could emerge from being subject to emotion AI, as it would prevent her from "*being able to be [her] full self.*" Pn15 described how she "*would have been disappointed*" in herself for suppressing who she was and how she felt.

In summary, emotion AI's automatic surveillance of worker emotions affords employers the continuous, perfect enforcement of emotional labor, which can amplify its negative effects to workers' wellbeing. While this harm shares similarities to psychological and possibly physical privacy harms [194], it entails harms of worker alienation and self-estrangement that are amplified as a result of emotion AI-enabled workplace surveillance's enforcement of emotional labor compliance that are not captured by Citron and Solove's typology. Indeed, the experience of estrangement from one's own private self and emotions is an "important occupational hazard, because it is through our feelings that we are connected with those around us" [417].

Disparate Effects of Emotion AI's Emotional Labor Enforcement Our findings suggest the negative effects workers may experience under emotion AI-enabled workplace surveillance as an emotional labor enforcement tool may be disproportionately felt by workers of marginalized identities and backgrounds. In particular, the experiences of Black women with emotion AI in their workplaces suggests that the negative effects from its use as an emotional labor enforcement tool may be more severe for Black women, who disproportionately endure challenging customer interactions as doubly women *and* workers of color [230]. While emotion AI can amplify this discrimination harm [194], its interaction with emotional labor involves a nuanced effect whereby workers may disproportionately endure emotional labor to *confront* the discrimination that harms them.

Pc6 described how her employer monitored her video and call-based interactions with customers in real-time to ensure that workers "*stay upbeat and make [them] really be positive and energetic through the whole conversation.*" Pc6 reported that this expectation was enforced even in the face of challenging interactions, which for Pc6 included racist and sexist customers who met her with disdain and sometimes even refused her support upon recognizing her identity as a Black woman. Describing the distress of having to provide support to these customers, Pc6 shared how difficult it was to maintain positivity "*when your insides are crying because of the poor, poor attitudes*

that you have to deal with all day,” knowing that their emotions were monitored to make sure of it. Pc7, a Black woman and healthcare aide whose employer similarly used real-time video and audio-based emotion analytics to monitor interactions with patients, reported similar distress from enduring emotional surveillance in the face of racist customer interactions. Pc7 shared that “*there’s also some patients who don’t like Blacks...so they will like insult you, they’ll treat you badly*”; though Pc7 would always “*try [her] best*” to convey positivity and make the patient happy, she described how sometimes it was too much to endure when “*you cannot take it anymore.*”

Both Pc7 and Pc6 described how they made sense of their experiences enduring emotional labor as ways to challenge racism, spinning them in a positive light. For example, Pc6 shared that even if she had “*someone that’s racist, I want to provide the best experience ever so that I can make you change your viewpoint on how you feel about someone of my complexion*” and “*change the narrative that your experience with a Black person was the best that you have had in a long time.*” Similarly, Pc7 described how maintaining calmness and positivity toward difficult patients could challenge patient prejudice: by refusing to respond to racism and contempt with anger, Pc7 believed that she “*chose to do the right thing*” by concealing the negative emotions that such racist encounters provoke, allowing her to “*be the bigger person.*” Such sense-making processes demonstrate the additional burdens and consequent discriminatory effects Black women and possibly other workers of color may take on in order to reproduce the constant positive emotional labor required of their jobs under emotion AI-enabled workplace surveillance.

Harms from emotion AI’s disparate negative effects from emotional labor enforcement share similarities to established discrimination privacy harms [194] in that they may disproportionately affect workers of marginalized identities and backgrounds, yet differ in that it does not create the same mark of shame and stigma. Pc6 and Pc7’s experiences instead reveal how they perform emotional labor to *challenge* societal prejudices and their stigmatized associations. The disparate effects workers may experience from emotion AI then stem from the additional labor marginalized workers disproportionately endure on account of societal discrimination.

Emotional Surveillance’s Chilling Effects on Felt Emotion Concerned that emotion AI could detect that the emotions they outwardly expressed in accordance with their job’s emotional expectations did not align with their inner, felt emotions, participants with cognizance of emotion AI-enabled workplace surveillance experienced chilling effects to their own felt emotions in order to align their emotions with perceived workplace emotional expectations. More than amplifying constraints to workers’ autonomy and the psychological harms this restriction may involve [194], we find these chilling effects to workers’ felt emotion to involve concerns that may be ignored by a categorization that insufficiently captures the complexities of human emotion that include, but also exceed, limits to free choice and rational thought [341].

For example, Pc1 described how the continuous emotional surveillance and emotional labor enforcement they experienced under emotion AI prevented her from *experiencing*, not just displaying, human, negative emotions. Pc1 shared that under constant emotion monitoring to enforce expectations that teachers maintain a positive demeanor, she felt she was not even allowed to experience negative emotions while at work—regardless of how she expressed them outwardly. Contextualizing her experience as a high school teacher, Pc1 shared examples of everyday interactions that would reasonably induce negative feelings: “*teenagers, they’re going to try to tell you that you look fat one day, or they’re gonna...ask if you have a boyfriend, or they’re going to tell you that their mom is younger than you.*” Pc1 explained how these difficult interactions “*push you to learn how to handle [them]*” and not visibly “*get angry.*” But, under emotional surveillance and emotional labor enforcement, “*if you did get a little heated one day and have a bad day, definitely you would be investigated.*” As a result, Pc1 found it difficult to not even be able to *feel* negative emotion, out of fear her employer would investigate her as a result. Similarly, Pc7 shared how she was unable to feel certain emotions as a result of her employer’s emotion AI-enabled emotional surveillance, describing how the “*pressure [of] wanting to feel something that is outside the organization, or just something that you are just by yourself,*” but couldn’t, was “*overwhelming*” due to the “*constraining*” effects of emotional surveillance.

These experiences demonstrate how emotion AI-enabled workplace surveillance can chill worker autonomy over their inner, felt emotions. This harm extends beyond established definitions of autonomy harm [194] as the point of contention goes further than concerns of undermining peoples’ choices and restricting lawful human behavior, rather it involves manipulating and re-orienting worker affect and emotions in ways that limit the bounds of human emotional life.

5.5 Discussion

Emotion AI is often celebrated for its potential to improve the safety and culture of organizations and the wellbeing of the employees that compose them [694, 827]. Yet, our examination of workers’ perceptions of and experiences with emotion AI illustrates a starkly different story: one where workers are subject to invasive emotional surveillance that enhances the control employers have over workers’ emotional lives [247, 58, 417] and amplifies the adverse consequences workers may experience from emotional labor enforcement and privacy intrusion. Even in the increasingly privacy-invasive modern workplace [58, 925], we find that participants perceived emotion AI to enable an especially intolerable form of surveillance that erodes workers’ privacy and control over their own emotions. Employers’ unrestrained ability to monitor and manipulate their employees’ emotions with emotion AI-enabled workplace surveillance threatens to degrade the value of and shift social norms around privacy at perhaps the most fundamental level of human experience:

what we refer to as *emotional privacy*.

Our findings call for industry, policy, and research to contend with emotion AI’s erosion of emotional privacy. To that end, we first discuss our conceptual contribution of emotional privacy to illustrate how emotion AI destabilizes privacy over one’s emotional life, and argue that emotional information and freedom from emotional manipulation are worthy of preservation and protection—within and beyond the workplace. We conclude with implications of our findings regarding emotional privacy for policy and design.

5.5.1 Naming Emotional Privacy

Documenting how employers engage in surveillance practices to monitor and manage employee emotions, Arlie Hochschild introduced the sociological concept of “emotional labor” in 1979 to describe the phenomenon of corporate control and commodification of workers’ emotions. Hochschild’s arguments proved to be politically potent [147] and were followed by an impressive breadth of scholarship that largely focused upon emotional labor’s adverse effects [499]. Yet, less attention has been paid to the privacy implications of emotional labor, which Hochschild referred to as “the best account of how deep institutions can go into an individual’s emotional life while apparently honoring the worker’s right to ‘privacy’ ” [417].

Hochschild depicts the interiority that remains deep inside workers as an “inner jewel” that evades the gaze of even the most authoritative employer [417]. As our findings suggest, emotional labor can function as a mechanism to manage and preserve one’s privacy over this inner jewel, yet, emotion AI that automatically infers workers’ emotions enables employers to break this shield and access the inner jewel of workers’ interiority. In so doing, as our study finds, emotion AI erodes peoples’ ability to preserve the privacy of their emotions—what we refer to as their *emotional privacy*—restricting whether and to what extent people can manage what is known about their emotions to others by transgressing human boundaries between expressed and felt emotion. Throughout this paper, we show how emotion AI use can disrupt this desired quality for many workers, how workers attempt to manage their emotional privacy through emotional labor, and why emotional privacy is consequential due to the harms its invasion imposes on workers. Emotional privacy has implications beyond the workplace, as emotion AI technologies and applications span many contextual use cases, including healthcare, education, marketing, and law enforcement [601]. The breadth of scholarship aiming to improve the algorithmic detection of “fake” and “genuine” emotions [279, 873, 530, 470] highlights emotion AI’s threat to emotional privacy.

By exposing and manipulating human emotion, as our findings suggest, the consequences of this emerging technology’s privacy harms add a new quality to the current recognition of digital privacy harms [194]. While emotion AI-enabled workplace surveillance has much in common

with other surveillance infrastructures, our findings suggest that there is a different, deeper level of quality to its privacy invasiveness. Emotion AI-enabled workplace surveillance constitutes a deeper privacy intrusion into a person’s interior—surveilling and manipulating humans’ emotional selves and bodily interiority—than is the case with prior surveillance infrastructures that mostly monitor outward display acts. Regardless of its current technological limitations [92, 226, 719, 654], our findings show that emotion AI is perceived by those who are or may be subjected to it as a technology that reads and manipulates one’s inner thoughts and emotions, and those perceptions pose real and harmful consequences to workers as we show.

Our findings demonstrate the need to study privacy of emotions or *emotional privacy* in more depth — regarding both harms to emotional privacy as well as protections of and rights to emotional privacy. As we show, emotion AI, by definition and design, erodes emotional privacy. To address its invasions of emotional privacy, we must first recognize emotional privacy as part of the human right to privacy—legally and ethically—and acknowledge that people deserve protection against technology-enabled harms from emotional privacy violations. Echoing participants’ sentiments, we argue people ought to have a right to privacy over their emotional information and remain free from emotional manipulation.

Such recognition and protection of emotional privacy could take the form of a civil right and liberty, as argued by legal scholars introducing parallel forms of privacy, notably Citron’s *intimate privacy* [194] and Richards’ *intellectual privacy* [721], which argue that privacy over our intimate and intellectual lives—together encompassing our bodies, health, relationships, thoughts, and beliefs—are fundamental to human flourishing and thus ought to be protected. However, algorithmic inferences thereof have the potential to reveal novel insights due to emotions’ fundamental integration with human behavior and cognition [688]. As such, while emotional privacy may span parallel privacy forms such as intimate and intellectual privacy, the contested and sweeping nature of human emotion raises questions about what it means and what is at stake when *emotions* are inferred using computational means. Whether and how emotional privacy involves concerns of bodily and intellectual integrity, and where it might diverge from established privacy interests, is an area requiring further research and theoretical work to which this discussion serves as a starting point.

5.5.2 Governing Emotional Privacy

Our findings have implications for policy that begins to protect emotional privacy. Law and policy can act as counterweights to limit the otherwise boundless practice of worker surveillance [33, 58]. Yet, US federal law does not currently limit or address the general surveillance of workers [33], barring public employees who enjoy constitutional privacy protection against their government

employers [888]. As such, available legal avenues for workers regarding employer surveillance fall under a patchwork of state legislation and common law privacy torts [888], though both have proven woefully inadequate to protect against and remedy privacy harms workers endure in the workplace [888, 33, 475], and do not cover emotional privacy. Of note, the California Consumer Privacy Act (CCPA) mostly exempted employers from compliance under its “workforce data exemption” [476], though its successor as of 2023—the California Privacy Rights Act (CPRA)—extends protection to all personal information, including employee data [259], which may have implications for workplace surveillance practices.

What’s more, history has shown that new data practices and technologies can enable employers to evade worker privacy protections [33, 227]. In response to surveillance constraints, employers have shifted away from the discreet collection and processing of workers’ personal information and other data practices that are regulated to a participatory approach that engages workers to share their information with employers under the guise of progress and wellbeing [202], in effect normalizing extensive and invasive employee surveillance and silencing its legal objections [202, 33]. Emotion AI-enabled workplace surveillance goes further by no longer requiring workers’ participation to share their thoughts and feelings, instead circumventing worker disclosure of such information with automatic (claimed) inferences of worker emotion and affect. Absent of technological, legal, or normative constraints to restrict its use [33], emotion AI in the workplace stands to collect, process, and share deeply private and sensitive emotional information about workers, leaving them without adequate and explicit protection and vulnerable to the harms we identified in this work.

Of the available employment privacy statutes in the US, most focus on remedying particular harms [736]. Exceptions include a few state statutes that limit the surveillance itself (i.e., video surveillance with audio [324]) and restrict the collection of certain types of employee data (i.e., biometric data [808]). However, because of the breadth of the information emotion AI processes and the uniqueness of the information emotion AI claims to generate (i.e., automatically reading a person’s emotions and affective phenomena more broadly), it is difficult to appropriately classify it under existing regulatory schemes [84]. Open questions remain regarding whether information about human emotion and affect can be protected under existing categories, including thoughts and beliefs, biological and biometric data, sensitive information, and/or identifiable health information [84]; and whether the artificially intelligent nature of the inference’s origin and its ability to “derive the intimate from the available” demands a renegotiation of conventional understandings of individual privacy to capture its potential to enable mechanisms of large-scale, “hyper-targeted control,” [165] particularly at the hands of anthropomorphized, emotionally intelligent AIs [237, 536, 458]. These open questions pose significant barriers to the application of enforceable regulatory frameworks to mitigate, prevent, and remedy potential harms from emotion AI [84, 198], a matter of increasingly pressing public concern [225, 508].

Consequently, legal scholar Bard advocates for the development of a framework to prevent or mitigate emotion AI’s potential harms in particular, rather than AI broadly (i.e., a general AI code of ethics). The development and enforcement of mechanisms to address emotion AI’s harms, as Bard observes, necessarily begin with the task of identifying them [84]. Our identification of emotion AI’s privacy harms in the workplace provides a foundational contribution to this discourse.

At a more fundamental level, regulation and policy could strengthen worker power and expand worker rights. Surely, the lack of available worker protections has enabled the adoption of exploitative and invasive emotion AI-enabled workplace surveillance [925]. Through this work, we have recognized and advocated for a right to emotional privacy in the workplace and identified the potential harms to which workers may be exposed as a result of emotion AI’s erosion of emotional privacy—insights that labor rights advocates could use to take steps in protecting and preserving workers’ emotional privacy.

5.5.3 Designing for Emotional Privacy

There are several opportunities for industry actors to better protect emotional privacy, and mitigate or pre-empt some of emotion AI’s harms within and beyond its application to the workplace.

First, for collective rather than individual monitoring applications, techniques such as differential privacy can protect privacy by introducing noise that offers plausible deniability for any identifiable individuals in emotion AI datasets [466]. For instance, after initial backlash over privacy concerns, the most recent release of Microsoft’s Viva platform, which generates wellbeing-related insights about individual employees and makes that information visible to employees through an individual dashboard [797], uses differential privacy, de-identification, and aggregation [797] to ensure identifiable data is visible only to the employee, while providing “privacy-protected” wellbeing-related insights to management [797, 887]. In addition, decentralized federated learning techniques could prevent the centralized collection of individual, identifiable inferences of emotion, restricting harms from the unregulated and unconstrained flow of emotion data. However, the privacy guarantees of such techniques are limited and should not be regarded as a “silver bullet” to privacy problems [270].

Second, enterprise risk management practices that identify, categorize, assess, and prioritize privacy risks to minimize harm to consumers could recognize the harms of emotion AI and incorporate them into existing and future risk management processes, such as privacy or data protection impact assessments (PIAs/DPIAs) [897] and ethical impact assessments [562]. Given the acceleration of privacy laws and regulation, prudent organizations that handle personal data will adopt data protection and privacy risk minimization standards [361]. To mitigate harm from the collection and processing of emotion data, future work could build on this study to measure the

risk of emotional privacy harm, an important component of several risk mitigation frameworks.

It is important to emphasize that emotional privacy harms may remain even if such policy and privacy interventions to mitigate emotion AI’s harms were implemented. For example, efforts to improve the precision of emotion AI inferences may stem some of emotion AI’s harms (i.e., reputational harms), but the perfect emotional surveillance of a highly accurate emotion AI system may perpetuate or introduce other harms (i.e., psychological and emotional labor harms). While faulty emotion AI can harm people, as we show, machine accuracy improvement is an imperfect solution, as more accurate surveillance systems can indeed exacerbate privacy concerns [376]. Certainly, many of emotion AI-enabled workplace surveillance’s harms (i.e., direct psychological and autonomy harms) cannot be mitigated through either technical solutions or the governance of emotion data, but through the refusal [347] to adopt the emotion AI and prevent its emotional surveillance collecting emotional information in the first place. Surely, non-adoption decisions by organizations would pre-empt the identified emotion AI-enabled workplace surveillance harms all together.

Privacy enhancement, regulation, and risk mitigation all have limits; a failure to consider at a more fundamental level whether it is just to develop, design, and implement systems that implicate the privacy of our inner, emotional lives can expose and exacerbate social injustices for all. These are questions of ethics and justice [120, 254], and to that end we contribute *emotional privacy* to advocate for addressing the many harms posed by technologies that aim to infer emotions and other affective phenomena, and last but not least, an individual right to privacy over one’s emotional information and to remain free from emotional manipulation.

5.6 Conclusion

In examining workers’ experiences and perceptions of emotion AI in the workplace, we find that emotion AI violates workers’ emotional privacy, erodes workers’ ability to manage privacy over their emotional information, and exposes workers to a wide range of privacy harms stemming from its emotional surveillance into workers’ interiority and its enforcement of workers’ compliance with emotional labor expectations. Our results call for the recognition of a human right to *emotional privacy*, which can better guide researchers, policy makers, and industry practitioners to make ethical and responsible decisions regarding emotion AI that protect and preserve peoples’ ability to maintain privacy over their emotional information.

Part III: Measuring Emotional Privacy Across Contexts

Part II established emotional privacy as a foundational privacy interest in emotion AI systems with context-relative stakes. Part III thus turns to the task of measuring emotional privacy by drawing on Helen Nissenbaum’s theory of Contextual Integrity (CI), a widely adopted framework for evaluating the appropriateness of privacy-relevant information flows relative to their social context.

CI evaluates information flows based on five interdependent parameters: subject, sender, recipient, information type, and transmission principle. These parameters define the structure of an informational norm. When flows align with entrenched social expectations across these parameters, they are considered presumptively appropriate. Privacy violations arise when these expectations are breached—when established privacy norms are disrupted. Importantly, CI is not only descriptive but normative: it includes a layered justificatory analysis to determine whether identified privacy violations—or novel flows without established social benchmarks—are normatively justified: (1) identifying the interests at stake, (2) weighing benefits and risks, and (3) evaluating whether the flow advances or undermines the social ends of the context in question [646, 649]. Part III adopts this model to analyze emotional privacy in the context of emotion AI systems in healthcare and employment, applying CI as both a measure framework and a normative diagnostic for emotional privacy.

To operationalize this analysis, I drew upon Kirsten Martin’s work (e.g., [577, 575]) to design a mixed-methods factorial vignette survey. Participants were presented with scenarios in which CI’s five parameters were fixed, while the social context and type of data input were systematically varied. Open-ended prompts followed each vignette to elicit perceived benefits and risks, and a post-survey questionnaire collected demographic data and information about participants’ privacy beliefs. This approach enabled empirical measurement of emotional privacy judgments grounded in CI, while also testing the influence of contextual and individual-level factors on those judgments. I employed a dual-sampling strategy: one nationally representative U.S. sample stratified by race, sex, and age, and one purposive sample of individuals with lived experience of mental illness and/or minoritized racial, ethnic, or gender identities—groups potentially more vulnerable to impacts from the flow of inferred emotional information in these domains.

Participants were presented with scenarios in which an *emotion inference* was held constant across vignettes as the information type parameter. Each scenario clarified that the employer or healthcare provider derived the inference from existing contextual records. Vignettes varied systematically by two source modalities (speech/text patterns vs. facial expressions) and across 14

distinct purposes. This design allowed empirical isolation of participant judgments regarding the privacy of interpreted emotional information, applying CI principles to inference-based systems. By varying both the data source and the purpose of use, the study also illuminated how *meaning* is ascribed to data flows: perceived appropriateness hinges not only on the modality of data collection but on the *intended purpose* of the inference—underscoring the centrality of purpose in privacy judgments involving inferred personal information.

The quantitative data empirically affirm CI's normative heuristic. Judgments by purpose generally aligned with CI's expectation that data flows are more appropriate when they reinforce the contextual ends of a given domain. However, divergences emerged. Regression results and sample comparisons suggest that participants with minoritized identities were more likely to perceive both greater benefits and greater risks than participants in the general population. Effects by purpose were not just stronger, but in some cases, directionally distinct. These results show that normative privacy judgments can diverge across individuals occupying the same roles within a given context—especially when those individuals face differential risks or histories of marginalization.

The qualitative findings reveal that while participants acknowledged potential benefits of emotion AI—such as improved diagnosis, early intervention, or support for workplace mental health—these were overshadowed by concerns that such inferences crossed a moral line. Respondents described risks of excessive surveillance that could erode their ability to meaningfully shape their work or care environments, while simultaneously augmenting the power of already hierarchically advantaged institutions. These concerns reflected a common theme: that machine interpretation of human emotion, particularly when deployed in institutional settings, risks undermining individual dignity and agency.

CHAPTER 6

Emotion Inferences in the Workplace and Healthcare: Workers' and Patients' Emotional Privacy Judgments and the Relative Influence of Contextual, Socio-demographic, and Individual Privacy Belief Factors¹

6.1 Introduction

Emotion AI technologies are now increasingly embedded in workplace and healthcare settings [164, 594, 408, 139, 492, 616]. Deployments promise similar aims across contexts—improved safety, performance, and wellbeing [594, 408, 139, 492, 616]. Alongside the transformative promises emotion AI makes to the workplace and healthcare are substantial privacy and ethical trade-offs. Facilitating unprecedented flows of emotional and affective data, such systems may reinforce power asymmetries, reproduce demographic biases, and enable surveillance and other misuses of sensitive emotional information, with effects that erode institutional trust and harm those subject to the technology [925, 735, 916, 655]. For workers and patients whose jobs or care

¹This chapter is based on three publications: (1) Shanley Corvite*, Kat Roemmich*, Tillie Rosenberg, and Nazanin Andalibi. 2023. Data Subjects' Perspectives on Emotion Artificial Intelligence Use in the Workplace: A Relational Ethics Lens. Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 124 (April 2023), 38 pages. <https://doi.org/10.1145/3579600>. *Co-first authors contributed equally. This material is based upon work supported by the National Science Foundation under Grant No. 2020872.; (2) Kat Roemmich, Shanley Corvite, Cassidy Pyle, Nadia Karizat, and Nazanin Andalibi. 2024. Emotion AI Use in U.S. Mental Healthcare: Potentially Unjust and Techno-Solutionist. Proc. ACM Hum.-Comput. Interact. 8, CSCW1, Article 47 (April 2024), 46 pages. <https://doi.org/10.1145/3637324>. This material is based upon work supported by the National Science Foundation under Grant Nos. 2020872 and 2236674; (3) Kat Roemmich and Nazanin Andalibi. 2024. Emotion Inferences in the Workplace and Healthcare: Workers' and Patients' Emotional Privacy Judgments and the Relative Influence of Contextual, Socio-demographic, and Individual Privacy Belief Factors. (under review at ACM Transactions on Computer-Human Interaction).

may hinge on opaque inferences generated by these technologies, the stakes can be profound—yet existing applications frequently overlook adequate privacy considerations [916, 864, 655].

Understanding when and how privacy is transgressed is essential to grasping the social impacts of emerging technologies and mitigating their harms [194, 793]. Yet empirical knowledge on how workers and patients—the individuals most exposed—judge the appropriateness of emotion AI data flows, and how those judgments vary with context, social position, and privacy beliefs, remains scarce. In its absence, technology research, policy, and practice risk privileging dominant norms while overlooking the heightened vulnerabilities of minoritized groups.

Theoretically and empirically grounded in the understanding that privacy norms are interdependently bound by contextual variables [573, 574], vary by socio-demographic and individual privacy belief factors [478, 117, 520, 112, 559], and may differ between dominant (i.e., nationally representative) and minoritized perspectives [590], this study contributes a deeper understanding of the benefit and risk perceptions of emotion AI held by data subjects, their emotional privacy judgments, and the factors that shape them, by answering the following research questions: *What is the relative influence of contextual, socio-demographic, and individual privacy belief factors on workers' and patients' emotional privacy judgments of emotion AI data flows? What benefits and risks do they anticipate?*

To answer this question, we designed a factorial vignette survey based on Helen Nissenbaum's theory of contextual integrity, which normatively justifies data flows when they uphold the legitimate goals of the context and serve its broader social ends [646]. To conceptualize emotional privacy as the appropriateness of emotional information flows, we structured vignettes by fixing contextual integrity's five canonical parameters: information type, subject, sender, recipient, and transmission principles. Guided by the principle of purpose limitation—which restricts data use to specific, legitimate aims [412, 334]—we systematically varied vignettes by 14 emotion data *purposes* (e.g. safety, diagnostics, performance management) and two *input* modalities (image/video vs. speech/text). Participants rated their comfort across 56 scenarios in total (2 contexts x 2 inputs x 14 purposes). We also collected *socio-demographic* factors and individual *privacy beliefs* (e.g., institutional trust, perceived sensitivity of emotional information) to model their influence on privacy judgments. Recognizing that nationally representative samples may obscure the privacy needs and vulnerabilities of underrepresented groups [590], we conducted the study across two U.S. adult samples: a nationally representative cohort by race, sex, and age ($n=300$) and a targeted oversample of minoritized participants by race/ethnicity, gender, and mental health status ($n=385$). We analyzed cohorts separately to reveal patterns that pooled, weighted analyses might miss.

After considering their comfort levels to each set of context-relative scenarios, participants then answered open-ended questions regarding what benefits, harms, undesired impacts, or concerns, if any, they anticipated from emotion AI use in the domain (employment, healthcare) corresponding

to that of the vignettes to which they had just finished responding. These open-ended questions were intentionally broad to allow participants to conceptualize the impacts most meaningful to them.

Our results yield four key insights for emotional privacy theory, system design, and governance:

1. Purpose is a dominant, context-specific driver of privacy judgments. Holding contextual integrity’s actor, attribute, and transmission principle parameters constant, varying the stated *purpose* of an emotion AI flow shifts mean comfort by -7.9 to $+7.0$ points—the largest swings observed. Purpose shows an interdependent effect: its direction and magnitude vary with the institutional goals of each domain, with some purposes producing the strongest effects across the model. In contrast, *input* modality shows a consistent main effect across contexts and cohorts: replacing image/video with speech/text raises comfort by $+2\text{--}5$ points, reflecting generalized discomfort with facial emotion analytics.

Employment. Flows supporting the workplace’s social mission—keeping workers safe, cared for, and productive—raise comfort: risk-of-harm predictions ($\approx +7$), group mental-health monitoring ($+2.6/+2.2$), and automated acute support ($+1.7/+1.9$). Conversely, evaluative flows importing clinical diagnoses or expanding individual surveillance—early medical diagnoses, individual mental health inferences, and performance scoring—lower comfort (-1.3 to -3.7). While these flows might, in principle, aid productivity and care, participants appear to weigh disclosure risks to employers more heavily, recognizing the power asymmetries such flows may intensify—thus contravening employment’s higher-order social goals (e.g., dignity, fair treatment [57]).

Healthcare. Early neurological screening is the only purpose rated positively in both samples ($+2.2/+3.5$), aligning with healthcare’s aim of improving clinical outcomes. Yet other clinical purposes—early mental health diagnosis, individual mental health inference, and automated interventions—register the strongest negative effects (-3.5 to -7.9), suggesting that granting machines interpretive authority over emotional states heightens vulnerabilities and undermines bodily and decisional autonomy—core to healthcare’s contextual integrity [859].

These purpose-specific variations underscore contextual integrity’s normative claim: a flow is judged appropriate when its purpose furthers the context’s institutional goals and the broader social ends they serve, and inappropriate when it strains or distorts those ends. Our findings therefore support governance efforts to bind the flow of emotion inferences to narrowly articulated, context-serving purposes and apply purpose limitation principles by default.

- 2. Socio-demographic variation shapes emotional privacy judgments in context.** Our dual-sampling approach highlights differing privacy judgments between representative and minoritized cohorts. Across both employment and healthcare settings, participants in the minoritized cohort tended to follow the same directional trends as the nationally representative cohort—but with magnified effect sizes, both positive and negative. These differences suggest heightened perceived susceptibility to emotional inferences and greater judgment intensity, consistent with the idea that position-related *vulnerability* shapes privacy expectations [590]. Importantly, divergences emerged at the purpose level. For example, in employment, early diagnosis for mental illness had a more negative effect in the minoritized sample (-2.5 vs. -1.3), as did neurological disorder screening (-1.7 vs. $+0.5$). These divergences were also context-specific. For instance, in healthcare, early mental health diagnosis was rated less negatively by the minoritized cohort (-0.5 vs. -2), while neurological screening was rated more positively ($+3.5$ vs. $+2.2$). These results suggest that participants more acutely attuned to systemic disparities (e.g., in care access, stigma) may evaluate flows through a different lens of contextual appropriateness—recognizing, for example, how a given use might support or undermine a context’s broader social ends. This position-sensitivity is further evident in the one statistically significant reversal of effect direction: identifying patients in need of support in healthcare, which was rated negatively by the representative cohort (-1.2) but positively in the minoritized cohort ($+1.3$). Such divergences highlight how differing lived experiences inform privacy judgments about whether a data flow upholds or violates contextual integrity. These divergences underscore how lived experience shapes judgments of contextual appropriateness. Concerning socio-demographic factors, while not all effects were statistically significant, patterns by race, gender, mental health status, and education were nonetheless illuminating, wherein we observed patterns consistent with position-related vulnerability. In both contexts, Black participants and those without a Bachelor’s degree tended to report greater comfort, suggesting heightened sensitivity to flows that promote wellbeing and dignity. Meanwhile, gender minorities and participants with mental health histories exhibited sharper negative responses in healthcare, highlighting where emotion AI systems may exacerbate existing vulnerabilities. These findings underscore the need for demographic sensitivity in the design and governance of emotion AI, and caution against assuming nationally representative samples capture the full range of emotional privacy concerns.
- 3. Trust and perceived sensitivity are decisive belief factors.** Institutional trust and perceived sensitivity of emotional data strongly influence comfort judgments. Each one-unit increase in trust, or decrease in perceived sensitivity, shifts comfort by 0.4 to 0.5 points, comparable to many mid-range purpose effects. Specifically, institutional trust increased comfort by $+0.44$ to $+0.54$ per scale unit, while perceived sensitivity decreased it by -0.25 to -0.3 . Notably,

perceived sensitivity of emotional information varied by context and, in some cases, was even rated higher than traditionally recognized sensitive categories of data such as biometric, genetic, or union membership information. This underscores the need to treat emotional information as a first-order privacy concern. Because these are continuous variables, their cumulative effect may exceed that of any single contextual or demographic variable we tested. As trust varies widely across workplaces and healthcare settings, meaningful protections should therefore be built into systems by design—not deferred to assumed institutional goodwill—and governed according to the heightened sensitivity of emotional information.

4. **Workers and patients fear erosion of dignity and contextual values.** While participants acknowledged that emotion AI use by employers or healthcare providers could, in principle, support positive outcomes (as framed in the survey vignettes), these potential benefits were consistently overshadowed by a wide range of perceived risks and harms. In both contexts, participants feared that the introduction of emotion AI would exacerbate the very challenges they already face—challenges that are often interconnected in the privatized landscape of the U.S., where access to care frequently depends on employer-provided subsidies and health plans. These concerns included the difficulty of maintaining work-life boundaries, the stigmatization of mental health issues, the quality and timeliness of care, and the lack of meaningful voice in shaping these institutional environments as both employees and patients.

Participants also emphasized contextual and relational concerns. Even when emotion AI was introduced for ostensibly pro-social aims—such as promoting safety, wellbeing, or mental health—it was often seen as degrading the relationships that gave these settings their meaning. In healthcare, this meant the loss of the “human” in human medicine, as the values of the patient-provider relationship was perceived to be displaced by those of the commercial market. In the workplace, it meant further erosion of worker agency and dignity in environments already characterized by intrusive surveillance, leading to strained peer relationships and deepened hierarchical power asymmetries.

These findings underscore that the core affiliations that make these social contexts meaningful are themselves at stake. As Contextual Integrity reminds us, privacy norms are grounded in how information flows shape the roles, responsibilities, and values embedded in specific social contexts [649]. When emotion AI purpose imply support to those values—by enhancing agency or preserving dignity—participants responded more favorably. But when it more obviously threatens them, concerns were more prominent. The moral evaluation of emotion AI, then, hinges not only on contextual purpose, but also on whether it sustains or erodes the normative structures that make these contexts worth protecting in the first place.

By grounding emotional privacy as a normative judgment shaped by contextual goals, individual

beliefs, and position-related vulnerabilities, this study extends contextual integrity by incorporating diverse participant perspectives and empirically testing the influence of purpose—alongside fixed contextual integrity parameters—on privacy judgments. In doing so, we offer both a theoretical refinement and an actionable model for evaluating the acceptability of emotion AI systems within the contextual integrity framework. These insights lay the foundation for designing and governing emotion AI technologies that respect autonomy, support dignity, and advance the broader social ends these systems claim to serve. At the same time, they surface a critical challenge: ensuring that the full socio-technical pipeline—from purpose specification to transmission constraints and system safeguards—consistently upholds these normative commitments across specific deployments and contexts. Our findings also offer empirical support for recent regulatory developments, such as the EU AI Act, and provide a model for anticipating future governance needs. In the sections that follow, we elaborate this framework, present our empirical findings, and discuss implications for the responsible development, deployment, and regulation of emotion AI grounded in heightened safeguards for emotional information, purpose-aware design and governance, and sensitivity to contextual and positional vulnerabilities in privacy research and practice.

6.2 Background and Related Work

Despite growing deployments of emotion AI in the workplace and healthcare, the privacy implications of these technologies remain poorly understood—particularly from the perspective of those most affected. While calls for empirical attention to privacy in technologies handling emotional data are growing, particularly in applications of AI to the workplace [735, 563, 922] and healthcare [125, 778, 655, 916], two persistent gaps limit progress. First, there is a structural gap: the actors designing, deploying, and evaluating these technologies often lack insight into the situated norms and vulnerabilities of the individuals over whom they hold power [576, 265]. Second, there is a conceptual gap: emotional privacy remains empirically under-theorized and thus difficult to operationalize [735], with limited empirical attention to how emotional information flows are judged across contexts, identity characteristics, and belief systems.

The present study addresses both gaps by investigating how three interdependent dimensions—(1) contextual factors, (2) socio-demographic characteristics, and (3) individual privacy beliefs— influence emotional privacy judgments. Our analytic framework models the relative influence of these factors on workers’ and patients’ judgments of emotion AI data flows *alongside* the structural parameters of contextual integrity. In doing so, we build on contextual integrity theory to clarify what emotional privacy means in practice, identify the factors that shape its protection or violation, and inform governance efforts to align emotion AI with both human values and contextual demands. This section reviews literature motivating our inclusion of these factors as explanatory variables

alongside contextual integrity parameters, with attention to how identity-based vulnerabilities can influence emotional privacy judgments in work and healthcare domains.

6.2.1 Contextual Factors

According to the theory of contextual integrity, privacy norms are shaped by the interdependent parameters of a given social context, including actors, attributes, and transmission principles. An information flow is judged appropriate when it aligns with these norms and supports the context's institutional goals [646]. What is acceptable in healthcare may not generalize to the workplace; these domains differ not only in place but in politics, conventions, and cultural expectations [646, 871]. Attending to the specific contextual configurations of employment and healthcare is thus essential for evaluating whether, and to what extent, emotional privacy is preserved or violated. Such analysis can reveal gaps between normative ideals and lived experience, enabling more socially responsive design and policy aligned with the values of those most affected by emotion inference technologies. In addition to measuring emotional privacy through contextual integrity's core parameters, this study examines the relative influence of two further contextual factors: the modality of data *input*, and the stated *purpose* of its use. These are particularly salient in emotion AI, where inferences are drawn from multimodal signals and often used for opaque or poorly justified ends [139, 521, 734].

6.2.1.1 Data Input

In both workplace and healthcare contexts, emotion recognition frequently relies on text, speech, and facial data, often in combination with additional contextual or biometric information [139, 837, 261, 506, 72]. The specific input modality may influence emotional privacy judgments, as privacy perceptions are known to vary across data types. Facial and bodily data captured through cameras and facial recognition systems raise concerns about visibility, identification, and biometric surveillance [922, 804, 80, 915]. Speech data collected via continuous microphones, such as those in smart speakers, elicit fears of ambient surveillance and constant monitoring [514]. Text-based inputs, including monitored emails and messages, prompt concerns about intrusions into private communication and intent inference [880]. In emotion AI, such data are not shared directly but processed to infer emotional states and then automatically disseminated. Under the theory of contextual integrity, these input modalities form part of the broader "sender" of emotional information [648].

Empirical studies support the idea that input type shapes privacy perceptions. Lee et al.'s qualitative study on mobile affective computing found sensor-specific concerns, with users worried that certain data types could expose personal traits and lead to profiling and surveillance [521]. Similarly, Zhang et al. showed that inferences about mental health from mobile data triggered

privacy concerns that varied by data source and contextual framing [916]. These findings suggest that the input source used to generate emotion inferences may directly shape how workers and patients evaluate emotional privacy.

6.2.1.2 Purpose

The purpose for which information is collected and used is known to shape privacy perceptions—particularly when the stated purpose offers a personal or collective benefit [295, 258, 464, 592, 731, 750, 896, 629]. Individuals are often more willing to share sensitive information, including emotional or health-related data, when they perceive it as contributing to their own wellbeing [117] or advancing a broader social good [430, 846, 521]. However, purpose-driven framing can be leveraged by powerful actors to normalize surveillance and downplay privacy risks. In workplace contexts, for instance, employers increasingly frame monitoring tools as enhancing productivity or wellbeing, thereby encouraging participation while limiting dissent [34]. Similarly, in healthcare, optimistic narratives about digital tools can obscure underlying privacy threats [113]. While positive framing may mask certain risks, people remain concerned about their emotional privacy even when purported benefits are emphasized. Zhang et al. found that privacy concerns persisted in mobile mental health apps, despite framing these tools as beneficial [916].

These findings highlight the contextual salience of *purpose* in shaping emotional privacy perceptions. Indeed, U.S. privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) embed purpose specificity as a normative constraint on sensitive data use [348, 22]. While purpose is not explicitly included among contextual integrity's five canonical parameters, applications of the theory often treat purpose as an optional transmission principle constraining data use [646]. Given that emotion AI systems infer emotions from diverse data inputs and for a range of ends, we expect emotional privacy judgments to be shaped not only by the structural features defined by contextual integrity, but also by the *purpose* for which emotion inferences are used. By empirically assessing the effect of purpose across fixed contextual integrity parameters, our study extends contextual integrity by measuring how purpose meaningfully contributes to emotional privacy judgments in context-specific ways.

6.2.2 Socio-Demographic Variations

While contextual factors define the descriptive and normative boundaries of privacy within a given setting, individuals' socio-demographic characteristics shape how those boundaries are perceived, enforced, and contested.

Empirical work shows that privacy perceptions vary across socio-demographic dimensions such as education [117], race/ethnicity [117], and gender [520, 112]. Although the relationship between

privacy perceptions and socio-demographic status remains understudied [478], research in privacy and HCI suggests that identity-based characteristics influence both one's exposure to surveillance and sensitivity to its harms [112, 382, 390].

6.2.2.1 Education

Recent Pew findings indicate that public concern over AI applications varies by educational attainment. Individuals with postgraduate education expressed greater concern about facial recognition by police, while those with a high school diploma or less were more concerned about AI-enabled misinformation detection and autonomous vehicles [709]. Similarly, Bhatia and Breaux found that individuals with doctorate degrees reported lower concern about sharing personal information than those with lower educational attainment [117].

6.2.2.2 Race/Ethnicity

Research shows that Black and Latine populations are afforded less privacy in U.S. society, in part due to the long-standing normalization of racialized surveillance practices [151, 204]. This structural disparity may contribute to privacy resignation and a diminished perception of risk, despite disproportionately high levels of vulnerability to privacy intrusions [382]. People of color may face heightened risks from emotion AI technologies, as studies have found that emotion recognition algorithms using speech, facial analysis, and natural language processing are often less accurate for people of color—increasing the likelihood of misclassification and harm [719, 905, 416].

6.2.2.3 Gender

Gendered surveillance also shapes privacy experiences and concerns. Women, who face disproportionate exposure to gender-based harassment [468, 788] and workplace monitoring [804], consistently report heightened privacy concerns in these contexts [330, 95, 59, 398, 858]. While research on the privacy perceptions of transgender and non-binary individuals remains limited, existing scholarship suggests that gender minorities have distinct privacy needs—shaped by elevated vulnerability to technological harms from exclusion and exposure [760, 865], as well as a heightened reliance on safe, supportive, and affirming technology-mediated interactions [393, 525, 388].

6.2.2.4 Mental Health Status

The privacy perceptions of individuals with mental illness warrant special consideration, as this population may be more vulnerable to the impacts of automated emotion inference. While emotion

AI applications may offer mental health benefits in some cases [492], individuals with mental illness also face heightened risks from stigmatization, disability discrimination, and inaccurate inferences [616]. Research centered on the privacy perspectives of individuals with mental illness highlights a constrained choice architecture: despite recognizing personal privacy risks, individuals report feeling compelled to trade privacy for access to potentially beneficial mental health technologies, including those that rely on emotion inferences [221, 125].

These risks are magnified in employment and healthcare contexts. Workers with mental illness frequently hesitate to disclose their conditions due to fears of discrimination, damaged professional reputations, and lack of confidentiality. Disclosure decisions often involve weighing the potential benefits of accommodations against risks of stigma and exclusion [840, 149, 293]. Similarly, patients with mental illness report distinct privacy concerns about health information sharing—even in clinical settings—stemming from prior experiences of mental health-related mistreatment [778]. These concerns extend to mobile mental health applications, where participants express particular discomfort with the sharing of sensitive data types such as social interaction information—potentially due to lived experiences with isolation and vulnerability [916]. The emergence of emotion AI in digitized workplace and healthcare settings may compound these challenges. Emotion AI systems have roots in problematic efforts to pathologize affective difference, such as the stigmatization of emotional expression in autistic individuals [631]. These technologies often prioritize the extraction of emotionally legible signals, despite weak epistemological foundations, and risk medicalizing affective variance [459, 802]. Such concerns are especially acute for those with conditions marked by atypical affective expression and regulation, who may be subject to mental illness predictions generated by models that circumvent personal disclosure decisions [258, 366].

The literature reviewed here suggests that emotional privacy judgments may vary by socio-demographic characteristics—including education, race/ethnicity, gender, and mental health status—each of which can shape individuals’ exposure to risk and capacity for control in digital environments. Our study builds on this body of work by empirically examining the relationship between socio-demographics and emotional privacy judgments concerning emotion inferences in workplace and healthcare contexts. We do so using two U.S. samples: a nationally representative cohort stratified by sex, age, and race, and a minoritized cohort composed of people of color, gender minorities, and individuals with lived experience of mental illness. Further details on recruitment and sampling are provided in Section 6.3.0.4.

6.2.3 Privacy Belief Influences

Beliefs about privacy can impact how individuals perceive and evaluate technologies, including general privacy concerns, perceived sensitivity of information sensitivity, risk perceptions, and

levels of institutional trust [559, 777, 610, 867].

6.2.3.1 General Privacy Concerns

Many studies measuring privacy rely on the concept of privacy concern, in part due to enduring disparities in how privacy is defined and conceptualized [652, 148]. While early scales captured privacy concern as a generalized construct [559, 854, 884, 791], a growing body of work shows these concerns are highly sensitive to context [20, 24, 622]. Consequently, general privacy concerns have limited utility in explaining context-specific privacy preferences and outcomes [569]. Still, general concern measures like the Internet User's Information Privacy Concerns (IUIPC) scale remain widely used—both as controls in privacy perception studies [559] and as predictors of related constructs such as privacy decision-making [469] and expectations [575].

6.2.3.2 Perceived Risk

Closely related to privacy concern is perceived privacy risk [502, 117]. While sometimes conceptualized as a global construct [350, 790], perceived risk can also be framed in terms of specific potential harms [117]. It can affect privacy judgments on an affective level: when a technology or data practice is viewed positively, it tends to be perceived as less risky and more beneficial, reducing privacy concerns overall [677, 323]. Risk perception is also a known mediator in privacy judgments. For example, in healthcare settings, Alraja et al. found that attitudes toward emerging technologies were shaped by perceptions of privacy, security, and trust, mediated through individual risk assessments [43]. This underscores the value of accounting for perceived risk in models of privacy perception.

6.2.3.3 Trust

Privacy and institutional trust are mutually reinforcing constructs [867, 571]. Trust in a specific institution can influence both general and context-specific privacy judgments [571], often by reducing the perceived risk of information misuse [737, 904]. At the same time, individuals' baseline privacy dispositions may precede and shape their levels of institutional trust [469, 571]. In both workplace and healthcare settings, trust plays a key role. Tolsdorf et al. found that workers' privacy perceptions in digitized workplaces were shaped by trust in their employers' handling of personal data [838], while Shen et al. showed that patients' willingness to share health information was similarly influenced by trust in healthcare organizations [778]. These findings highlight the importance of modeling institutional trust when evaluating privacy perceptions in contextually sensitive domains.

6.2.3.4 Data Sensitivity

Data sensitivity is best understood not as a static property, but as a belief that varies by individual traits and situational context [610, 565]. Sensitivity is closely linked to privacy risk: more sensitive data is seen as riskier and requiring stronger protections [559, 99]. Prior work suggests that people perceive emotional information [731, 53] and related data such as mental health status [856] as highly sensitive, particularly in commercial or surveillance contexts [221]. Importantly, people often underestimate the sensitivity of data *inputs*—like sensor data—while expressing strong concern about the *inferences* drawn from them. In Lee et al.’s study on mobile affective computing, participants viewed raw sensor data as relatively non-sensitive and often failed to recognize how it could be processed to reveal emotional or psychological traits. When made aware that such data could be used to reveal personal traits, however, participants expressed greater concern [521]. These findings suggest that emotional privacy judgments may be more accurately captured when people are explicitly informed about how inferences are generated—i.e., from which inputs, and for what purposes. We expand on these implications for vignette design in Section 6.3.

Our study examines how emotional privacy judgments are shaped by individual privacy beliefs. By analyzing workers’ and patients’ evaluations of emotional information flows in workplace and healthcare contexts, we offer insight into how individual beliefs interact with contextual features and socio-demographic characteristics. These findings inform the design of policies and systems that more closely align with the privacy judgments and needs of diverse people and groups.

6.3 Methods

A useful method to uncover individuals’ privacy perceptions about a technology which may otherwise be difficult to examine [646, 447, 653, 898], we designed a factorial vignette survey to elicit workers’ and patients’ emotional privacy judgments concerning automatic emotion inferences, and allow us to investigate how their emotional privacy judgments vary by individual and situational factors. From their perspectives as workers and patients, participants rated their level of comfort to a series of vignettes in which their employers and healthcare providers processed data already collected about them to automatically infer their emotions. We varied the vignettes by contextual factors, and issued a post-test for participants to report their socio-demographic information and privacy beliefs. Our analysis contributes an understanding of whether and to what extent workers’ and patients’ emotional privacy judgments concerning automatic emotion inferences vary by contextual, socio-demographic, and individual privacy belief factors. In this section, we describe our survey’s theoretical underpinnings, design, recruitment and data collection efforts, and data analysis procedure, followed by a reflection on our research’s limitations and opportunities for

future work.

Contextual Integrity and Privacy Vulnerability as Theoretical Frameworks to Measure Emotional Privacy Judgments

Two theoretical frameworks for privacy underlie our study design: (1) Nissenbaum's *contextual integrity* [646], which defines privacy "as respecting the appropriate norms of information flow for a given context" [573]; and (2) McDonald and Forte's *privacy vulnerability*, a theoretical perspective to surface the privacy risks vulnerable people face in the operation of privacy norms [590].

Contextual Integrity

Under contextual integrity, privacy violations occur when information flows transgress contextually specific privacy norms. To establish a privacy norm, five specifications are necessary: (1) information type (about what); (2) subject (about whom); (3) sender (by whom); 4) recipient (to whom); and (5) transmission principle (flow under what conditions) [573]. Together, these parameters "predict a complex dependency between privacy judgments on the one hand, and the values for all five parameters on the other" [574]. As such, it was important that our study recognized the combined interdependency of these contextual parameters (in addition to individual differences) when investigating workers' and patients' emotional privacy judgments by using this framework to establish emotional privacy norms in the workplace in healthcare.

Methodologically, factorial vignette surveys are well-suited to account for a set of interdependent contextual parameters to surface privacy perceptions, enabling researchers to study the effect of factors *in combination* on privacy perceptions by asking participants to report their perceptions to various scenarios that are bound within contextual specifications and vary by a researcher's factors of interest. Informed by prior work specifying contextual, socio-demographic, and individual privacy belief parameters in factorial vignettes to study privacy perceptions [646, 117, 574], our vignette design uses contextual integrity principles to measure emotional privacy judgments by defining contextual specifications that govern norms surrounding emotional information sharing in the workplace and healthcare as follows in Table 6.1.

Privacy Vulnerabilities

Though contextual integrity is a leading theoretical framework for privacy scholarship in Human-Computer Interaction (HCI), McDonald and Forte argue that it often overlooks how privacy norms can function unevenly—benefiting privileged groups while disadvantaging vulnerable or minoritized groups. Drawing upon critical theories that expose how norms themselves can perpetuate

Contextual Parameter	Emotional Privacy Norms
Information Type	emotional: including but not limited to mood, stress, anxiety, depression, boredom, calmness, fear, fatigue, attentiveness, happiness, sadness, disgust, surprise, anger
Subject*	employees/patients
Sender	emotion inference system
Recipient*	employer/healthcare provider(s)
Transmission Principles	<ul style="list-style-type: none"> – recipient retains subject's emotional information indefinitely, as allowed by law – recipient will not share subjects' emotional information, unless otherwise noted – subject consented to monitoring by recipient

Table 6.1 Privacy Norms – Fixed Parameters in Vignettes, Adapted from Martin and Nissenbaum, 2016 [573] **Factorial Vignette Condition*

exclusion and oppression [526, 335, 579, 229, 526, 205], they propose that HCI research move “beyond norms” to center *privacy vulnerability* as both an analytic and normative lens. This perspective recognizes how individuals’ identities and social positions shape their privacy risks, which may not be reflected in dominant norms, and seeks to advance a socially just understanding of privacy that accounts for all [590].

We aligned our study with this perspective in two ways. First, our study design accounted for socio-demographic differences by quantifying the relative influence of education, race/ethnicity, gender, and mental health status on emotional privacy judgments measured using contextual integrity theory. Second, we conducted the study across two samples: a U.S. nationally representative cohort by race, sex, and age ($n=300$), and a cohort oversampling participants by minoritized identity statuses—race/ethnicity, gender, and mental health status ($n=385$).

Analyzing these groups separately allowed us to identify comparative patterns that could otherwise be obscured in pooled analyses. This methodological choice was not intended to monolithize minoritized perspectives, but rather to empirically examine McDonald and Forte’s theoretical proposition: that privacy norms may obscure, reinforce, and therefore systematically disadvantage certain groups based on intersecting vulnerabilities. By disaggregating socio-demographic factors and comparing privacy judgments across dominant and minoritized cohorts, we use contextual integrity not as a static normative framework but as an empirical tool to reveal how privacy expectations may differ across social identities. This approach supports, rather than replaces, contextual integrity’s foundational principles while advancing McDonald and Forte’s call to empirically surface privacy vulnerabilities as both analytic and normative concerns. While our study focused on education, race/ethnicity, gender, and mental health status, we recognize that other minoritized statuses (e.g., disability, assistive technology use) are also relevant and warrant future empirical

attention.

Normatively, we conceptualize vulnerability as referring to groups requiring additional protections or safeguards beyond those conventionally provided, consistent with medical and research ethics standards [67, 332]. Empirically, we define vulnerability as encompassing groups known to face significant disparities and unmet needs (e.g., risk factors, access, outcomes) in labor and health domains—including the economically disadvantaged by education; racial, gender, and ethnic minorities; and individuals with chronic health conditions including mental illness [829, 262]. Through this lens, our study centers *privacy vulnerabilities* by: (1) incorporating socio-demographic factors known to shape privacy judgments (education, race/ethnicity, gender, mental health status), and (2) adopting a dual-sampling strategy to comparatively assess emotional privacy judgments between socially dominant perspectives (i.e., U.S. representative sample) and minoritized groups known to be disproportionately surveilled, more vulnerable to privacy harms, or otherwise possessing distinct privacy needs. The literature supporting this approach is reviewed in Section 6.2.2.

By integrating *contextual integrity*'s normative parameters with a theoretically grounded and empirically informed *privacy vulnerabilities* lens, our study investigates both how emotional information flows are evaluated and how privacy judgments vary across social positions with privacy experiences, expectations, and needs.

Survey Design

Privacy skeptics often point to what is commonly referred to as the *privacy paradox*: though people say they have privacy concerns, behaviors implicating their privacy suggest otherwise [795]. One way to explain the privacy paradox relates to how we measure privacy in the first place, with privacy research often failing to specify and account for the variables upon which privacy judgments so crucially depend [573]. Other explanations include an individual's lack of awareness regarding the extent to which data is collected and repurposed, and how said collection and use may impact them [21, 570, 524]. Certainly, how we measure privacy also has important societal implications, as public policy often relies upon conceptualizations of privacy as employed in research [573, 502] to inform privacy regulation, and it is therefore important to attend to factors that can influence privacy perceptions and norms when conceptualizing, operationalizing, and measuring privacy [573].

6.3.0.1 Factorial Vignettes

Conventional privacy research often overlooks the contextual and individual variables that shape privacy expectations, the perception of privacy violations, and for whom those violations are most

salient [573, 590]. Grounded in contextual integrity theory (Section 6.3), our factorial vignette design systematically incorporated these variables to investigate workers' and patients' emotional privacy judgments.

Vignette Structure. Each vignette described a scenario in which an employer or healthcare provider used data already collected about the participant to automatically infer emotions. To ensure clarity and standardization, we fixed the contextual parameters concerning consent, data retention, and sharing practices (Table 6.1) and provided participants with the following reference statement at the start of each vignette set:

“Emotional state” refers to your emotions and moods, including but not limited to stress, anxiety, depression, boredom, calmness, fear, fatigue, attentiveness, happiness, sadness, disgust, surprise, and/or anger. Unless otherwise noted, assume that:

1. your employer/healthcare provider will not share your information;
2. your information is retained indefinitely, as allowed by law;
3. you have consented to this monitoring through a consent form.

Participants were instructed to consider their willingness to be the subject of the described technology, taking into account the type of data, its intended use, and the social context.

Within-subjects Experimental Design. The vignette design followed a 2 (Context: workplace, healthcare) x 2 (Data Input: speech/text, image/video) x 14 (Purpose) within-subjects design. All participants responded to all 56 scenarios. Vignettes were split into two sets by context. To avoid ordering effects, we randomized vignette presentation order across the three nested dimensions: (1) context, (2) data input, and (3) purpose.

Dependent Variable: Comfort. For each vignette, participants rated their comfort using a Visual Analog Scale (VAS) ranging from 0 (“very uncomfortable”) to 100 = (“very comfortable”). The VAS permitted 1-unit increments, treating comfort as a continuous measure and allowing participants to respond in line with mental models of subjective experience rather than ordinal categories. VAS is widely recommended for measuring subjective phenomena due to its metacognitive sensitivity and ability to capture fine-grained judgments [713, 338, 404], while avoiding common limitations of ordinal (e.g., Likert-type) scales such as clustering and data loss [841, 187, 191, 446, 42, 124].

Although no consensus exists on the optimal dependent variable for privacy perceptions research [646, 447, 653, 898], the appropriate measure depends on the construct of interest. Studies

focused on *behavioral privacy* often use willingness-to-use (e.g., [449, 117]). However, such constructs are less suitable for *normative privacy judgments*, especially in power-imbalanced settings like employment and healthcare, where choice constraints and institutional pressures may shape expressed willingness and risk obscuring underlying privacy concerns—a dynamic consistent with the bounded rationality and malleability of privacy perceptions identified by Acquisti et al. [21].

Studies eliciting normative privacy judgments commonly use either participants' comfort levels (e.g., [535, 667]) or judgments of acceptability (e.g., [574]). We selected comfort because acceptability can be shaped by adaptive preferences or resignation to constrained choices, particularly among workers and patients with limited agency over emotion inference technologies. Comfort also correlates strongly with perceived privacy risk [117]. Although Bhatia and Breaux operationalized perceived privacy risk through willingness-to-share measures, their factorial vignette studies treated these ratings as reflecting both behavioral intent and normative judgments of risk acceptability. Their finding that discomfort ratings explained up to 79% of the variance in perceived privacy risk supports the use of comfort as a valid single-item measure of privacy judgments in scenario-based designs, offering a practical and efficient proxy where more complex outcome measures may be prohibitive. Finally, selecting comfort aligns with contextual integrity's emphasis on individuals' intuitive judgments of normative appropriateness relative to contextual norms [648]. While we regard comfort as the most suitable dependent variable for this study's focus and design, future work could explore alternative constructs or multi-item measures to assess variations in emotional privacy judgments across contexts, identity characteristics, and measurement approaches.

Vignette Prompt. We asked participants to rate their level of comfort with each scenario using the following text, adapted to context. Variable levels are summarized in Table 6.2.

As a \$C1, rate your comfort (0 = very uncomfortable to 100 = very comfortable) with your \$C2 using a computer program to automatically detect your emotional states using records of \$I collected from your daily activities and device use, for the purpose of \$P.

We used the phrase “computer program to automatically detect emotional state” to promote neutrality and comprehension. This phrasing avoided technical jargon (e.g., “emotion AI”), stigmatizing proxies (e.g., “mental health state”), or unfamiliar terms (e.g., “affective state”) that could bias or confuse participants.

Participants saw only the text relevant to the given context (“*as an employee...*” or “*as a patient...*”) crossed with the assigned data input, followed by a separate VAS slider for each of the 14 purposes. This design minimized cognitive load and improved response efficiency. Although

Vignette Variable Levels	
Context (\$C)	(1) employee* (\$C1), employer** (\$C2), work performance (\$C3) (2) patient* (\$C1), healthcare provider** (\$C2), overall health (\$C3)
Data Input (\$I)	(1) what you say (either verbally or written/typed) and how you say it (e.g., speed, tone) (2) images or video of what you look like, based on your facial expressions
Purpose (\$P)	(1) giving (\$C2) data-driven insights into (\$C1)'s wellbeing (2) sharing that information with academic researchers (3) diagnosing mental illness in (\$C1) earlier than otherwise possible (4) diagnosing neurological disorders (e.g., dementia, ADHD) in (\$C1) earlier than otherwise possible (5) avoiding subjectivity in other methods (\$C2) may use to learn about your emotional state (e.g., surveys, observations) (6) inferring mental health state of (\$C1) individually (7) inferring mental health at the group level only (8) identifying (\$C1) needing mental health support to better plan organizational mental health resources (9) inferring (\$C1) at risk of harming others (10) inferring (\$C1) at risk of self-harm (11) developing intelligent computer therapy programs for (\$C1) (12) detecting moments (\$C1) may need emotional support and responding to help (13) alerting (\$C2) when (\$C1) may need support (14) assessing (\$C3) of individual (\$C1)

Table 6.2 Variables and Levels by Contextual Factor.

**Contextual Integrity Parameter: Data Subject;*

***Contextual Integrity Parameter: Data Recipient*

participants rated 56 vignettes (28 per context), the consistency of the purposes across data input conditions allowed participants to develop familiarity with the response format and proceed quickly. Figure 6.1 shows an example vignette from the employment context.

Purpose Selection. The 14 purposes included in this study were selected through a two-stage process to ensure relevance, validity, and comparability across contexts. First, we identified common and emerging uses of emotion AI documented in industry practice through patent analyses of workplace [139] and healthcare [462] applications. Second, we cross-referenced these industry uses with scholarly literature describing potentially beneficial applications of emotion AI across both settings. This literature emphasizes purposes such as providing general wellbeing insights and

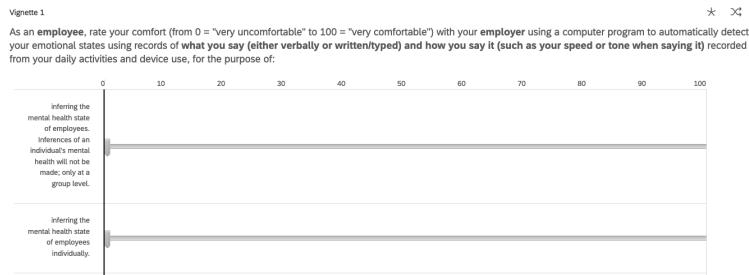


Figure 6.1: Presentation of Vignettes for the Employment Context, Partial Example

more specific mental health inferences at both individual and group levels [184, 492]; detecting or preventing self-harm or harm to others [731, 258, 750, 833]; and enabling early detection of mental and neurological illness with the goal of improving mental health support, safety monitoring, and research [258, 366, 592, 731, 833, 492].

While comprehensive, this set is not intended to be exhaustive. Rather, the purposes represent a theoretically and empirically grounded sample of common and proposed emotion AI uses relevant to workplace and healthcare contexts. In our analysis, purpose is modeled as a fixed effect, reflecting these specific uses, rather than as a random sample intended to generalize to all conceivable purposes. This design choice supports the validity of our findings while acknowledging that other uses warrant future empirical attention.

6.3.0.2 Open-ended questions

After completing each of the two vignette sets (employment and healthcare contexts), participants answered four open-ended questions:

1. In what ways, if any, do you think these systems could benefit you? Please describe and provide examples and as much detail as you are comfortable with.
2. In what ways, if any, do you think these systems could harm you or have other undesired impacts on you? Please describe and provide examples and as much detail as you are comfortable with.
3. What other concerns, if any, do you have about these systems? Please describe and provide examples and as much detail as you are comfortable with.
4. In what ways, if at all, do aspects of who you are (for example, your race/ethnicity, gender, sexuality, employment status, class, education, mental health conditions, physical health conditions, or any other features of your identity) shape your responses to the use of computer programs to infer your emotional states?

The qualitative data gathered through these questions provided rich insights into participants' perceived benefits, risks, and personal contexts influencing their privacy judgments.

6.3.0.3 Post-test

After completing the vignette ratings and open-ended questions, participants responded to a post-test that gathered additional information about individual characteristics. Following best practices for inclusive survey data collection [799, 159, 328, 356], the post-test collected socio-demographic information, including race/ethnicity, gender, age, subjective socio-economic status, mental health status, employment status, and educational attainment.

The post-test also assessed individual privacy beliefs. Participants responded to items measuring general information privacy concerns, perceived risks of employer and healthcare provider access to sensitive personal information, institutional trust in those entities, and the perceived sensitivity of emotional information relative to other commonly recognized sensitive data types [1, 198]. These items adapted the Internet Users' Information Privacy Concerns (IUIPC) scale [559] to our specific contexts of employment and healthcare. Full item wording for the socio-demographic and privacy belief measures appears in Appendices A.4 and A.5. Participants used the same Visual Analog Scale (VAS) ranging from 1 to 100 to report privacy beliefs, maintaining consistency with the vignette ratings. To avoid potential priming effects, we administered the post-test only after participants completed all vignette responses.

These measures allowed us to analyze whether, and how, socio-demographic and privacy belief factors shaped emotional privacy judgments alongside the contextual integrity parameters and additional contextual factors varied in the vignettes.

6.3.0.4 Pilot Study

To ensure the survey consistently measured the intended constructs, we conducted a pilot study ($n=25$) in which participants completed the survey vignettes and provided feedback on any confusing elements. Analysis of the pilot data indicated that no substantive design changes were necessary. Participants' responses confirmed that their comfort ratings reflected perceptions of employer or healthcare provider use of computational emotion inferences specifically, rather than general monitoring—supporting the survey's construct validity.

The pilot also assessed potential participant fatigue. We included attention check questions and monitored completion time. While factorial vignette designs often entail a learning curve due to their novelty rather than respondent fatigue [576]; participants became familiar with the vignette format quickly. Despite evaluating 56 vignettes, the average completion time was 24 minutes, and only two participants failed the attention check. These results indicated the survey length was

appropriate for the study's objectives.

Recruitment and Data Collection

6.3.0.5 Sampling

We collected two samples to assess emotional privacy judgments: (1) a U.S. nationally representative sample by age, sex, and race ($n=300$), and (2) a sample oversampling individuals with one or more minoritized identities (person of color, minority gender, and/or mental illness status; $n=385$). As described in Section 6.3, this sampling strategy allowed us to investigate how privacy judgments vary both within and between socially dominant and minoritized perspectives.

6.3.0.6 Recruitment

Participants were recruited via Prolific, using pre-screening criteria for age, sex, race, minoritized identities, and other relevant characteristics. The nationally representative participant group was recruited in October 2021 using Prolific's automatic balancing feature. The minoritized participant group was recruited between December 2021 and February 2022 using targeted pre-screening. Participants completed the survey through Qualtrics and were compensated \$3.80, following Prolific's recommended rate. We note that some under-represented gender and ethnic minority groups could not be analyzed separately due to small sample sizes. Summary statistics are reported in Table 6.3.

6.3.0.7 Ethical Oversight

Our institution's IRB determined that this study qualified for exemption from oversight under 45 CFR 46.104(d)(2)(i), which applies to survey procedures where information is recorded such that subjects cannot readily be identified, directly or indirectly [853]. Data were collected anonymously via the Prolific platform, which compensated participants directly, eliminating the need for researchers to collect linkable personal information. The study was determined to involve no more than minimal risk to participants, consistent with federal research ethics standards.

Exemption from oversight does not preclude ethical responsibility. The research team followed best practices to protect participant privacy, data security, and dignity, including obtaining informed consent, ensuring anonymity in survey responses, minimizing participant burden, and reviewing pilot study results for potential ethical concerns. The pilot study identified no design or content issues, and participants' open-ended responses indicated high engagement and willingness to reflect on the study topics. Although Prolific assigns participant IDs, these were not published or linked to study results, and data access was restricted to the research team.

Factor	Level	Rep. Sample	Min. Sample
Race / Ethnicity	Additional ethnicities	11	26
	Asian	26	47
	Black	51	104
	Latine	15	42
	White	197	194
Gender	Trans and/or non-binary	6	44
	Woman	148	232
	Man	146	139
Mental Health Status	Under treatment for 1+ mental illness	67	115
	Untreated / resolved mental illness	50	101
	No mental illness	183	140
	Did not report	0	57
Age Group	18–27	55	170
	28–37	55	120
	38–47	49	42
	48–57	52	30
	58+	89	36
	Did not report	0	15
Education	Bachelor's degree or higher	170	167
	No bachelor's degree	130	190
	Did not report	0	56

Table 6.3 Sample Statistics by Socio-demographic Level

Data Analysis

Pre-processing

We prepared our dataset for analysis by removing 49 respondents that did not complete both sets of vignettes, 13 respondents that did not provide any demographic information, and one respondent that failed the attention check. We additionally removed one low quality (i.e., same answer for every question without justification in the open-ended questions) submissions and 12 duplicate submissions. For those that had one incomplete and one complete submission, we preserved the complete submission and discarded the incomplete one; for those that had two complete submissions, we preserved the first submission and discarded the second. We imputed missing responses (i.e., randomly skipped questions) using the mice package in R, a common method in

social science research to handle missing data [277, 844].

Due to the size of our sample, it was necessary to condense groupings of the socio-demographic levels collected in the post-test (provided in Appendix A.4). Due to race/ethnicity mapping and value differences between our pre-screener and Prolific's categories described in Section 6.3.0.6, participants reporting mixed or multiple race/ethnicities were grouped according to either their non-white race/ethnicity or primary ethnicity in order to preserve the most data integrity. For example, participants identifying as white and Latine in the prescreener had inconsistent race/ethnicity values reported by Prolific (e.g., some "white", some "mixed", some "other"); to ensure data consistency and in acknowledgement of historical controversies in US reporting of Latine racial categories as white [729], we coded these participants' race/ethnicity as Latine. Participants reporting multiple non-white ethnicities in our pre-screener were grouped according to the primary race/ethnicity reported in their Prolific profile, as data in these cases did not have the same inconsistencies.

Factors

For both our representative and minoritized samples, we regressed the contextual, socio-demographic, and individual privacy belief variables of interest on participants' reported comfort level to each scenario. Table 6.4 lists the factors used in our analysis:

Factors	Levels
Contextual	Context (Employment vs. Healthcare)
	Data Input
	Purpose
Socio-demographic	Race/Ethnicity
	Gender
	Mental Health Status
	Educational Attainment
Privacy Belief	General Privacy Concerns
	Trust in Employer/Healthcare Provider Handling Sensitive Information
	Perceived Sensitivity of Emotional Information in Employment/Healthcare

Table 6.4 Factors and Levels

For the socio-demographic categorical variables, we re-leveled the reference categories so that the results would compare levels to the most socially dominant group in each category, which we defined as white race/ethnicity, male gender, age 58+, no mental illness experience, and educational attainment of Bachelor's degree or higher. For the contextual categorical variable of

purpose, we defined the reference category as “giving employers/healthcare providers data-driven understanding into employee/patient wellbeing” given the prevalence of organizational initiatives to drive employment and healthcare decisions with data, including those providing insights into workers’ [637] and patients’ [533] emotional state.

For individual privacy beliefs reported in the post-test, we averaged participants’ reported value (ranging from 0-100) across each construct: general privacy concerns, trust in employer/healthcare provider handling of sensitive information, and perceived sensitivity of emotional information handled by employer/healthcare provider. Responses to some questions were first reverse-coded as necessary (e.g., if the higher value for the question indicated the opposite direction of the belief measured).

Mixed Effect Modeling

Our analysis takes a comprehensive approach to understanding how each of the contextual, socio-demographic, and individual privacy belief factors interdependently influence emotional privacy judgments. We conducted the quantitative analysis in R using multi-level modeling techniques with the lme4 package. As our factorial vignette design obtained multiple observations from each participant, the multi-level modeling approach clusters the analysis by participant, which allowed our analysis to account for individual variation within participant responses and avoid violating the independence assumption in traditional linear regression approaches [319]. This structure specifies individual participants as a random effect to account for subject to subject variability, thus limiting biased covariance estimates for each participant, and specifies our independent variables of interest as fixed effects [355, 319].

We fitted four multivariable linear mixed-effects models: one for responses to each employment and healthcare vignette sets, for both the representative and minoritized samples. To facilitate comparisons between samples, and because our individual privacy belief variables collected responses that were specific to and varied by either the employment or healthcare context, it was necessary to run separate models for both samples and vignette contexts.

To assess the best model fit for each dataset [98], we used Anova to compare various model combinations that specified individual participants as a random effect, included fixed effects for our contextual variables of interest (purpose and data input), and additionally included fixed effects combinations that varied by what socio-demographic and individual privacy belief variables were included. The Anova function conducts likelihood ratio tests (LRTs) to compare the likelihoods of multiple models and assess whether including or excluding certain fixed effects significantly improve model fit. We used LRTs along with Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values for each model to select models based on fit [693, 98]. We fit our models using maximum likelihood (ML), which means that estimates for the specified random

and fixed effect parameters were generated by maximizing the likelihood of the observed data; fitting models with ML rather than lmer's default restricted maximum likelihood (REML) criterion is necessary to meaningfully compare models with varying fixed effects structures [693].

To facilitate model comparisons to investigate our research questions, we chose to employ a model with the same fixed effects across all four models. We included a variable as a fixed effect if our ANOVA analysis showed it was a significant predictor in at least one of the four datasets and a variable that contributed to the best model fit, with the exception of respondents' general privacy concerns. Respondents' general privacy concerns were not a predictor in any of the four models, but we chose to include given our interest in privacy perceptions. During this process we opted to exclude the socio-demographic variables for age group, perceived socio-economic status, and employment status as fixed effects in our models. We conducted t-tests using the Satterthwaite's method to assess statistical significance [755], as it is generally inappropriate to use traditional p-values to assess the significance of fixed effects in mixed effect models [545].

We additionally used Anova to compare our chosen model to each of the four datasets' respective null models (containing no predictors), confirming our final models' fitness. To assess the proportion of variance explained by the model's structure, we computed the intra-class correlation (ICC) for the null models of both datasets [432]; the ICC for the representative and minoritized models was .72 and .67 respectively, indicating fair to good reliability [783]. For all models, we plotted the residuals to the quantiles of the standard normal distribution to confirm that the normality assumption was met [860]; although not all variables were normally distributed, the linear mixed effect analysis we employed is suitable for both normal and non-normal variables [431, 71].

We ultimately selected (and report on) a random slope mixed effects model for all datasets. This model provides a nuanced understanding of how our study's contextual, socio-demographic, and individual privacy belief variables of interest influence workers' and patients' emotional privacy judgments by recognizing that individuals may have unique responses to the explanatory variables, and that the relationship between the independent variables and participants' reported comfort can differ from person to person, by assigning distinct baseline values for each participant and allowing the effects of the independent variables to have a different effect for each participant. Specifically, our chosen model treats individual participants as random effects, and as described in Table 6.4, includes the following explanatory variables as fixed effects: contextual variables of data input and purpose; socio-demographic variables of gender, race/ethnicity, mental health status, and educational attainment; and individual privacy beliefs concerning general privacy, trust toward employer/healthcare provider handling of sensitive information, and perceived sensitivity of emotion data use in employment/healthcare. Participants' individual privacy beliefs regarding the riskiness of their employer/healthcare provider handling of sensitive information was found to be a predictor, but removed from the analysis due to multicollinearity with individual trust beliefs.

For each factor, we compare the relative magnitude and strength of the relationship between samples using Z-tests; a positive Z-score indicates the factor effect is greater in the U.S. representative group compared to the minoritized sample, while a negative Z-score indicates the factor effect is relatively greater in the minoritized group. We identify significantly different variable effects between U.S. representative and minoritized samples where the absolute value of the Z-score is greater than the critical value (e.g., 1.96 for a 5% significance level). The variation between samples reveals meaningful differences in how distinct contextual, socio-demographic, and individual privacy belief factors influence U.S. representative and minoritized perspectives differently, even where the variation is not significantly different between samples or where the effect of some predictors is not strong enough within each sample to be significant on its own.

Qualitative Responses

We conducted iterative qualitative analysis to analyze participants' answers to three open-ended questions described in 6.3.0.2.

Codebooks were developed separately for each of the two contexts. We developed a codebook through several coding exercises to create a common understanding among the research team. Five coders trained in qualitative coding individually and independently open-coded a random subset of 50 participants' responses, followed by a meeting to discuss and refine codes. In a second exercise, the team then applied the revised codebook to a separate random sample of 35 responses, and met to finalize the codebook and ensure team agreement.

Once we finalized the codebook, we established inter-rater reliability (IRR) [591] as follows. Two coders separately coded a newly selected random subset of 20 responses using the final codebook. Using functionality available in ATLAS.ti, we measured IRR with Krippendorff's alpha binary. Tables 6.5 and 6.6 include the alpha binary for the codebook themes for each context and the average of the relevant alpha binary values. We established IRR after reaching a score above .75 [591], deemed as "acceptable," after two rounds of coding data and measuring IRR. To identify and resolve disagreements after the first round, the two coders met to discuss any discrepancies, shared perspectives and rationales, and reached consensus to ensure similar understanding and application of codes moving forward.

After we established IRR, we divided the remaining data among the same two coders. Though they used the established codebook in this final coding round, the two coders could add new codes to mark for discussion with the rest of the team. This choice ensured that our analysis remained open and flexible. However, no new codes surfaced in these processes. After coding the remaining data, the whole research team met to identify and refine resulting themes surrounding data subjects' perceived risks and benefits associated with emotion AI in the workplace and healthcare.

Codebook themes	Alpha binary
Perceived potential benefits of emotion AI use in the workplace	0.735
Perceived potential concerns of emotion AI use in the workplace	0.85
Average alpha binary across relevant themes	0.7925

Table 6.5 Alpha binaries and average of alpha binaries of codebook themes relevant to the workplace context

Codebook themes	Alpha binary
Perceived potential benefits of emotion AI in healthcare	0.881
Perceived potential concerns of emotion AI in healthcare	0.837
Average alpha binary across relevant themes	0.859

Table 6.6 Alpha binaries and average of alpha binaries of codebook themes relevant to the healthcare context

Reflections, Limitations, and Opportunities

6.3.0.8 Vignette Responses

Our design elicited workers' and patients' self-reported comfort with being subject to various applications of automatic emotion inferences as a measure of their emotional privacy judgments. We framed vignettes as neutrally as possible, avoiding references to potential harms. However, some purposes (e.g., enhancing safety or mental health support) may have implied benefits, which could have influenced judgments [117]. Future work could test framing effects more explicitly.

Our use of a continuous Visual Analog Scale (VAS) for the dependent variable reduced common limitations of ordinal scales, such as data loss and clustering (see Section 6.3.0.1). While standard limitations of self-reported data apply, factorial vignette designs mitigate respondent bias by varying factors across scenarios, making it difficult for participants to systematically adjust responses [573].

6.3.0.9 Model Variables and Missing Factors

We recognize that privacy judgments are shaped by a wide range of contextual and individual factors. Our models focused on contextual integrity parameters, socio-demographic identities, and privacy beliefs most relevant to our research questions. Our vignettes specified consent and data handling parameters consistent with typical workplace and healthcare data practices. However, real-world implementations may involve organizational and institutional cultures, different consent dynamics, data sharing policies, or types of emotional information, which could affect privacy judgments in addition to individual variables such as privacy awareness or technological literacy. Such factors were beyond the scope of this study but merit future investigation.

6.3.0.10 Generalizability and Sample Limitations

While our U.S. representative sample followed standard demographic balancing procedures and our minoritized sample intentionally centered minoritized perspectives, neither fully captures the diversity of experiences within these participant groups.

Participants were recruited from Prolific—click workers who are often over-represented in research and whose privacy perceptions may differ from the broader population. Nonetheless, recent scholarship indicates that Prolific samples are generally representative in studies of privacy perceptions [823], supporting the validity of findings drawn from our U.S. representative sample.

Finally, our decision to combine data input types and to examine emotion inferences without specifying emotion categories facilitated a manageable vignette design and minimized participant fatigue. However, these choices necessarily limit the granularity of our findings. Future research should explore how privacy judgments vary across more specific data modalities and emotion types.

6.3.0.11 Statistical Considerations

Our mixed-effects models balanced theoretical relevance with statistical rigor, accounting for individual variability and interdependent predictors. As expected in models incorporating multiple variables and random effects, some factors showed non-significant associations [126, 693]. We interpret these conservatively and report confidence intervals to avoid dichotomous significance testing [126, 693]. Where appropriate, we discuss notable patterns that may have theoretical significance [126, 545].

6.4 Comparing Normative Judgments of Emotional Privacy

Our study systematically dissects the complex interplay of factors influencing emotional privacy judgments toward technologies that infer and interact with human emotion in workplace and health-care settings. Using mixed-effects modeling, we examine how contextual, socio-demographic, and individual privacy belief factors differentially influence workers' and patients' comfort levels. Our findings synthesize insights crucial to privacy theory, human-computer interaction, and technology policy, enhancing our understanding of emotional privacy amid growing AI-driven practices.

Recognizing that privacy perceptions vary across contexts and between dominant (U.S. representative) and minoritized groups [590, 646], we highlight these variations to underscore the multi-dimensional nature of emotional privacy judgments. Our rigorous methodological framework (see Sections 6.3 and 6.3.0.7) enables meaningful comparisons in emotional privacy judgments and identification of significant trends and differences.

Regression results, summarized in Tables 6.7 (employment) and 6.8 (healthcare), present coefficients, standard errors, and statistical significance across key variables.

Regression Results for Employment Context

	Representative (n=300)	Minoritized (n=385)	Z-Test (comparison)
(Intercept)	36.64 (6.44)***	34.24 (6.08)***	0.27
<u>Contextual Factors</u>			
Data Input (baseline: image/video)			
speech/text	2.69 (0.35)***	4.25 (0.34)***	-3.21
Purpose (baseline: data-driven insights)			
(2) academic research	4.18 (0.93)***	1.26 (0.89)	2.28
(3) early diagnosis – mental illness	-1.32 (0.93)	-2.49 (0.89)**	0.91
(4) early diagnosis – neurological	0.55 (0.93)	-1.70 (0.89)	1.75
(5) avoid human subjectivity	0.57 (0.93)	-1.39 (0.89)	1.52
(6) indiv. level inference	-3.48 (0.93)***	-3.70 (0.89)***	0.17
(7) group level inference	2.63 (0.93)**	2.16 (0.89)*	0.36
(8) identify those needing support	2.15 (0.93)*	3.78 (0.89)***	-1.27
(9) infer risk to others	6.39 (0.93)***	7.03 (0.89)***	-0.50
(10) infer risk of self-harm	3.20 (0.93)***	2.60 (0.89)**	0.47
(11) auto. intervention – therapy	1.68 (0.93)	1.92 (0.87)*	-0.19
(12) auto. intervention – acute support	1.66 (0.93)	1.85 (0.89)*	-0.14
(13) alert employer	0.28 (0.93)	-0.05 (0.89)	0.26
(14) assess performance	-0.88 (0.93)	-2.55 (0.89)**	1.30
<u>Socio-demographic Factors</u>			
Race/Ethnicity (baseline: white)			
Asian	-3.15 (4.31)	-8.05 (3.55)*	0.88
Black	5.64 (3.19)	7.38 (2.79)**	-0.41
Latine	8.27 (5.44)	4.45 (3.84)	0.57
Other races/ethnicities	6.78 (6.27)	3.53 (4.69)	0.41
Gender (baseline: male)			
trans and/or non-binary	-0.63 (8.71)	-4.63 (4.08)	0.42

	Representative (n=300)	Minoritized (n=385)	Z-Test (comparison)
woman	-2.99 (2.41)	-0.06 (2.45)	-0.85
Mental Health (baseline: no history)			
under treatment	6.47 (3.13)*	-0.79 (3.06)	1.66
resolved/untreated	-3.67 (3.36)	2.17 (2.97)	-1.30
Education (baseline: Bachelor+)			
no Bachelor's degree	0.14 (2.46)	6.16 (2.37)**	-1.76
<u>Privacy Beliefs</u>			
general privacy concerns	-0.04 (0.07)	-0.07 (0.07)	0.34
emotion data sensitivity	-0.30 (0.05)***	-0.25 (0.05)***	-0.71
trust in employer - sensitive info.	0.54 (0.05)***	0.40 (0.05)***	2.09
AIC	71657.54	93685.87	
BIC	71861.58	93911.72	
Log Likelihood	-35799.77	-46811.94	
Observations	8400	10780	
Groups (participants)	300	385	
Var: Intercept	394.19	417.98	
Var: Residual	257.45	303.61	

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; † $p < 0.1$. **Bold** Z-scores indicate statistical significance between samples at $p < 0.05$.

Table 6.7: Regression results: Comfort with Emotion Inferences in Employment

Regression Results for Healthcare Context

	Representative (n=300)	Minoritized (n=385)	Z-Test (comparison)
(Intercept)	28.91 (8.21)***	26.96 (6.72)***	0.18
<u>Contextual Factors</u>			
Data Input (baseline: image/video)			
speech/text	4.13 (0.37)***	5.35 (0.36)***	-2.37
Purpose (baseline: data-driven insights)			
(2) academic research	-2.61 (0.98)**	-2.43 (0.95)*	-0.13
(3) early diagnosis – mental illness	-1.98 (0.98)*	-0.55 (0.95)	-1.05
(4) early diagnosis – neurological	2.19 (0.98)*	3.47 (0.95)***	-0.94
(5) avoid human subjectivity	-3.73 (0.98)***	-2.44 (0.95)**	-0.94
(6) individual-level inference	-5.52 (0.98)***	-4.72 (0.95)***	-0.59
(7) group-level inference	-4.32 (0.93)***	-4.74 (0.95)***	0.31
(8) identify those needing support	-1.17 (0.98)	1.34 (0.95)	-1.84
(9) infer risk to others	-0.26 (0.98)	-0.56 (0.95)	0.22
(10) infer risk of self-harm	-0.27 (0.98)	-1.10 (0.95)	0.61
(11) auto. intervention – therapy	-7.91 (0.98)***	-7.88 (0.95)***	-0.02
(12) auto. intervention – acute support	-3.49 (0.98)***	-3.18 (0.95)***	-0.23
(13) alert provider	-3.89 (0.98)***	-2.09 (0.95)*	-1.32
(14) assess overall health	-1.91 (0.98)	0.23 (0.95)	-1.57
<u>Socio-demographic Factors</u>			
Race/Ethnicity (baseline: white)			
Asian	3.33 (5.44)	-3.90 (3.92)	1.08
Black	10.65 (4.02)**	6.66 (3.05)*	0.79
Latine	4.60 (6.87)	5.47 (4.29)	-0.11
Additional races/ethnicities	3.01 (7.93)	0.95 (5.18)	0.22
Gender (baseline: male)			
trans and/or non-binary	-6.26 (10.99)	-15.32 (4.55)***	0.76
woman	-1.52 (3.03)	-0.07 (2.71)	-0.36
Mental Health (baseline: no history)			

	Representative (n=300)	Minoritized (n=385)	Z-Test (comparison)
under treatment	3.69 (3.94)	1.70 (3.33)	0.39
resolved/untreated	-0.12 (4.23)	3.32 (3.31)	-0.64
Education (baseline: Bachelor+)			
no Bachelor's degree	2.06 (3.06)	2.12 (2.63)	-0.01
<u>Individual Privacy Beliefs</u>			
general privacy concerns	-0.04 (0.08)	-0.08 (0.07)	0.35
emotion data sensitivity	-0.10 (0.05)	-0.11 (0.04)**	0.15
trust in provider - sensitive info.	0.44 (0.06)***	0.53 (0.05)***	-1.07
AIC	72732.10	94520.25	
BIC	72936.15	94745.93	
Log Likelihood	-36337.05	-47229.12	
Observations	8400	10724	
Groups (participants)	300	385	
Var: Intercept	632.11	511.98	
Var: Residual	288.95	342.41	

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; † $p < 0.1$. **Bold** Z-scores indicate statistical significance between samples at $p < 0.05$.

Table 6.8: Regression results: Comfort with Emotion Inferences in Healthcare

Together, these analyses provide a comprehensive overview of how contextual, socio-demographic, and privacy belief factors influence emotional privacy judgments about emotion AI in the workplace and healthcare.

Table 6.9 distills these results, complemented visually by the coefficient plot in Figure A.2 (Appendix A.7).

Factor	Key Findings	Sample Differences (Rep. vs Minoritized)
Context	Lower baseline comfort in employment than healthcare. Contextual factors exert greater influence in healthcare.	Minoritized groups generally reported lower comfort across contexts.
Data Input	Speech/text preferred over image/video across both contexts.	Stronger preference for speech/text in minoritized samples.
Purpose	Group-level inferences, harm prevention, and academic research (employment only) raised comfort. Individual assessments and mental-health diagnostics reduced comfort, especially in healthcare.	Minoritized groups showed lower comfort across most purposes; slightly higher trust in neurological diagnostics in healthcare.
Race/Ethnicity	Black participants reported higher comfort across contexts; Asian participants lower comfort in employment.	Black participants more positive; Asian and other minoritized participants more cautious, particularly in employment.
Gender	No significant effects except trans/non-binary participants reporting lower comfort in healthcare (especially in the minoritized sample).	Strong negative effect for trans/non-binary in healthcare.
Mental Health Status	Current treatment increased comfort in employment (representative sample only). No significant effects in healthcare.	Effect attenuated or reversed in the minoritized group.
Education	Lower educational attainment linked to higher comfort in employment (significant in minoritized sample).	Larger positive effect in minoritized sample.
General Privacy Concerns	No significant effects.	N/A
Trust Beliefs	Higher trust increased comfort in both contexts.	Stronger effect in representative sample (employment); stronger in minoritized sample (healthcare).
Perceived Sensitivity of Emotional Data	Higher perceived sensitivity linked to lower comfort across contexts.	Effect confirmed in minoritized sample (healthcare); similar trend elsewhere.

Table 6.9 Summary of Key Quantitative Findings Across Factors and Sample Comparisons
153

6.4.1 Contextual Vulnerability: Setting, Data Input, and Purpose

In examining emotional privacy judgments concerning emotion inferences, we focused on three key contextual variables: context ($\$C$), data input ($\I), and purpose ($\$P$). Participants evaluated tailored vignettes that varied by these variables as follows:

As a $\$C1$, rate your comfort (from 0 = “very uncomfortable” to 100 = “very comfortable) with your $\$C2$ using a computer program to automatically detect your emotional states using records of $\$I$ recorded from your daily activities and device use, for the purpose of $\$P$

Our analysis assesses how data input ($\$I$) and purpose ($\P) shape emotional privacy judgments within each context.

6.4.1.1 Context: Emotional Privacy Judgments More Susceptible to Factor Influences in Healthcare than in Employment

Privacy perceptions differed substantially between employment and healthcare contexts (Table 6.10).

Context	Sample	Mean	Mean StdDev	Regression Intercept
employment	representative	32.50	32.59	36.64
employment	minoritized	32.55	32.11	34.24
healthcare	representative	49.70	32.45	28.91
healthcare	minoritized	50.02	32.54	26.96

Table 6.10 Summary Statistics – Mean and Estimated Comfort Levels by Context and Sample

Mean comfort was markedly lower in employment (32.50/32.55) than in healthcare (49.70/50.02). However, the regression intercepts—which control for all other variables—reveal that in healthcare, baseline comfort was even lower (28.91/26.96). This suggests that factors in our model had a greater impact on comfort levels in healthcare than in employment, where intercepts were closer to the mean comfort levels.

These differences reflect distinct power dynamics and privacy expectations. Healthcare is anchored in trust and confidentiality, particularly around mental health, where subjective emotional disclosures are central. This reliance may amplify privacy sensitivities, especially among minoritized groups who have faced inequitable care or stigma. Our related qualitative findings confirm heightened concerns about emotion AI’s potential to undermine autonomy, care access, and the patient-provider relationship [733].

By contrast, employment reflects normalized surveillance and limited worker autonomy, contributing to baseline discomfort with emotion inferences. Qualitative data indicate that workers—especially from minoritized groups—view such technologies as likely to exacerbate existing privacy and power disparities [219].

These patterns underscore the contextual variability of emotional privacy judgments and validate our use of mixed-effects modeling to disentangle how contextual, socio-demographic, and belief factors shape these judgments. Lower healthcare intercepts indicate that such factors exert stronger influences in healthcare, where participants expressed overall higher comfort yet rejected most specific purposes for emotion inference—especially those that undermined autonomy or discretion. In contrast, while employment settings evoked lower baseline comfort, participants differentiated sharply between acceptable and unacceptable uses. Some purposes—such as group-level inferences or harm prevention—elicited positive responses, suggesting conditional acceptance even in surveillance-prone environments. Yet overall, emotion inferences in employment remained a source of concern, particularly given the risks of employer misuse and the potential to reinforce existing power asymmetries.

6.4.1.2 Data Input: Workers and Patients Favor Speech/Text Emotion Recognition Over Facial Emotion Recognition, though Emotional Privacy Judgments Remain Low with All Modalities

We examined whether and how participants' comfort with emotion inferences varied by the type of data input to the emotion recognition algorithm. From their perspectives as employees and patients, participants rated their comfort (from 0 = “very uncomfortable” to 100 = “very comfortable) with their employers and healthcare providers using a computer program to detect their emotional states from either (1) speech/text records—what they say (verbally or written/typed) and how they say it (e.g., speed or tone)—or (2) images/video of their facial expressions, for various purposes.

Our regression results show that both workers and patients were significantly more comfortable with speech/text-based emotion inferences than with those based on facial recognition. Compared to the baseline category of image/video records, speech/text inputs were associated with significantly higher comfort in both employment (representative: $\beta = 2.69$, $SE = 0.35$, $p < 0.001$; minoritized: $\beta = 4.25$, $SE = 0.34$, $p < 0.001$) and healthcare (representative: $\beta = 4.13$, $SE = 0.37$, $p < 0.001$; minoritized: $\beta = 5.35$, $SE = 0.36$, $p < 0.001$). This may reflect public discomfort with facial recognition technologies and their attendant accuracy and privacy concerns [922].

Notably, although speech/text inputs raised comfort relative to facial recognition, predicted comfort levels across all data inputs remained low—ranging from 32.31 to 39.33 on a 0–100 scale (Figure 6.2). The more pronounced positive effect of speech/text was statistically significant in both employment and healthcare, with Z-scores of -3.21 and -2.37, respectively.

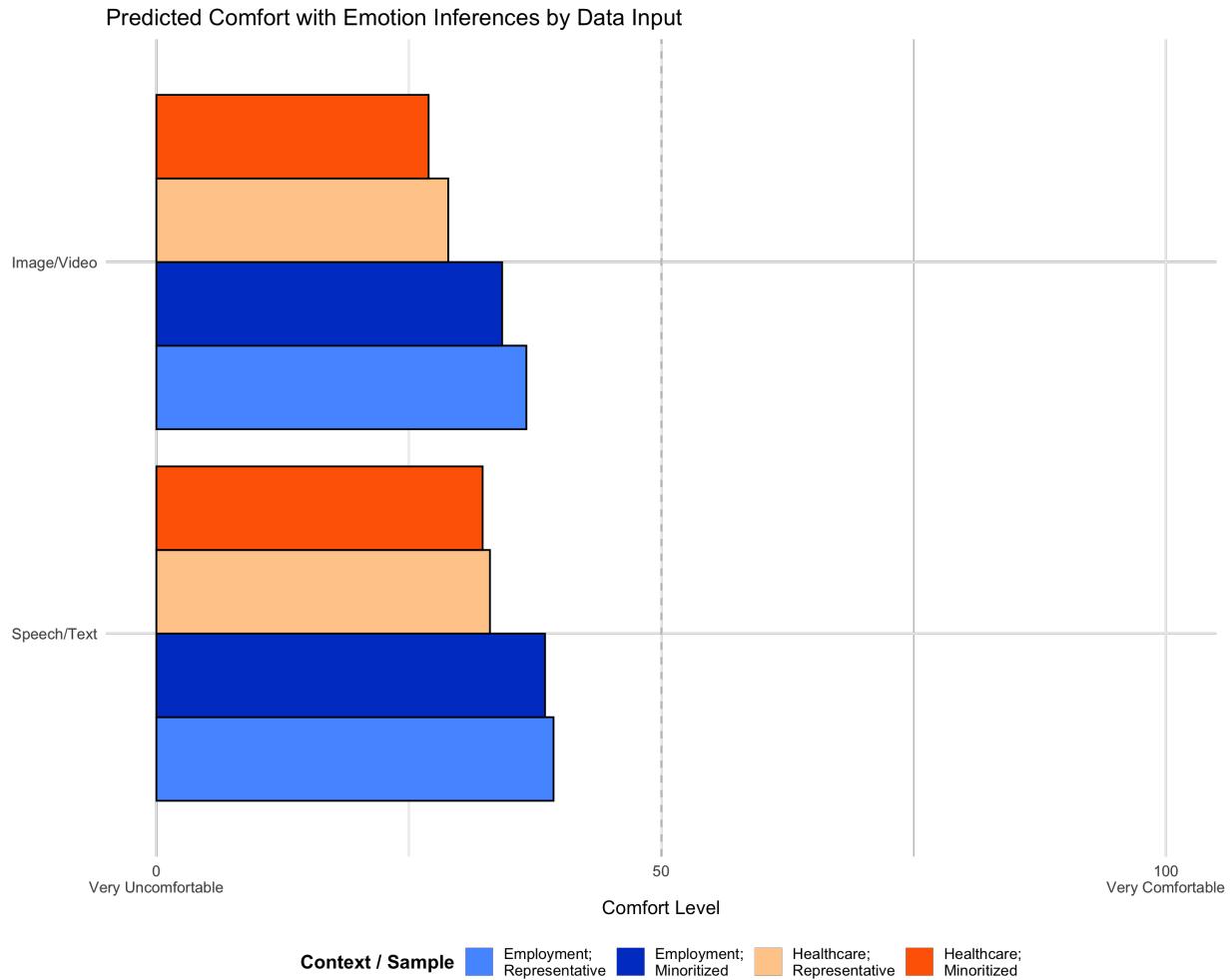


Figure 6.2: Predicted Comfort Levels by Data Input Type. This figure illustrates the predicted comfort levels by combining the data type variable coefficients to each mixed-effects regression model intercept, derived by analyzing respondent comfort on a scale from 1 (very uncomfortable) to 100 (very comfortable). Bars with black borders indicate statistically significant results.

These findings support a growing recognition that facial recognition technologies—including facial *emotion* recognition—are widely viewed with suspicion. They also confirm that data input is a meaningful and statistically significant contextual factor shaping emotional privacy judgments. However, this does not suggest that emotional privacy can be preserved by avoiding facial inputs alone. Even with speech/text, predicted comfort remained low.

Importantly, the effect of data input was more pronounced for participants in the minoritized sample, who consistently reported lower comfort across both contexts and all input types. This pattern underscores greater emotional privacy concerns about all forms of emotion recognition—speech, text, and facial—among people of color, people with mental illness, and/or minority genders compared to the U.S. representative cohort.

6.4.1.3 Purposes for Which Employers and Healthcare Providers Use Emotion Inferences Shape Emotional Privacy Judgments

To assess the influence of purpose on emotional privacy judgments, we examined whether and how participants' comfort varied across fourteen distinct purposes for which employers and healthcare providers might use emotion inferences (Table 6.2). We report how participants rated their comfort (0 = “very uncomfortable” to 100 = “very comfortable) relative to a common baseline: providing data-driven insights into employee or patient wellbeing. For interpretive clarity, we grouped the fourteen purposes into higher-level themes (Table 6.11).

Our findings demonstrate that purpose significantly shapes emotional privacy judgments, with effects varying by specific purpose, context, and participant group. Generally, purposes that reinforced each context's social mission or aligned with its privacy expectations were judged more positively, while purposes that strained those expectations were judged more negatively. As prior work shows, perceived technological benefits and risks can both influence privacy perceptions [117].

To contextualize these findings, we draw on qualitative analyses of perceived benefits and risks voiced by study participants, drawn from their open-ended responses (Section 6.3.0.2) and published in related studies on employment [219] and healthcare [733].

Figure 6.3 visualizes predicted comfort levels for each purpose.

Facilitating Earlier Diagnosis of Neurological Disorders and Mental Illness One proposed use case for emotion inferences involves facilitating earlier medical diagnosis. This application has been suggested for healthcare and, increasingly, the workplace—given the extensive time people spend at work and the rise of surveillance systems already collecting data from which emotional features might be extracted [497, 679, 164]. We examined how using emotion inferences to detect mental illnesses and neurological disorders earlier than otherwise possible influenced participants'

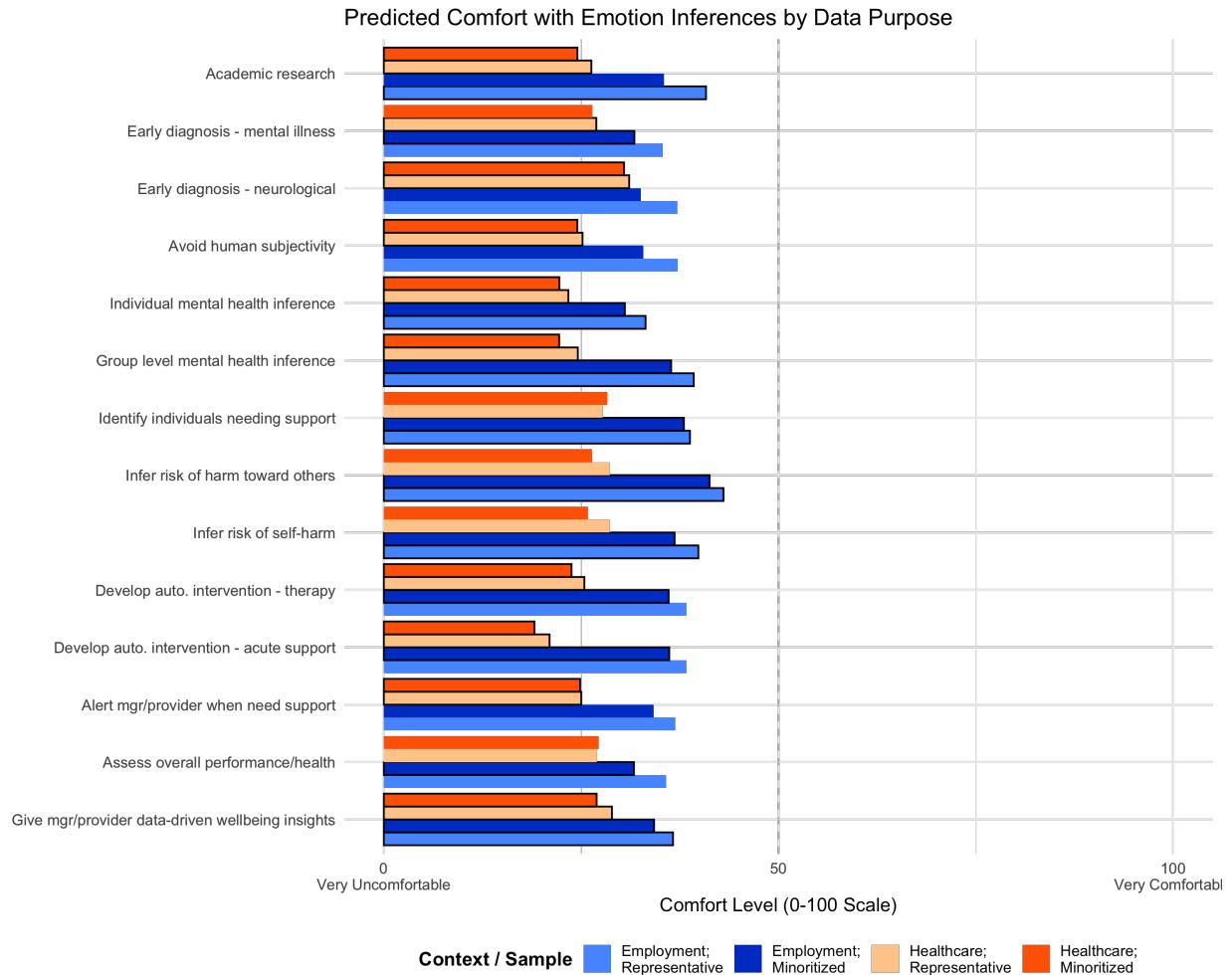


Figure 6.3: Predicted Comfort Levels by Purpose. This figure illustrates the predicted comfort levels by combining the purpose variable coefficients to each mixed-effects regression model intercept, derived by analyzing respondent comfort on a scale from 1 (very uncomfortable) to 100 (very comfortable). Bars with black borders indicate statistically significant results.

Purpose Grouping	Purpose Levels
Early diagnosis of mental illness and neurological disorders	(3) Diagnose mental illness in (\$C1) earlier than otherwise possible. (4) Diagnose neurological disorders (e.g., dementia or ADHD) in (\$C1).
Augment employee and patient assessments	(5) Avoiding subjectivity in other methods your (\$C2) may use to learn about your emotional state, like a survey or your (\$C2)'s observation. (14) Assessing the (\$C3) of individual (\$C1).
Individual and group-level mental health inferences	(6) Inferring the mental health state of (\$C1) individually. (7) Inferring the mental health state of (\$C1). An individual's mental health will not be inferred; only group-level inferences will be made.
Societal benefit	(2) Sharing that information with academic researchers to help them learn more about mental health, as part of a research partnership. (8) Identifying (\$C1) in need of mental health support, to better plan organizational mental health resources.
Harm prevention	(9) Inferring whether (\$C1) are at risk of harming others. (10) Inferring whether (\$C1) are at risk of harming themselves.
Supportive interventions	(11) Developing an intelligent computer program, such as a chatbot, that can conduct mental health therapy with (\$C1), including you. (12) Inferring moments (\$C1) may be in need of emotional support and responding with an intelligent computer program designed to help (\$C1) improve their wellbeing, such as offering wellbeing tips. (13) Automatically alerting your (\$C2) when (\$C1)s may need support, including you.
Baseline purpose	(1) Giving (\$C2) data-driven insights into (\$C1) wellbeing.

Table 6.11 We examined the impact of 14 purposes for which employers and healthcare providers may use emotion inferences, grouped into higher level themes to aid interpretation.

comfort. Participants rated their comfort (0 = “very uncomfortable” to 100 = “very comfortable”) with their employers or healthcare providers using emotion inferences for:

- (3) *diagnosing neurological disorders, such as dementia or ADHD, in employees/patients earlier than otherwise possible; and*
- (4) *diagnosing mental illness in employees/patients earlier than otherwise possible.*

Predicted comfort levels for both diagnostic purposes remained low across contexts and samples—ranging from 31.75 to 37.19 for workers and 26.41 to 31.1 for patients. As shown

in Figure 6.3, comfort was consistently lower in healthcare than in employment, reflecting heightened privacy concerns about emotion inferences in clinical settings.

Earlier diagnosis of mental illness. Across both contexts and samples, using emotion inferences to detect *mental illness* had a negative effect on comfort compared to the baseline purpose. While both workers and patients expressed discomfort with this application, differences emerged by context and sample.

Employment context. For employment, the negative effect was statistically significant only in the minoritized sample (representative: $\beta = -1.32$, $SE = 0.93$, not significant; minoritized: $\beta = -2.49$, $SE = 0.89$, $p < 0.01$). Qualitative findings suggest this may reflect greater privacy concerns about employer access to mental health information among minoritized participants [219].

Healthcare context. In healthcare, the negative effect was more pronounced and statistically significant only in the representative sample (representative: $\beta = -1.98$, $SE = 0.98$, $p < 0.05$; minoritized: $\beta = -0.55$, $SE = 0.95$, not significant). The smaller effect in the minoritized group may relate to disparities in mental healthcare quality. Participants from minoritized backgrounds reported difficulties getting providers to recognize their mental health concerns and noted that emotion inferences might help legitimate issues that might otherwise be ignored [733]. Nonetheless, predicted comfort remained low in both samples (26.41 for the minoritized sample and 26.93 for the representative sample), and the difference between them was marginal.

Notably, although early mental health diagnosis aligns with healthcare's broader goals, participants still viewed this purpose as diminishing their emotional privacy.

Earlier diagnosis of neurological disorders. By contrast, using emotion inferences to detect *neurological* disorders had markedly different effects.

Healthcare context. In healthcare, this purpose produced a significantly positive effect on comfort in both samples (representative: $\beta = 2.19$, $SE = 0.98$, $p < 0.05$; minoritized: $\beta = 3.47$, $SE = 0.95$, $p < 0.001$). It was the *only* purpose across all fourteen tested to have a significant positive effect on patient comfort. Despite generally low comfort with emotion inferences, participants viewed this use case as a limited exception that had a positive effect on emotional privacy judgments.

Although our qualitative data did not explicitly address this finding, it may reflect the greater availability of objective measures in neurological diagnostics (e.g., imaging, neurological exams), which could reduce perceived risks to patient autonomy compared to subjective mental health assessments. The stronger positive effect in the minoritized sample suggests these participants may perceive greater potential benefits—or lower risks—from this specific application. However, estimated comfort remained low overall (30.43 for the minoritized sample and 31.1 for the representative sample).

Employment context. In employment, effects were mixed. This purpose had no significant

effect in the representative sample ($\beta = 0.55$, $SE = 0.93$, not significant) but showed a weakly significant negative effect in the minoritized sample ($\beta = -1.70$, $SE = 0.89$, $p < 0.1$). The larger and significant negative effect in the minoritized sample suggests workers from minoritized backgrounds perceived higher risks—or fewer benefits—from employer use of emotion inferences for neurological diagnosis.

Coefficient plots (Table A.2) show that, within a 95% interval, the effect for the representative sample crossed zero, while the effect for the minoritized sample did not. This indicates that although the direction of the effect is uncertain for representative participants, it can be confidently interpreted as negative for minoritized workers.

Qualitative data from minoritized participants underscore this discomfort, citing fears of negative personal and professional consequences tied to health disclosures in the workplace [219].

Employee and Patient Assessments Scholars and technologists have proposed automatic emotion inferences as a potentially objective method to reduce bias in both employee [734] and patient [919] assessments. Rather than relying on self-reports or human observations, incorporating presumably objective emotion inferences into work performance evaluations and health assessments is thought to minimize human subjectivity and the biases involved in understanding individuals' emotional states and their relation to overall work performance and health. We examined how purposes related to augmenting employee and patient assessments influenced participants' comfort with emotion inferences. Participants rated their comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with their employers or healthcare providers inferring their emotions for the following purposes:

(5) avoiding subjectivity in other methods of your employer/healthcare provider learning about your emotional state, like a survey or your employer/healthcare provider's observations; and (14) assessing the work performance/overall health of individual employees/patients.

Predicted comfort levels for these purposes were generally low. On a scale from 0–100, worker comfort ranged from 31.69 to 37.21, while patient comfort ranged from 24.52 to 27.19 (see Figure 6.3).

Employment context. Employers using emotion inferences to assess overall work performance had a negative effect on worker comfort compared to the baseline purpose, with significance observed only in the minoritized sample (representative: $\beta = -0.88$, $SE = 0.93$, insignificant; minoritized: $\beta = -2.55$, $SE = 0.89$, $p < 0.001$). The larger and significant negative effect in the minoritized sample likely reflects both the general trend in our results—that this sample perceives greater invasions to emotional privacy—and, possibly, increased statistical power from

the larger sample size. Qualitative insights from our adjacent studies suggest this discomfort may also reflect concerns among minoritized participants that emotional surveillance could impair work performance or lead to negative employment outcomes [219].

Employers using emotion inferences to reduce subjectivity in understanding workers' emotional states did not yield statistically significant effects in either sample. Taken together, these findings suggest that workers view employer use of emotion inferences for performance assessments as negatively affecting emotional privacy.

Healthcare context. Healthcare providers using emotion inferences to avoid human subjectivity in evaluating patients' emotional states had a significant negative effect on patient comfort in both samples (representative: $\beta = -3.73$, $SE = 0.98$, $p < 0.001$; minoritized: $\beta = -2.44$, $SE = 0.95$, $p < 0.01$). Although patient comfort was lower in the minoritized sample (24.52 vs. 25.18), the smaller and less significant negative effect in this sample suggests that people of color, individuals with mental illness, and/or minoritized genders may associate relatively higher benefit or reduced risk with this purpose. Our qualitative findings support this interpretation: minoritized participants expressed a desire for more objective, less biased evaluations of their emotional and mental health but also voiced concerns that algorithmic inferences could exacerbate provider bias in practice [733].

By contrast, healthcare providers using emotion inferences to assess overall patient health had no statistically significant effect in either sample, though a weakly significant negative effect (at the $p < 0.1$ level) was observed in the representative sample.

In summary, these results indicate that workers judged employer use of emotion inferences for performance assessments as negatively affecting their emotional privacy. Similarly, patients judged healthcare providers' use of emotion inferences to avoid subjectivity in emotional evaluations as negatively affecting their emotional privacy.

Inferring Mental Health at Individual and Group Levels We examined how participant comfort was affected by employers and healthcare providers using emotion inferences for the purpose of inferring workers' and patients' mental health at both individual and group levels. Participants were asked to rate their comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with their employers/healthcare providers inferring their emotions using various data inputs for the purposes of:

- (6) *inferring the mental health state of employees/patients individually; and*
- (7) *inferring the mental health state of employees/patients. Inferences of an individual's mental health will not be made; only at a group level.*

As Figure 6.3 illustrates, predicted comfort levels for both purposes were low across contexts and samples, with worker comfort ranging from 30.54—39.27 and patient comfort ranging from

22.22—24.59. In both contexts, predicted comfort levels were lower in the minoritized sample than in the U.S. representative sample.

Employment context. Employers using emotion inferences to infer individual workers' mental health had a significant negative effect on comfort in both samples, compared to the baseline purpose (representative: $\beta = -3.48$, $SE = 0.93$, $p < 0.001$; minoritized: $\beta = -3.70$, $SE = 0.89$, $p < 0.001$), with a slightly more pronounced negative effect in the minoritized sample. By contrast, employers inferring workers' mental health at a group level had a significantly positive effect on comfort in both samples (representative: $\beta = 2.63$, $SE = 0.93$, $p < 0.01$; minoritized: $\beta = 2.16$, $SE = 0.89$, $p < 0.05$), with a somewhat smaller effect in the minoritized sample. These results indicate that individual-level mental health inferences are discomforting and perceived as privacy invasive, whereas group-level inferences may be welcomed and seen as relatively privacy-preserving. Consistent with these results, participants in our adjacent qualitative study expressed enthusiasm for the potential of emotion AI to improve mental health support by helping employers identify workplace improvements or resources, tempered by deep concerns about misuse of individual-level emotional information—especially the risk of negative employment outcomes such as termination or loss of opportunities [219]. Our regression results suggest that aggregating emotion inferences may mitigate risks linked to identifiability, balancing perceived benefits with protections for workers' emotional privacy.

Healthcare context. In contrast, healthcare providers using emotion inferences to infer patients' mental health—at either the individual level (representative: $\beta = -5.52$, $SE = 0.98$, $p < 0.001$; minoritized: $\beta = -4.72$, $SE = 0.95$, $p < 0.001$) or the group level (representative: $\beta = -4.32$, $SE = 0.98$, $p < 0.001$; minoritized: $\beta = -4.74$, $SE = 0.95$, $p < 0.001$)—had a significant negative impact on patient comfort across both samples. The negative effect of individual-level inferences was slightly smaller in the minoritized sample than in the U.S. representative sample; the effects for group-level inferences were similar across samples. These results indicate that patients are significantly discomforted by mental health inferences regardless of identifiability. Qualitative insights from our related study provide explanatory context: patients expressed concerns that emotion AI could facilitate or worsen harmful mental healthcare practices, such as biased assessments, reduced patient voice, strained provider interactions, and misuse of sensitive information at both the individual and collective levels [733]. Notably, the positive effect associated with group-level inferences observed in the employment context was absent in healthcare. Although our qualitative data did not directly explain this pattern, we suggest that the inherently individualized nature of the patient-provider relationship may account for participants' reluctance to view group-level inferences as alleviating privacy concerns in healthcare.

Societal and Collective Benefit We examined how employers and healthcare providers using emotion inferences for purposes of societal or collective benefit—specifically, to benefit society by supporting academic research and to benefit workers and patients by identifying individuals requiring additional support to improve mental healthcare resource planning—affected participants’ comfort with emotion inferences. Participants rated their comfort (from 0 = “very uncomfortable” to 100 = “very comfortable”) with their employers/healthcare providers inferring their emotions using various data inputs for the following purposes:

- (2) *sharing that information with academic researchers to help them learn more about mental health, as part of a research partnership; and*
- (8) *identifying employees/patients in need of mental health support, to better plan organizational mental health resources*

Predicted comfort levels remained low overall. As shown in Figure 6.3, patient comfort (ranging from 24.53—28.3) was consistently lower than worker comfort (ranging from 35.5—40.82) across both purposes and samples.

Employment context. Employers using emotion inferences to support academic research had a positive effect on worker comfort relative to the baseline purpose, with larger and statistically significant effects in the U.S. representative sample only (representative: $\beta = 4.18$, $SE = 0.93$, $p < 0.001$; minoritized: $\beta = 1.26$, $SE = 0.89$, insignificant). While our adjacent study [219] did not surface specific insights explaining this pattern, the result aligns with prior qualitative work indicating that while people hold predominantly negative views toward automatic emotion recognition, their attitudes are less negative in specific use cases involving societal benefit, such as supporting academic research [53, 731]. The positive effect was significantly larger in the representative sample than in the minoritized sample, with a Z-score of 2.38, possibly reflecting heightened mistrust of academic research in minoritized communities due to historical patterns of exclusion and mistreatment [132].

For the purpose of identifying individuals in need of mental health support to inform organizational planning, this use case had a significantly positive impact on worker comfort relative to the baseline in both samples, with a larger effect in the minoritized sample (representative: $\beta = 2.15$, $SE = 0.93$, $p < 0.05$; minoritized: $\beta = 3.78$, $SE = 0.89$, $p < 0.001$). In contrast to the negative effects observed for individual-level emotion inferences in Section 6.4.1.3, this result suggests that workers’ discomfort can be mitigated when emotion inferences are used for purposes that do not assess individual mental health states directly and are instead linked to collective worker benefit. Qualitative results from our related study support this interpretation: nearly one-third of participants, most with minoritized identities, acknowledged potential benefits of using emotion inferences to improve organizational mental health resources and accommodations [219].

Overall, these results suggest that inferences of worker emotion—when used strictly for societal or collective worker benefit—may represent a limited acceptable use case that positively influences emotional privacy judgments. However, sample-level differences also underscore the importance of nuanced, personalized approaches to collecting, using, and sharing emotion inferences that respect diverse privacy needs and preferences.

Healthcare context. By contrast, purposes framed as societal or collective benefit did not preserve patient emotional privacy. Healthcare providers sharing emotion inferences with academic researchers had a significantly negative effect on patient comfort in both samples (representative: $\beta = -2.61$, $SE = 0.98$, $p < 0.01$; minoritized: $\beta = -2.43$, $SE = 0.95$, $p < 0.05$). For the purpose of identifying patients needing support to inform mental healthcare resource planning, the results were not statistically significant at the .05 threshold; however, sample comparisons revealed a statistically significant difference, with a negative effect in the representative sample and a comparatively positive effect in the minoritized sample ($Z = -1.84$).

Qualitative insights from our related study help explain these patterns. Participants, including many with minoritized identities, acknowledged the potential value of emotion inferences for advancing mental health research. However, they expressed strong concerns about data sharing practices, particularly fears that sharing inferred emotional information with third parties could compromise privacy and create barriers to care [733]. Additionally, higher expectations of confidentiality in the patient-provider relationship likely contributed to the negative effect of this purpose, in contrast to the more positive evaluations observed in employment.

Taken together, these results indicate that even when framed as benefiting society or patients collectively, sharing patient emotion inferences is discomforting and perceived as violating emotional privacy.

Harm Prevention We investigated whether and how employers and healthcare providers inferring workers' and patients' emotions for the purpose of preventing self-harm and harm toward others influenced comfort with emotion inferences. Participants rated their comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with their employers/healthcare providers inferring their emotions using various data inputs for the following purposes:

- (6) *inferring whether employees/patients are at risk of harming themselves; and*
- (7) *inferring whether employees/patients are at risk of harming others*

Predicted comfort levels for harm prevention were consistently lower in the healthcare context (25.86—28.65) than in the employment context (36.84—43.03), with lower estimates observed in the minoritized sample than in the U.S. representative sample, as shown in Figure 6.3.

Employment context. Employers' use of emotion inferences for harm prevention had a significantly positive effect on worker comfort relative to the baseline purpose. Notably, inferring risk of harm toward others had the largest positive effect on worker comfort of any purpose tested in both samples (representative: $\beta = 6.39$, $SE = 0.93$, $p < 0.001$; minoritized: $\beta = 7.03$, $SE = 0.89$, $p < 0.001$). Employers using emotion inferences to infer self-harm also had a significantly positive effect in both samples (representative: $\beta = 3.20$, $SE = 0.93$, $p < 0.01$; minoritized: $\beta = 2.60$, $SE = 0.89$, $p < 0.01$). Positive effects were similar between samples, suggesting that workers may view employer use of emotion inferences as acceptable for harm prevention purposes, provided that use is limited and justified.

Our adjacent qualitative study offers a nuanced interpretation of these findings. While some workers expressed support for monitoring employee emotions to prevent workplace violence or self-harm—acknowledging potential safety benefits—they emphasized that this would only be acceptable if emotion inferences were restricted strictly to this purpose and proven accurate. Participants expressed deep concern that employers might repurpose emotion inferences for unrelated or punitive uses, or that biased or inaccurate inferences could lead to false flags and unwarranted interventions, ultimately compromising worker safety rather than protecting it [219]. These findings underscore the importance of weighing the potential safety benefits of harm prevention against the serious risks of misuse and error.

Healthcare context. In contrast to the positive effects observed in employment, we found no statistically significant effects of harm prevention purposes on patient comfort in either sample. Trends indicated negative effects for both self-harm (representative: $\beta = -0.27$, $SE = 0.98$, insignificant; minoritized: $\beta = -1.10$, $SE = 0.95$, insignificant) and harm toward others (representative: $\beta = -0.26$, $SE = 0.98$, insignificant; minoritized: $\beta = -0.56$, $SE = 0.95$, insignificant).

Our qualitative analysis provides explanatory insight. Most participants did not perceive benefits to healthcare providers using emotion inferences for harm prevention and expressed substantial privacy concerns. They feared that such uses could legitimize over-surveillance of already vulnerable mentally ill patients and worried that inaccurate inferences could lead to severe consequences, such as unwarranted coercive interventions or involuntary commitment [733].

Supportive Interventions We examined how emotion inferences used for supportive interventions influenced participants' comfort with emotion inferences in workplace and healthcare contexts. Participants rated their comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with employers and healthcare providers inferring their emotions using various data inputs for the following purposes:

- (11) developing an intelligent computer program, such as a chatbot, that can conduct mental health therapy with employees/patients, including you;
- (12) inferring

moments employees/patients may be in need of emotional support and responding with an intelligent computer program designed to help employees/patients improve their wellbeing, such as offering wellbeing tips; and (13) automatically alerting your employer/healthcare provider when employees/patients may need support, including you.

Predicted comfort levels for these purposes were generally low. Worker comfort ranged from 34.19—38.32 and patient comfort ranged from 19.08—25.42, with lower levels observed in the healthcare context and the minoritized sample (Figure 6.3).

Employment context. Emotion inferences for supportive interventions had a positive impact on worker comfort when used to develop (representative: $\beta = 1.68$, $SE = 0.93$, $p < 0.1$; minoritized: $\beta = 1.92$, $SE = 0.89$, $p < 0.05$) and deliver (representative: $\beta = 1.66$, $SE = 0.93$, $p < 0.1$; minoritized: $\beta = 1.85$, $SE = 0.89$, $p < 0.1$) automated interventions that provided direct support, relative to the baseline purpose. However, these effects were only weakly significant. Effects were similar between samples. Our analysis did not identify a statistically significant effect for interventions involving third-party alerts to managers or employers.

These results suggest that workers may perceive potential benefits from emotion inferences used to develop or deliver direct wellbeing interventions—especially when such interventions remain private and do not involve employer oversight. Our qualitative study did not yield direct insights into this specific finding. However, workers expressed a general desire for improved wellbeing support while also voicing concerns that employer access to inferred emotional information could lead to negative personal and professional consequences [219]. This suggests that workers may cautiously welcome automated wellbeing interventions provided they protect privacy and are not shared with employers or third parties.

Healthcare context. In contrast, healthcare providers using emotion inferences for supportive interventions had a significantly negative and substantially larger impact on patient comfort across all three purposes. Developing automated mental health therapy had the largest negative effect of any purpose tested (representative: $\beta = -7.91$, $SE = 0.98$, $p < 0.001$; minoritized: $\beta = -7.88$, $SE = 0.95$, $p < 0.001$). Delivering acute wellbeing support, such as wellbeing tips, also had a significant negative effect (representative: $\beta = -3.49$, $SE = 0.98$, $p < 0.001$; minoritized: $\beta = -3.18$, $SE = 0.95$, $p < 0.001$). Automatically alerting a healthcare provider when support was needed had a significant negative effect as well (representative: $\beta = -3.89$, $SE = 0.98$, $p < 0.001$; minoritized: $\beta = -2.09$, $SE = 0.95$, $p < 0.05$).

Our qualitative analysis suggests several factors contributing to this discomfort. Participants expressed concern that automated wellbeing interventions could harm patients' mental health through inaccurate inferences or inadequate responses, reduce human interaction between patients and providers, lower the quality of mental healthcare, and breach confidentiality—particularly

troubling in a healthcare context characterized by strong expectations for privacy [733].

6.4.2 Identity-Based Effects

We examined the effect of socio-demographic factors on participants' comfort with emotion inferences in employment and healthcare, specifically race/ethnicity, gender, mental health status, and educational attainment as described in Section 6.3.0.3 and justified in Section 6.2.2.

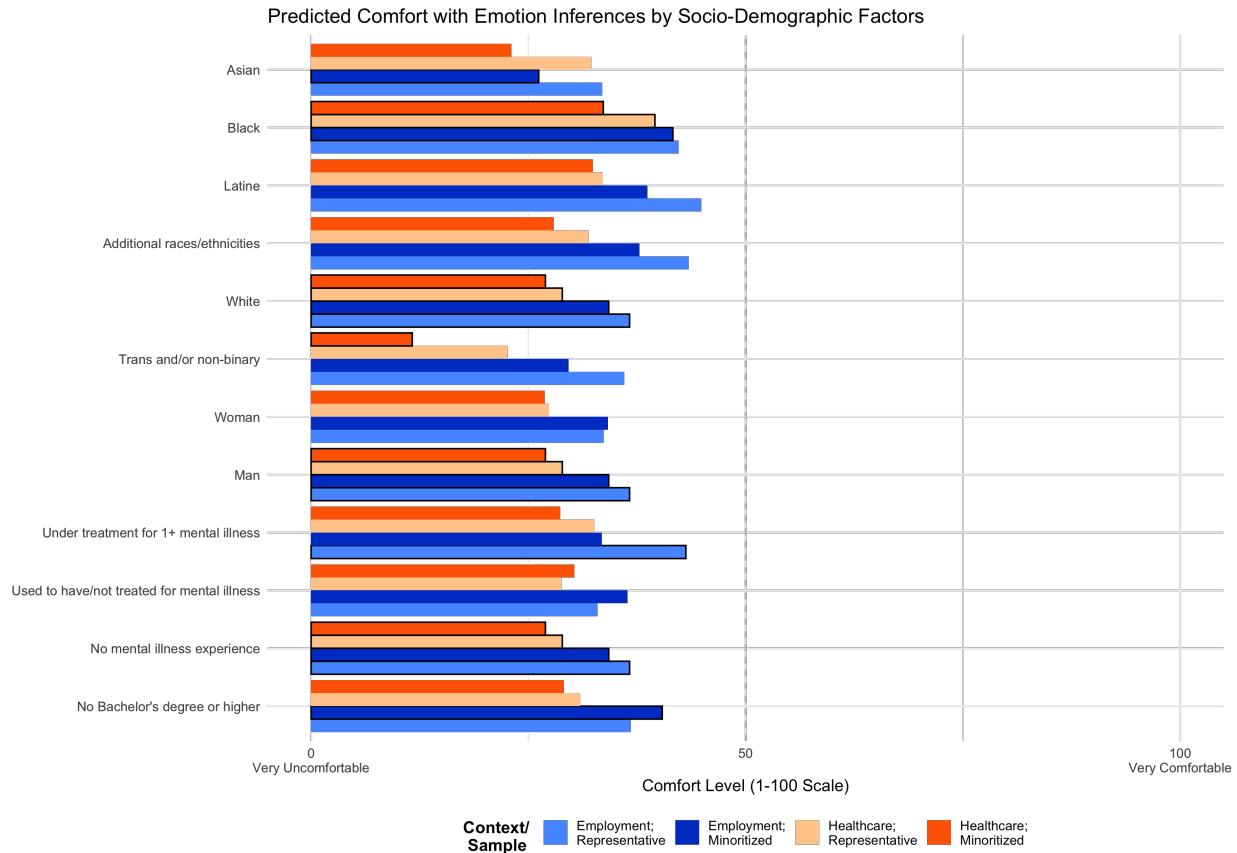


Figure 6.4: Predicted Comfort Levels by Socio-demographics. This figure illustrates the predicted comfort levels by combining the socio-demographic variable coefficients to each mixed-effects regression model intercept, derived by analyzing respondent comfort on a scale from 1 (very uncomfortable) to 100 (very comfortable). Bars with black borders indicate statistically significant results.

6.4.2.1 Race/Ethnicity

Compared to white participants, Black participants reported higher comfort with emotion inferences in both employment and healthcare contexts. In employment, this effect was statistically significant

in the minoritized sample (representative: $\beta = 5.64$, $SE = 3.19$, $p < 0.1$; minoritized: $\beta = 7.38$, $SE = 2.79$, $p < 0.01$). In healthcare, higher comfort was significant in both samples (representative: $\beta = 3.33$, $SE = 5.44$, $p < 0.01$; minoritized: $\beta = 6.66$, $SE = 3.05$, $p < 0.05$).

Asian participants reported lower comfort with employers inferring their emotions compared to white participants, particularly in the minoritized sample where the effect was statistically significant (representative: $\beta = -3.15$, $SE = 4.31$, insignificant; minoritized: $\beta = -8.05$, $SE = 3.55$, $p < 0.05$). We did not observe significant race/ethnicity effects for Latine or other categories, and no significant effects for Asian participants in the healthcare context.

Across all race/ethnicity categories, Black participants reported the highest comfort levels in both contexts, while Asian participants reported the lowest comfort in employment and white participants reported the lowest comfort in healthcare.

While higher comfort among Black participants may appear surprising given the documented racial and cultural biases present in emotion recognition datasets—leading to potential harms through inaccuracy or discriminatory use [719, 905, 416]—this group may attribute greater potential benefits to emotion inferences or perceive lower risk. Our qualitative studies [219, 733] provide some support for this interpretation. Black participants often highlighted the potential for emotion AI to mitigate racial discrimination and improve emotional support in both employment and healthcare. Yet, they also expressed concern about the risk of perpetuating existing inequities. Importantly, our qualitative analyses did not explicitly investigate the influence of race/ethnicity on participants' perceived risks and benefits. Future work is needed to understand how Black workers' and patients' nuanced perspectives on emotion AI shape their emotional privacy judgments.

6.4.2.2 Gender

In employment scenarios, we did not observe a statistically significant influence for any gender category on participants' comfort in either sample. While prior work suggests that privacy perceptions are often gendered, including in workplace contexts [804], larger sample sizes may be needed to confirm whether gender meaningfully influences emotional privacy judgments concerning emotion inferences in employment and healthcare.

In the healthcare context, however, trans and/or non-binary participants reported significantly less comfort than men on average, with this trend confirmed in the minoritized sample, which included a larger number of trans and/or non-binary participants (representative: $\beta = -6.26$, $SE = 10.99$, insignificant; minoritized: $\beta = -15.32$, $SE = 4.55$, $p < 0.001$). No statistically significant differences were observed for women compared to men in either sample.

Notably, the discomfort reported by trans and/or non-binary participants regarding healthcare providers' use of emotion inferences represents the largest negative effect observed for any socio-demographic factor in our analysis, underscoring substantial emotional privacy concerns about

healthcare applications of emotion AI in this group.

6.4.2.3 Mental Health Status

In the employment context, participants currently under treatment for one or more mental illnesses reported significantly higher comfort with emotion inferences compared to participants with no mental illness, but only in the U.S. representative sample (representative: $\beta = 6.47$, $SE = 3.13$, $p < 0.01$; minoritized: $\beta = -0.79$, $SE = 3.06$, insignificant). While minoritized participants currently under treatment reported lower comfort on average, the result was not statistically significant. As the coefficient range in Table A.2 shows, the direction of this variable's impact remains inconclusive in the minoritized sample.

We did not observe statistically significant differences in comfort for participants with resolved or untreated mental illness in either sample.

In the healthcare context, no statistically significant effects were found for any level of mental health status.

The significantly higher comfort observed among participants currently receiving mental health treatment in the U.S. representative sample—but not in the minoritized sample, which included a comparatively higher proportion of such participants—suggests a complex relationship between mental health status and emotional privacy judgments that may vary by intersectional identities. Since our sampling did not differentiate based on specific mental health diagnoses, we recommend future research explore perceptions of emotion inferences among people with particular mental illnesses to better understand and address these perspectives.

6.4.2.4 Educational Attainment

Compared to participants with a Bachelor's degree or higher, those without a Bachelor's degree reported, on average, higher levels of comfort with emotion inferences across both contexts and samples.

For employer use of emotion inferences, participants without a Bachelor's degree reported higher comfort in both samples. This relationship reached statistical significance only in the minoritized sample, where the positive effect size was substantially larger—a difference likely influenced by the minoritized sample's greater representation of participants without a Bachelor's degree (representative: $\beta = 0.14$, $SE = 2.46$, insignificant; minoritized: $\beta = 6.16$, $SE = 2.37$, $p < 0.01$). In the healthcare context, the relationship between lower educational attainment and comfort with emotion inferences was positive but statistically insignificant in both samples.

The consistently higher comfort reported by participants with lower educational attainment, especially in the employment context, suggests that this group may be less likely to recognize

potential risks associated with emotion inferences and/or may perceive greater potential benefits. More research is needed to better understand how educational attainment shapes emotional privacy judgments and risk-benefit perceptions related to emotion AI.

6.4.3 The Role of Privacy Beliefs, Trust, and Data Sensitivity

We investigated whether and how individual privacy beliefs—including general privacy concerns, trust in employers’ and healthcare providers’ handling of sensitive information, and perceived sensitivity of emotional information—affected participants’ comfort with emotion inferences.

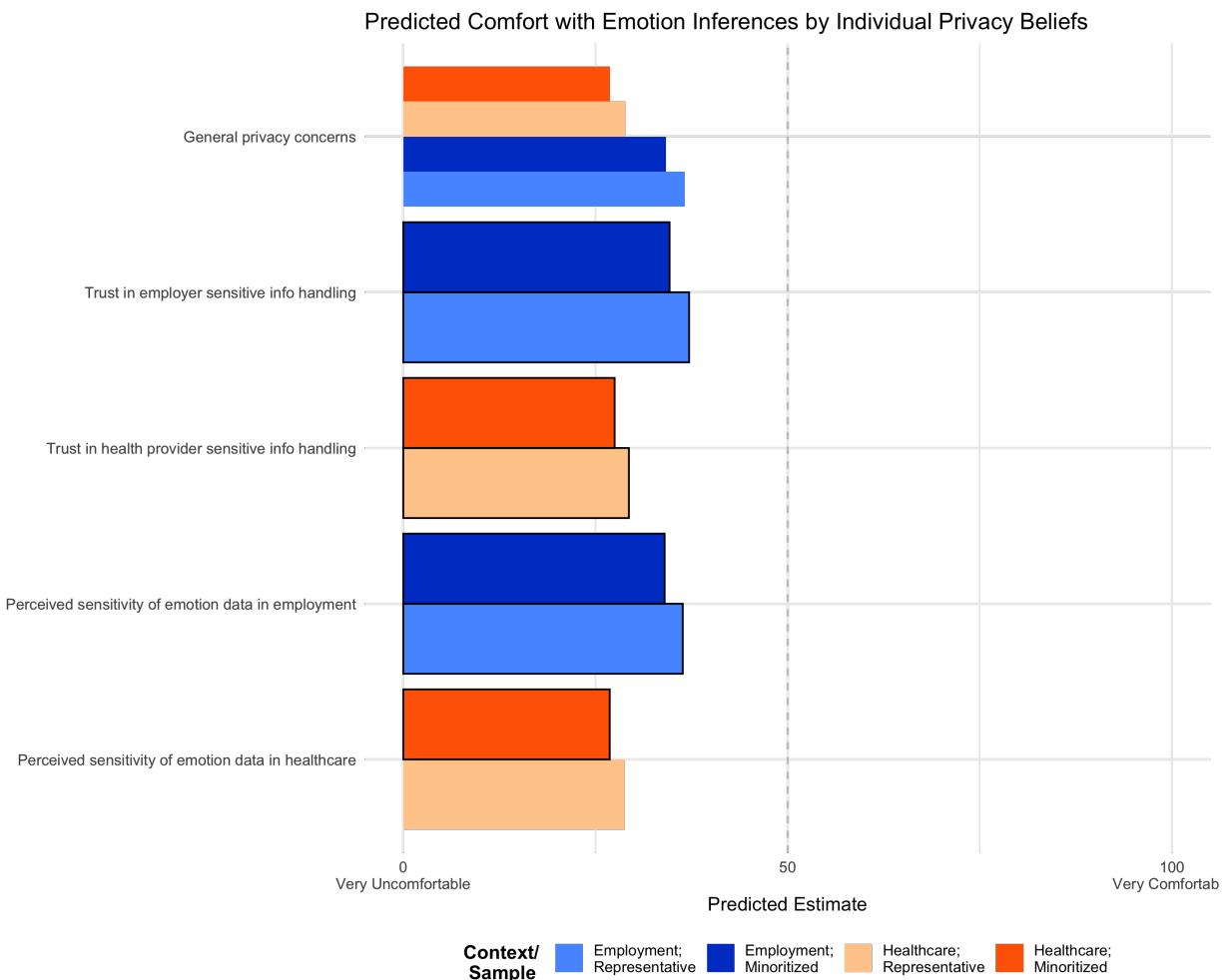


Figure 6.5: Predicted Comfort Levels by Individual Privacy Beliefs. This figure illustrates the predicted comfort levels by combining the individual privacy belief variable coefficients with each mixed-effects regression model intercept, derived by analyzing respondent comfort on a scale from 1 (very uncomfortable) to 100 (very comfortable). Bars with black borders indicate statistically significant results.

6.4.3.1 General Privacy Concerns

Participants' level of general privacy concerns did not have a statistically significant effect on their comfort with emotion inferences in either context or sample.

6.4.3.2 Context-relative Trust in Sensitive Information Handling

The level of trust participants attributed to their employers' and healthcare providers' handling of their sensitive information significantly and positively influenced their comfort with emotion inferences in both contexts. Participants reporting higher levels of trust reported significantly higher comfort with emotion inferences in both employment (representative: $\beta = 0.54$, $SE = 0.05$, $p < 0.001$; minoritized: $\beta = 0.40$, $SE = 0.05$, $p < 0.001$) and healthcare (representative: $\beta = 0.44$, $SE = 0.08$, $p < 0.001$; minoritized: $\beta = 0.53$, $SE = 0.05$, $p < 0.001$) contexts.

Of note, this effect was significantly different between samples for the healthcare context; the Z-score of 2.09 indicates that positive trust beliefs had a greater influence on patient comfort in the U.S. representative sample than in the minoritized sample.

6.4.3.3 Context-relative Perceptions of Emotion Data Sensitivity

Participants rated the level of sensitivity they associated with emotional information along with other information types already categorized in law and literature as sensitive – political opinions, religious beliefs, biometric data, health information, sex life/sexual orientation, genetic information, and union membership [1, 198] – when handled by one's employer and healthcare provider. As participants answered this question in a post-test after responding to vignettes that described various uses of their emotion inferences, we expect that responses are indicative of participants' perceptions of emotion inferences.

Employment context. As the box plot in Figure 6.6 illustrates, participants rated the sensitivity of emotional information handled by one's employer similar to data types already recognized as sensitive. The median level of perceived sensitivity of emotional information handled by employers for participants in the representative sample ranks higher than that for genetic information, health information, and union membership. The median sensitivity rating for emotional information handled by employers in the minoritized sample ranked among the lowest of sensitive data types, with a similar sensitivity to political opinions.

Healthcare context. Participants rated the sensitivity of emotional information handled by healthcare providers higher than when handled by employers, as shown in Figure 6.7. Participants in the representative sample perceived the sensitivity of emotional information handled by healthcare providers higher than biometric data, health information, political opinions, religious beliefs, and union membership. In contrast to their relatively lower perceived sensitivity of emotion data

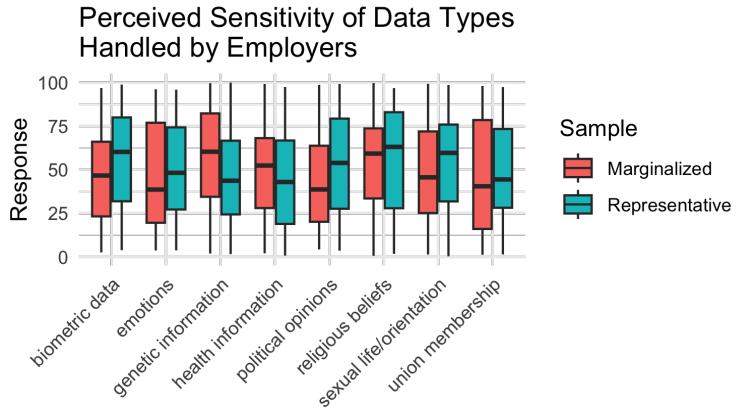


Figure 6.6: Perceived sensitivity ratings of emotional information compared to other sensitive data types in the employment context. Box plots show the distribution of sensitivity ratings for each data type by sample, on a scale from 1 (not sensitive) to 100 (extremely sensitive).

information handled by employers compared to other data types, participants in the minoritized sample rated emotion data information's sensitivity higher when handled by healthcare providers than all other sensitive information types.

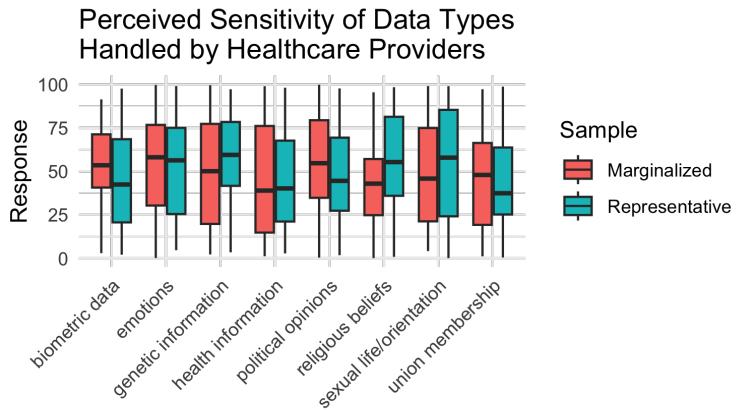


Figure 6.7: Perceived sensitivity ratings of emotional information compared to other sensitive data types in the healthcare context. Box plots show the distribution of sensitivity ratings for each data type by sample, on a scale from 1 (not sensitive) to 100 (extremely sensitive).

In addition, our analysis examined whether and how participants' perceived sensitivity of emotional information when handled by employers and healthcare providers affected their comfort with emotion inferences. We found that participants' perceived sensitivity of emotion data had a significant effect on their comfort with emotion inferences in both contexts. Participants associating emotional information with higher sensitivity reported significantly less comfort with

emotion inferences in the employment context (representative: $\beta = -0.30$, $SE = 0.05$, $p < 0.001$; minoritized: ($\beta = -0.25$, $SE = 0.05$, $p < 0.001$). Participants similarly reported less comfort with emotion inferences in healthcare (representative: $\beta = -0.10$, $SE = 0.05$, $p < 0.1$; minoritized: ($\beta = -0.11$, $SE = 0.04$, $p < 0.01$) contexts, with a significant effect confirmed in the minoritized sample.

6.5 Relational and Contextual Impacts: Perceived Benefits and Risks

A relational ethics lens centers the needs of those most impacted by a technology to critically examine its implications, challenging ideas of using AI to solve complex, social problems that often perpetuate harmful patterns of injustice and discrimination [121]. This approach is aligned with a growing body of scholarship [731, 122, 376, 53, 220] advocating to center the voices of those subjected to technologies in determining their impact, design, use, and regulation. Investigating data subjects' perceptions of emotion AI's impacts can uncover whether its applications are indeed beneficial for its contextually-situated data subjects, and expose implications that can inform future policy, regulation, and decision-making.

This section takes a relational ethics approach to examine the perceived impacts data subjects associate with the integration of emotion AI into two contexts: the workplace and healthcare. After answering survey vignettes self-rating comfort to various potentially beneficial applications of emotion AI in each respective context, participants answered open-ended questions about the benefits and risks they anticipate from using these systems. While the survey vignettes primed participants to anticipate positive impacts, they received no priming about potential risks, harms, or other unintended consequences.

In presenting qualitative results from this study, this section emphasizes the necessity of examining the relational and distributive implications of emotion AI, particularly how it interacts with power dynamics within workplace and healthcare contexts. By centering the voices of workers and patients, this analysis reveals that the deployment of emotion AI may inadvertently reinforce existing labor and healthcare inequities rather than resolve them. The findings underscore the need for a normative framework that prioritizes individual vulnerabilities and emotional privacy, setting the stage for a more comprehensive understanding of how emotion AI impacts the capabilities necessary for a just and equitable society.

6.5.1 Emotion AI in the Workplace

After answering survey vignettes about their comfort with employers using emotion AI for various purposes, participants answered open-ended questions about the benefits and risks they anticipate with these applications.²

While some participants acknowledged potential benefits, such as enhanced support for well-being and bias reduction, the overwhelming sentiment was one of caution and concern. Many workers anticipated that emotion AI would exacerbate existing challenges in their work environments, highlighting fears of heightened surveillance, loss of autonomy, and the amplification of biases that disproportionately affect minoritized communities.

Notably, when asked what benefits to them, if any, they associated with emotion AI in the workplace, 32% of participants responded that they did not foresee *any* benefits – 71.7% of whom identified with a minoritized identity, which we define for purposes of this paper as data subjects that identified with at least one of the minoritized groups for which we oversampled: (1) person of color; (2) gender minority; (3) current or past lived experience with mental illness. For those participants that did acknowledge how emotion AI could potentially benefit them in the workplace, we find that emotion AI's potential benefits are overshadowed by a myriad of ethical and justice-related concerns that expose data subjects to potential harms. Using a relational ethics lens [121], we reveal how data subjects acknowledge potential benefits of emotion AI in the workplace yet fear employers' use of emotion AI may enact unjust and disproportionate emotion AI-inflicted harms to data subjects. We describe three ways data subjects perceive employers' use of emotion AI could potentially benefit or harm their 1) wellbeing; 2) work environment, performance, and employment status; and 3) employers' (im)partiality. Moreover, we found that as a result of these potential risks, participants foresee conforming to or refusing emotion AI in the workplace. Throughout the following sections, we include (1) the overall percentages of all participants who shared each findings' theme and (2) what percentage is represented by participants who identified as having at least one minoritized identity to further amplify the voices and perspectives of minoritized participants.

In this section, we first describe findings from the large number of participants that did not perceive *any* benefit to them associated with emotion AI in the workplace. We then share findings organized around areas where data subjects anticipate potential benefits and risks of emotion AI in the workplace. We conclude by describing data subjects' potential practical responses to emotion AI in the workplace.

²This section is based on: Shanley Corvite*, Kat Roemmich*, Tillie Rosenberg, and Nazanin Andalibi. 2023. Data Subjects' Perspectives on Emotion Artificial Intelligence Use in the Workplace: A Relational Ethics Lens. Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 124 (April 2023), 38 pages. <https://doi.org/10.1145/3579600>. *Co-first authors contributed equally. This material is based upon work supported by the National Science Foundation under Grant No. 2020872.

6.5.1.1 No Perceived Benefit

Although participants acknowledged a variety of potential benefits of using emotion AI in the workplace (described in later sections), many participants reported that there would be no benefit to them at all. Even after responding to vignettes that suggested potentially beneficial workplace applications of emotion AI, about 32% of participants, 71.7% of whom were participants who identified with a minoritized identity (i.e., person of color, woman, transgender, non-binary, having or had a mental illness) did not note any benefit when asked to describe potential benefits they might receive from emotion AI in the workplace. Some participants pointedly responded with answers such as “*None*” and “*No benefit*,” while others shared that they viewed no benefit to emotion AI use in the workplace and, instead, raised concerns regarding how it can pose risks onto them. Participants specified a variety of concerns (outlined in later sections) including the potential for emotion AI use to harm their wellbeing, work environment, and employment status, and to create and amplify bias and stigma against them, especially for those with minoritized identities. Furthermore, participants expressed distrust against emotion AI use and anticipate conforming to or refusing its implementation in practice. For example, P103 said, “*I don't [see a benefit]. computers and such have not advanced enough to take the place of people in these things and every persons expressions and things are more subtle.*” Additionally, P86, who has a mental health condition, stated, “*i don't think anything could be beneficial from the intrusion of employers into employees' personal health.*” Participants such as P103 and P86 did not perceive any value to being subjected to emotion AI in the workplace, and their remarks point to their skepticism of the technological capabilities and employers’ use of emotion AI. We also note that, although 32% of participants explicitly noted no benefit to emotion AI use at all, even participants who acknowledged a potential benefit went on to describe a myriad of potential risks that they may become exposed to.

6.5.1.2 Contextual Impacts: Work Environment, Performance, and Status

Many participants commented on how employers’ use of emotion AI-generated inferences and associated data could improve or harm data subjects’ work environment, performance, and employment status. While participants noted the potential for employers to use emotion AI in ways that could improve the work environment such as through maintaining a safe workplace and managing their workloads, participants contrasted these potential benefits with potential risks of employers using emotion AI in harmful ways (e.g., the potential for employers’ use of emotion AI to impair their work performance, potential negative employment outcomes such as denying benefits or even termination).

Work environment. Emotion AI is touted for the possibility to enhance the workplace environment and employees' performance through purposes such as insider threat detection [683] and managing employees' work [362]. Participants acknowledged similar potential benefits, noting how employers could use emotion AI to improve the work environment by helping to maintain a safe work environment and improve data subjects' performance through performance management. Such conceptualizations are in conversation with survey vignettes positing emotion AI as a means to infer employees' risk of self-harm and harm towards others, assess employees' work performance, and alert employers when they need more support.

Workplace safety. 10.9% of participants remarked upon how emotion AI's inferences could be used to identify and protect people who may pose harm to themselves or others, 69.8% of whom were participants with a minoritized identity. P164 who was a woman looking for work stated, "*Personally, I think the most beneficial uses of these programs would be to detect potentially harmful/violent behavior,*" suggesting that emotion AI could help with sustaining a safe work environment. Some participants noted how emotion AI could be used to detect potential danger for both self-harm and harm posed on others. P77, a full-time employee, stated that emotion AI could potentially "*infer if employees could harm themselves and others*" and P391 who is looking for work stated, "*I do think in terms of safety it would be beneficial. If it's able to see if a coworkers has the potential to hurt themselves or another person I would want to know. It would prevent injury or in some cases death,*" pointing to a potential benefit of emotion AI to recognize potential harm and maintain safety which may, in turn, improve the workplace.

Performance management. 15.4% of participants also acknowledged the potential for employers to use emotion AI to improve the workplace for purposes such as assessing data subjects' performance and alerting employers when data subjects may be experiencing work overload, 75.4% of whom identified as having a minoritized identity. Some interpreted emotion AI as a form of performance management, suggesting how such purposes could improve data subjects' work conditions and outcomes. For example, P148 an African-American or Black man who was a full-time employee, stated "*Potentially, this system could help me when I am overrun with work and burnt out. The system could help alert my employers that I need something to improve my emotional well-being,*" suggesting that using emotion AI as part of work performance assessments could prompt employers to reduce or adjust data subjects' workloads without negative repercussions for the data subject, as implied by P148. To add, P245 stated that she has "*issues when it comes to working that could use some legitimate support. This could help [her] employer understand or at least be obligated to comply with offering that support or lenience,*" demonstrating how some participants recognize the potential for emotion AI use in the workplace to result in an increase in employers' support for data subjects.

Some participants noted that emotion AI as a performance management tool could be mutually beneficial for both data subjects and employers. As P120 described, using emotion AI to promote “*greater productivity in the work place*” could “*be a win-win.*” A “*win-win*” in this context demonstrates how, for some participants, using emotion AI to identify gaps in support that data subjects may need to manage their workload and better perform, may benefit both data subjects and employers, potentially leading to improved work outcomes.

Impaired work performance. Participants raised concerns regarding the potential for emotion AI use to impair their work performance.

8.1% of participants noted how employers’ use of emotion AI would impair their work performance, 80% of whom were participants who identified with a minoritized identity. P339 who is employed full-time, noted how “*Monitoring employee’s attention or work with close scrutiny has been proven to lower productivity,*” contesting previously mentioned potential benefits of emotion AI use to monitor work to manage and improve data subjects’ workloads and productivity. Some participants also mentioned how employers using emotion AI could potentially wrongly treat data subjects, deteriorating the workplace environment for data subjects. P42, who was a part-time employee, noted that employers’ use of emotion AI could result in “*Lower self-esteem and bring embarrassment from being called out by an employer,*” pointing to how employers’ emotion AI use could potentially harm how she performs and feels in her workplace, in contrast to promises of using emotion AI to promote better work conditions and outcomes.

6.5.1.3 Relational Impacts: Worker Dignity, Agency

Worker wellbeing. Participants’ remarks pointed to how, as data subjects, they anticipated their wellbeing would be directly impacted by emotion AI use in the workplace. Perceptions of whether data subjects’ wellbeing would be positively or adversely impacted by emotion AI were dependent upon the employers’ use of the personal and sensitive emotional data the emotion AI application generated. Participants acknowledged the potential for employers to practice care when interacting with data subjects’ emotion data (e.g., improving their wellbeing by identifying individuals in need of support and taking supportive action when support is needed); however, participants contrasted these potential benefits with concerns that employers would use emotion AI in ways that could potentially inflict harm to their wellbeing (e.g., inaccurate inferences, loss of privacy).

Enhance worker wellbeing. Producers and adopters of emotion AI claim to use data regarding peoples’ emotional states to promote workplace and data subjects’ overall wellbeing [670, 626, 616]. Resembling factorial vignettes that posited the purpose of emotion AI in the workplace to diagnose health conditions early, identify individuals in need of support, and provide data-driven understanding about employees, participants acknowledged emotion AI’s potential to improve their wellbeing (e.g., early health detection, diagnosis, and intervention; increased support and awareness of data

subjects' wellbeing) with the qualification *if* employers used emotion AI-generated inferences in ways that demonstrate care towards data subjects' wellbeing.

Early health detection, diagnosis, and intervention. 15.9% of participants, 79.4% of whom were participants who identified with a minoritized identity, noted emotion AI's potential to benefit them if emotion AI could diagnose a health condition early through health monitoring and detection. For instance, P314, who reported having a mental health condition being treated with medication, stated, “*... there are always undertones but at the same time it can detect if you really need help. Not only that but to also have your employer care about your mental health? That seems very beneficial because taking care of mental health will most likely help with performance of that individual.*” P314 acknowledges how emotion AI in the workplace could potentially benefit her if employers used the technology in ways that provided care for her mental health, such as by detecting health conditions early (an outcome with positive consequences for one's work performance as well). P27 who was a multi-racial man looking for work, also stated, “*Anything that could diagnose cognitive decline earlier would be helpful not only to me, but to anyone with a brain. It could lead to much more positive health outcomes and a higher quality of life for those nearing retirement age,*” suggesting how participants anticipate the potential benefit of emotion AI aiding in the earlier diagnosis of health conditions to promote better health outcomes that may not otherwise be possible for them.

Some participants also described how early health detection and diagnosis can serve as a means to intervene in one's health conditions discovered and diagnosed by emotion AI. P57, who reported having a mental health condition, described emotion AI's potential to detect health conditions early as beneficial by providing necessary information they could use to intervene in their own wellbeing: “*Used responsibly, these could be a serious boon to mental health. If a system sees I'm having say... an OCD relapse before I do, I could take action earlier to stop it.*” These examples highlight how participants acknowledge emotion AI use in the workplace to be potentially beneficial to them in detecting, diagnosing, and intervening in their potential health conditions which can potentially lead to an improved overall wellbeing in the workplace.

Increased support and awareness. 30.6% of participants acknowledged emotion AI's potential to provide the necessary information to employers to identify health conditions and provide support to manage said conditions, 82.6% of whom were participants who identified with a minoritized identity. P3, a non-binary person with a mental health condition, stated that emotion AI could be beneficial to them, “*if employers are fair and understanding of their worker's mental health and the importance of providing accommodations when needed...,*” illustrating how emotion AI may be potentially useful to data subjects if employers were to demonstrate care towards data subjects by providing resources to those in need.

P230, who disclosed having a mental health condition being treated with medication, noted how

emotion AI could lead to potential wellbeing benefits by increasing one's self-awareness about their health, stating "*Yes, i feel like these systems would 100% benefit me, not only as a employee but also as a person.. For example, i could find more ways to cope with my bipolar depression, while at home and work.*" In addition to promoting self-management through increased self-awareness, P230 further described how emotion AI health detection and monitoring could potentially lead to improved understanding about mental health and disabilities in the workplace as a whole, sharing "*I feel like its important for the work place to understand mental health and also disabilities, and I feel like this is amazing and so beneficial.*" Here, P230 highlights the lack of understanding surrounding mental health and disabilities in the workplace, and the allure of emotion AI's potential to indirectly benefit data subjects by promoting data-driven understandings about mental health between data subjects and employers in the workplace.

These findings demonstrate that data subjects view emotion AI as a tool to potentially promote employers' understanding and awareness of data subjects' wellbeing while allowing for data subjects' self-awareness of possible health conditions and potentially providing data subjects with necessary resources and care that can improve their wellbeing. However, the potential for emotion AI to benefit data subjects is dependent upon employers' use of emotion AI-generated data that demonstrates care towards data subjects in the workplace. As the following section describes, participants described concerns of how employers' use of emotion AI data beyond the stated beneficial purposes could inflict harm to data subjects' wellbeing.

Harm worker wellbeing. Despite the aforementioned potential benefits, our analysis reveals how data subjects contrasted their acknowledgement of emotion AI's potential to benefit their wellbeing with concerns that employers could use emotion AI-generated data in ways that would not demonstrate care toward data subjects, which could inflict harm to their wellbeing. Some noted the potential for negative wellbeing effects as a result of inaccurate inferences and subsequent misdiagnosed health condition(s), while others cited concerns that emotion AI would increase stress due to privacy loss. Therefore, many participants perceived the personal information emotion AI generates as irrelevant to their employers and that emotion AI in the workplace is contextually inappropriate.

Incorrect inferences of health conditions and misdiagnosis. 5% of participants noted that emotion AI use could lead to incorrect inferences of a data subjects' health condition, of whom 76.2% identified with a minoritized identity. These participants contested previous ideas of emotion AI use to aid in improving data subjects' wellbeing. As noted by P36 who reported having a mental health condition, "*I have a problem with the possibility of incorrectly assessing individuals. It would be a hell of a thing to get 5150³ into a psych ward just because a computer thought you needed*

³The number of the section of the Welfare and Institutions Code that allows a person with a mental illness to be involuntarily detained for up to 72 hours.

it," describing concerns regarding potential harms, including to one's wellbeing and autonomy, resulting from emotion AI's inaccurate inferences. Similarly, P84 said that emotion AI systems "could easily misdiagnose my condition or make it seem as if I had a poor work experience even if that is not the case," illustrating concerns around how emotion AI could lead to both a misdiagnosis and negative assumptions towards an individual's work.

Privacy loss. 50.9% of participants, 36.7% of whom identified as having a minoritized identity, noted their concerns surrounding their privacy and how emotion AI use by employers could contribute to their worsened wellbeing. P96, who has a mental health condition that has not been formally diagnosed, expressed how emotion AI systems could "*gather sensitive information the employee wishes to be kept private, and they [emotion AI] just generally overstep boundaries,*" suggesting that participants consider emotion data to be personal and sensitive information and that emotion AI's deployment in the workplace would violate privacy boundaries that data subjects hold for their emotions in the work context. By rendering visible personal and sensitive information that one wishes to remain private, emotion AI may psychologically harm data subjects by inducing emotional disturbance or distress [194]. For example, P363, who disclosed having multiple health conditions, shared concerns regarding negative wellbeing implications as a result of emotion AI-induced privacy intrusions: "*The awareness that I am being analyzed would ironically have a negative effect on my mental health.*" P363's concerns demonstrate that, despite emotion AI's claimed goals to infer and improve data subjects' wellbeing in the workplace [670, 626, 494, 616] as also observed in Section 4.2.1, emotion AI use can "*ironically*" lead to opposite effects in which data subjects' wellbeing may suffer as a result of losing control and privacy over their emotional states.

As many participants noted privacy concerns regarding emotion AI use, 20% therefore viewed emotion AI use as contextually irrelevant to employers and the workplace in general, 81% of whom identified with having a minoritized identity. P62, who has lived with anxiety and depression, noted, "*it makes me feel that it might be strange to have a system at work monitoring my mental health as these things may have nothing to do with my work, or what or how much I am accomplishing during my time at work.*" P62's remarks point to how monitoring data subjects' wellbeing using emotion AI could generate unrelated inferences regarding their performance at work. P56, a white woman who was a full-time employee, also stated, "*It's an invasion of privacy. It's an employee's responsibility to seek out help not an employer's responsibility to pry into someone's personal life,*" demonstrating the belief that emotion AI is both invasive to data subjects' privacy and inappropriate for employers to attempt to improve data subjects' wellbeing. Participants noted how emotion AI-generated inferences are unsuitable for the workplace as it could lead to potential negative impressions of data subjects and feelings of uneasiness at and towards work, including for those living with mental health conditions. Overall, participants' remarks suggest that emotion

AI's introduction in the workplace violates contextual integrity (i.e., by not adhering to contextually relevant and appropriate norms of information collection and sharing [646]).

Altogether, we find that participants acknowledged that employers' use of emotion AI could potentially improve or deteriorate data subjects' wellbeing, where either outcomes is dependent on how much employers care for, in practice, data subjects' wellbeing in the data subject-employer dynamic.

Negative employment outcomes. 19.7% of participants expressed concerns regarding the potential for employers' use of emotion AI to create or intensify existing power imbalances between data subjects and their employers, which could, in turn, harm data subjects' employment status, 76.9% of whom identified with a minoritized identity. P10, who was a part-time employee, noted how emotion AI systems could potentially "*give a little too much power or authority to the employers,*" pointing to how she finds employers' use of emotion AI concerning. P115, who noted that she was looking for work, describes how emotion AI could give "*employers more access to personal, private data on their employees*" which could result in employers having more power over data subjects by using their "*personal, private data.*"

Participants that mentioned the potential for exacerbated power imbalances were fearful of the dynamic they would have with employers if emotion AI were integrated into their workplace, pointing to how emotion AI use could potentially intensify already existing tensions in the employer-data subject relationship. For instance, P44, a full-time employee, describes how he viewed the current employer-data subject relationship in the workplace: "*The amount of control that employers already have over employees suggest there would be few checks on how this information would be used. Any 'consent' on employees is largely illusory in this context.*" P44's remarks on the dynamics in the workplace highlight the power employers currently have which emotion AI could potentially intensify. Additionally, P44's remarks point to the potential for creating a false sense of security whereby data subjects giving consent to the use of emotion AI may be useless and "*illusory*" as employers could misuse emotion AI regardless with "*few checks*" in place.

These potential risks regarding power imbalances between data subjects and employers led 33.4% of participants, 77.3% of whom were participants who identified with a minoritized identity, to express concerns regarding employers potentially using emotion AI and their authority – reinforced by relying on emotion AI-generated inferences – as a means to make unjust employment decisions such as firing or denying benefits and promotions with implications for equal opportunity and safety. P245, who has a felony and has "*been labeled a felon [since she] got out in 2008,*" expressed concerns about what this might affect her if the system would "*block people by accidentally saying they're dangerous,*" and thereby equating having a felony with being a risk to others in the workplace, a phenomenon observed in prior work [415]. P245's remarks acutely demonstrate the existing biases she faces as someone with a felony, suggesting how using emotion AI in workplace

safety initiatives may indeed not increase workplace safety and equal participation opportunities (or perceptions thereof) for some data subjects who already face unjust barriers in the workplace (e.g., those with a felony in their backgrounds). Whether a particular emotion AI system takes into account a data subject's past history or not, and whether it labels a former felon as dangerous is not our focus here, but what the concerns of a data subject in that position are. To add, P15, a full-time employee with a diagnosed mental health condition, mentioned that "*They [employers] could decide that I am no longer a good fit at work and fire me. Decide I'm not capable enough and not give me a raise, or think I'm not working enough,*" highlighting beliefs about several potential ways in which employers' use of emotion AI could impact data subjects' employment status. P50, also a full-time employee who has multiple health conditions, stated that "*This technology could very easily be used by employers to fire employees struggling with mental health issues or to hold them back from receiving promotions or raises. Employers could possibly use it to force employees into unpaid FMLA leave if the employer determines that the employee has mental health issues.*" These examples are in stark contrast with potential benefits described earlier surrounding employers using emotion AI to better manage data subjects' workloads as well as to support and care for their wellbeing.

Taken together, these examples highlight participants' concerns around employers using emotion AI as a means to impair work performance and justify undesired employment decisions or negative and consequential perceptions of data subjects, posing harm to data subjects' employment status rather than improving their work environment or performance. Furthermore, these findings demonstrate that data subjects are wary of how employers would use emotion AI in practice to make high-stakes decisions, such as firing and denying them benefits.

Bias and stigma. Participants noted the potential for employers to make impartial or partial decisions and perceptions when using emotion AI-generated inferences. We refer to (im)partiality as the degree to which employers make fair, unbiased decisions or perceptions regarding data subjects when using emotion AI. Participants acknowledged the potential benefit of employers using emotion AI in ways that demonstrate impartiality towards them such as reducing bias and removing barriers to disclosure surrounding mental health conditions. However, they also noted potential harms such as making false inferences regarding data subjects and creating and perpetuating bias and stigma against data subjects. In describing these concerns, participants highlighted their fear of employers overrelying on inaccurate and biased emotion AI systems.

Reducing bias and stigma. Research demonstrates that in general, the public's views are aligned with the popular allure of AI's potential to make unbiased and objective decisions [97, 243, 464]. Similarly, participants noted that emotion AI use in the workplace could lead to more unbiased decisions and perceptions (compared to traditional workplaces) concerning data subjects through

reducing bias and stigma, in line with claimed purposes of implementing emotion AI and other algorithmic approaches to *avoid* human biases and subjectivity when assessing an employee's emotional state (i.e., through employer observations or employee self-reports), as suggested by one survey scenario.

Reducing bias against data subjects. 3% of participants acknowledged how employers could potentially use emotion AI to make fair and unbiased decisions and perceptions regarding them if the systems were trained and de-biased to account for differences in identities, 66.7% of whom identified with a minoritized identity. As P180 stated, emotion AI might be beneficial, "*assuming the software adjusts for those differences [across different identities] and still outputs correct and reliable data,*" suggesting that, the potential benefit of emotion AI for data subjects is contingent upon its technical accuracy and lack of bias, particularly for groups emotion AI is known to generate less accurate inferences for including women, disabled people, and people of color [654, 719, 271]. If the system were to reliably remove demographic biases, some participants saw its potential to mitigate the bias they experience in the workplace. As P228, a transgender East Asian woman who was looking for work, shared, "*It might benefit me if it avoids employers' human biases in making judgments so as to be more objective,*" demonstrating how the promise of objective and reliable use of emotion AI leads to a belief that emotion AI use could reduce subjectivity and bias in the workplace. Sharing how such promises of emotion AI to reduce human judgement and subjectivity might benefit data subjects, P194, a Black woman who reported having a mental health condition, stated that "*It may be better in getting past the common problems of discrimination in giving better readings.*" These conceptions of emotion AI show the allure towards emotion AI's promises of objective and de-biased inferences to benefit data subjects by replacing human subjectivity and bias in the workplace.

Reducing stigma around data subjects' mental health and associated disclosures. 5% of participants, 80% of whom identified with a minoritized identity, acknowledged the potential for emotion AI use to help reduce stigma surrounding mental health and its disclosure in the workplace. As disclosure about mental health can often be a stigmatized topic [743, 296] with consequences including in the workplace, participants acknowledged how emotion AI use could potentially facilitate discussions surrounding mental health in the workplace, remove barriers to disclosure of mental health (or other health conditions) to employers, and lead to increased understanding and support for data subjects. For example, P28, who lives with multiple mental health conditions, stated, "*I think they [emotion AI systems] would benefit me in being able to speak about my wellbeing not directly . . . ,*" demonstrating how emotion AI could potentially allow data subjects to disclose their emotional or physical states without having to directly talk about it by virtue of emotion AI making inferences about their health. To add, P311, a full-time employee with a mental health condition, noted the potential for emotion AI to aid in reducing stigma around mental health in the

workplace, leading to more access to support and resources for data subjects: “*They could allow access to care that normally has a negative stigma attached to it without having to put yourself [out there.]*” These examples highlight participants’ beliefs that emotion AI use in the workplace could potentially directly or indirectly facilitate disclosure of their mental health or wellbeing broadly which may potentially lead to employer support without risking stigmatization and bias.

Stigma, as a form of prejudice and of discriminatory nature, is associated with negative wellbeing effects [537, 538]; as such, while we include themes surrounding stigma here, we note that participants’ concerns surrounding emotion AI’s impact on increasing or reducing stigma is also relevant to emotion AI’s potential wellbeing impacts suggested by participants, which we discuss in Section 4.2.

Creating and perpetuating bias and stigma. Although participants acknowledged the potential for emotion AI to be used in ways that could help reduce bias and stigma in the workplace, in contrast, they also raised concerns about the potential for employers to use emotion AI in harmful ways such as making false inferences about data subjects based on inaccurate emotion AI inferences, leading to bias, discrimination, and stigmatization against data subjects.

Incorrect and inaccurate emotion AI-generated inferences. 36% of participants noted the potential for emotion AI to produce inaccurate and incorrect inferences about data subjects that employers would then accept at face value, leading to partial and incorrect perceptions of data subjects, 72.7% of whom identified with a minoritized identity. P87, who was being treated for a mental health condition, described the understanding that “*all current AI systems depend heavily on their training material. They are often wrong when presented with information outside their training. It seems to me very difficult to provide a suitable large training set to cover the full gamut of human emotions,*” noting concerns regarding emotion AI’s inability to infer all human emotions correctly. Additionally, P94, a Hispanic or Latino/a/x man employed full-time, stated that data subjects “*have just one more thing to worry about, the employer thinking the employee is suicidal or something because a system saw them make a weird facial expression,*” illustrating the belief that emotion AI could potentially produce inaccurate inferences regarding data subjects that employers could take as true, leading to potentially harmful consequences against them. To add, P328, a Hispanic or Latino/a/x woman employed part-time stated, “*Another concern I would have is it [emotion AI] being inaccurate. Or perhaps not doing enough help than they [employers] think it will. If it is so accurate in diagnosing and bringing attention problems, that might also have a negative side effect. If you were somewhat feeling okay that day but the system reads something different, it could bring attention to things that people were not thinking of in the first place and make them aware of their true emotions at an inconvenient setting.*” P328’s remarks point to concerns around inaccurate emotion AI-generated inferences about her, but also harm such as lack of control and agency around whether and when she directs attention to her internal states, even if

accuracy concerns are resolved.

Bias and discrimination against data subjects. 15.9% of participants expressed concerns regarding employers using emotion AI in ways that could lead to bias against data subjects with a minoritized identity, disproportionately affecting data subjects along dimensions such as race, gender, class, disability, and sexuality, 85.7% of whom were participants who identified with a minoritized identity. For instance, P95, a disabled Black woman, mentioned that she would be concerned about emotion AI use in the workplace “*If they’re [emotion AI systems] not programmed properly to consider race & culture.*” She went on to describe how her identity as a “*poor/black/elderly/woman*” would lead to obstacles in “*getting real, honest, caring help from professionals... [and she has] to take into consideration that the bots are being programmed by people which most times, (maybe unintentional), use their bias.*” P95’s comments point to concerns regarding emotion AI’s potential to perpetuate bias and discrimination against data subjects with minoritized identities and how said bias may impact the support and resources one would have access to. To add, P7, a transgender and non-binary white person, described how emotion AI systems could “*have the potential for both racial and gender biases, particularly against POC [people of color], women, and trans individuals. Who is deciding what expressions ‘look violent’ and how can one determine people as a threat just from the look on their face?...*” P7’s remarks demonstrate concerns around how employers using emotion AI could potentially discriminate minoritized data subjects. As a result, participants noted emotion AI to be a potential means to take unfair actions towards minoritized groups and to justify such partial actions. P42, a white woman who was employed part-time, mentioned that “*there is already a bias in the workplace for minorities and women, these systems could be used as ‘evidence’ in any unjustice, or oppression, by blaming it on mental instability.*” These examples point towards data subjects’ concerns regarding employers using emotion AI to further perpetuate discrimination against minoritized data subjects in the workplace.

Perpetuating mental health stigma. 2.3% participants, 88.9% of whom were participants who identified with a minoritized identity, noted how emotion AI could potentially lead to perpetuating stigma of mental health and discrimination in the workplace, contrary to earlier acknowledgements of emotion AI reducing stigma and barriers to disclosure of mental health-related concerns in the workplace. P358, who has a formally diagnosed mental health condition stated that “*Mental health is stigmatized enough without allowing employers access to a computer program that thinks it can figure out mental health,*” implying, with skepticism about emotion AI’s capabilities, that implementing emotion AI into the workplace could further stigmatize mental health.

Others described concerns surrounding being stigmatized for their accurately emotion AI-inferred mental health conditions. For example, P234, a Black woman with multiple mental health conditions, stated that, “*Unfortunately, there’s a nasty stigma around mental health and you can*

be subjected to employee discrimination even though it's against the law. I can't afford to take that chance. I believe mental health is best left in the hands of medical professionals instead of employers with possible agendas." P234's remarks demonstrate the fear of employers becoming aware of her mental health status due to emotion AI use, leading to stigmatization in the workplace and potentially causing her unjust harm and discrimination. These examples highlight data subjects' concerns regarding emotion AI in the workplace and the potential to perpetuate bias and stigma, contesting ideas of reducing bias and stigma as described in Section 4.3.1. Overall, participants who noted emotion AI's potential to perpetuate bias and stigma demonstrate the harms that could be posed to data subjects, especially specific groups of people such as those with minoritized and/or stigmatized identities.

All in all, to situate concerns regarding inaccuracy, stigma, and bias resulting from emotion AI use in the workplace, it is worth noting that 1.3% of participants also expressed concerns regarding the potential for employers to overly rely on emotion AI, 80% of whom identified with a minoritized identity. Participants were especially concerned with overreliance on systems that have no human input, expressed by 4.3% of participants, 67.4% of whom identified with a minoritized identity. For example, P103, a full-time employee, stated that emotion AI "*will be relied upon too much.*" Similarly, P318 mentioned that implementing emotion AI would mean trusting "*employers to do the right thing too much,*" suggesting that relying on employers to effectively use emotion AI in supportive ways is not trivial. Employers' use of emotion AI systems without human input led to exacerbated concerns regarding overreliance. For example, P283, a white woman who was a full-time employee living with a mental health condition, stated that because they "*feel like computers cannot do what a human is able to do,*" employers over relying on emotion AI-generated inferences is problematic.

Worker choice. Participants described how they would respond to emotion AI if the systems were implemented into their workplace with many envisioning themselves to change their feelings or behavior, citing how they would partake in conforming to emotion AI expectations or rejecting the technology. Whether or not participants would actually engage in these activities in practice is difficult to know, especially with emerging technologies like emotion AI; however, these findings do illustrate participants' concerns regarding being subjected to emotion AI in the workplace.

Conforming. 7.6% of participants, 76.7% of whom identified with a minoritized identity, suggested that they would intentionally change their typical behavior and feelings in the workplace to conform to expectations of data subjects' behavior as prescribed by emotion AI, in an attempt to avoid being adversely affected by emotion AI. This is significant as it highlights how emotion AI implementation in the workplace may lead to data subjects losing control and autonomy over their own actions and emotions. For example, P360, a trans full-time employee with a mental health

condition, stated, “*it would cause me to act differently than I normally do at work,*” highlighting the potential for emotion AI use to lead data subjects changing their behavior at work. P272, also a full-time employee, described that “*You could not be yourself and roll your eyes at your Supervisor or co-worker if you felt the urge, you would have a constant feeling that big brother is watching and you are not alone.*” P272’s association of emotion AI with “Big Brother,” a fictional character from George Orwell’s dystopian novel *1984* [672], highlights her fear of surveillance from higher authorities and the consequences from implementing emotion AI into the workplace. To add, P185, employed full-time, shared: “*It would affect me in that if I had to use the app for work, I would fake a smile or otherwise try to fool the software because I would not want my employer to know my mental state unless I wished them to,*” pointing to concerns regarding emotion AI use to violate data subjects’ privacy, therefore changing her behavior by engaging in emotional labor [232, 371] to not provide her employer with private information about her emotional states. Similarly, P71 who is disabled and has multiple health conditions, stated that they would “*exert a massive amount of energy masking (engaging in neurotypical/expected behaviors when they aren’t natural to me) even when alone in my office, which would make me very distracted and unproductive*” if emotion AI was implemented in the workplace. However, if they were unsuccessful, P71 continues, they could “*constantly be flagged by the software.*” These remarks demonstrate the concern that conforming to emotion AI in the workplace may amplify negative impacts, especially on disabled data subjects, by constantly partaking in additional emotional labor to evade emotion AI inferences, leaving them exhausted. These examples illustrate that data subjects may live with emotion AI in the workplace by acting in ways that would deter the system’s inferences about them or by otherwise conforming to normative workplace behavior expectations as encoded by emotion AI. However, participants feared how this reaction could lead to harmful impacts on their health and productivity, especially when disabled or otherwise minoritized in the workplace.

Refusing. 4.1% of participants who shared concerns about emotion AI-inflicted harms responded with feelings of distrust towards emotion AI technology, 62.5% of whom identified with a minoritized identity. Some stated broad feelings of distrust such as P26 who was employed full-time saying, “*I would not trust such a system.*” Others explained why they felt such distrust towards emotion AI such as P12 who was employed part-time stating, “*I do not trust a computer program to accurately and benevolently diagnose and/or treat mental health issues. If I had a mental illness diagnosed through a computer, I would not trust that diagnosis,*” illustrating that reasons for participants’ distrust included their perceptions of the technology’s inability to accurately infer or diagnose health conditions.

Feelings of distrust and perceived emotion AI-induced risks led to anticipated refusal towards being subjected to emotion AI at the workplace. 3.5% of participants described that they would refuse emotion AI use such as by quitting their job, not accepting a job that uses such systems, or

not giving consent to its use, 78.6% of whom identified with a minoritized identity. For instance, P124 who is retired stated, “*I see no possible way ‘these systems’ could benefit me, since I will never accept employment with any organization that uses them,*” demonstrating how some participants envisioned themselves to maintain their power and autonomy by rejecting jobs and organizations that use emotion AI. To add, P155 stated that, using emotion AI systems in the workplace, “*is an invasion of privacy that [they] would never agree to,*” further highlighting how participants would potentially reject giving consent to the use of emotion AI if they have the option to do so. P292, a part-time employee, also stated that if their workplace began using emotion AI systems, it would be a sign for them that “*it is time to find a new employer,*”. While these anticipated responses highlight negative attitudes towards being subjected to emotion AI at work, it is important to note that declining a job offer or quitting a job are highly privileged acts which many data subjects do not have and, therefore, would continue to work under conditions they find harmful; that is, assuming data subjects would be aware of being subjected to emotion AI in the workplace in the first place.

6.5.2 Emotion AI in Healthcare

Focusing on both potential benefits and risks, participants provided open-ended responses after self-rating comfort with a series of survey vignette questions scoped to healthcare applications of emotion AI⁴.

Conceptualizing how emotion AI could reshape their experience with mental health services, participants acknowledged the promise of emotion AI to enhance mental healthcare assessments, improve diagnosis and treatment personalization, and facilitate the disclosure of sensitive information. However, alongside these potential benefits, participants expressed considerable concern regarding the accuracy of emotion AI assessments, the risk of bias in treatment, and the potential erosion of patient autonomy. Many voiced apprehensions about the possibility of emotion AI exacerbating existing disparities in mental healthcare access and quality, particularly for minoritized groups. Highlighting how emotion AI’s deployment can lead to new and worsened procedural, distributive, and interactional injustices in the provisioning of mental healthcare, this section emphasizes the importance of considering the contextually-situated relational and distributive implications of emotion AI. By illuminating these critical issues, this analysis contributes to a deeper understanding of the ethical considerations surrounding emotion AI in mental health, advocating for frameworks that prioritize emotional privacy and protect the dignity and rights of all patients.

Our findings illustrate participants’ perceptions of potential impacts that emotion AI in mental

⁴This section is based on: Kat Roemmich, Shanley Corvite, Cassidy Pyle, Nadia Karizat, and Nazanin Andalibi. 2024. Emotion AI Use in U.S. Mental Healthcare: Potentially Unjust and Techno-Solutionist. Proc. ACM Hum.-Comput. Interact. 8, CSCW1, Article 47 (April 2024), 46 pages. <https://doi.org/10.1145/3637324>. This material is based upon work supported by the National Science Foundation under Grant Nos. 2020872 and 2236674.

healthcare could have on data subjects in practice, surfaced from our analysis of their open-ended responses to survey questions regarding the benefits and harms/concerns they anticipate with integrating the technology in healthcare.

We identified four main perceived impacts emotion AI may pose to data subjects when used to address the following existing mental healthcare challenges: 1) improve mental healthcare assessments, diagnoses, and treatments; 2) facilitate data subjects' mental health information disclosures; 3) identify potential data subject self-harm or harm posed to others; and 4) increase involved parties' understanding of mental health. While participants shared perceptions that emotion AI may be beneficial, they also raised concerns regarding how the technology may, consequently, exacerbate extant challenges in mental healthcare and harm emotion AI data subjects: 1) increase inaccurate assessments, diagnoses, and treatments along with providers' biases; 2) reduce or remove data subjects' voices and interactions with providers in mental healthcare processes; 3) inaccurately identify potential data subject self-harm or harm posed to others with implications for negative wellbeing effects; 4) involved parties' misuse of emotion AI inferences with consequences to (quality) mental healthcare access and data subjects' privacy.

To contextualize our findings, we include 1) overall percentages and counts of participants who shared responses relevant to each finding; 2) percentage of the overall count representing participants who identified with at least one minoritized identity that we oversampled (i.e., a person of color, gender minority, lived experience with mental illness) to foreground participants who may experience exacerbated harms from emotion AI use [121, 829, 865, 616, 731, 720, 927]; and 3) the relevant factorial vignette(s) in footnotes that we interpret as related to each finding.

6.5.2.1 Contextual Impacts: Mental Healthcare Provisioning and Harm Prevention

Inaccurate, inefficient, and biased mental healthcare provisions contribute to inadequate healthcare [730, 23, 815, 189, 481, 914, 557]). Some participants recognized emotion AI's potential to mitigate these mental healthcare challenges by enabling accurate, efficient, and unbiased mental health assessments, diagnoses, and treatments.⁵ However, many participants also explicated ways that using emotion AI to improve existing mental healthcare provisions may exacerbate already inadequate mental healthcare by producing inaccurate mental health inferences and heightening providers' biases against patients (i.e., emotion AI's data subjects).

⁵Acknowledged potential uses of emotion AI to improve mental healthcare provisions map to factorial vignettes that posed the use of emotion AI in healthcare to infer the mental health state of patients (on a group level and individually), infer patients in need of wellbeing support, assess the overall health of individual patients, diagnose mental health illness and neurological disorders in patients earlier than otherwise possible, automatically alerting healthcare provider(s) when patients may need support, and avoiding human judgment and subjectivity.

Assessments, diagnoses, and treatments. Some participants acknowledged how emotion AI may improve mental healthcare provisions by rectifying inaccurate, inefficient, and biased mental health assessments, diagnoses, and treatments carried out by human mental healthcare providers. Providers' failure to meet data subjects' mental health needs can negatively affect their mental and overall health, which participants recognized emotion AI to potentially mitigate by addressing inaccuracies, inefficiencies, and biases within mental healthcare.

Addressing inaccurate mental healthcare provisions. 8.4% of participants ($n = 33$), 84.7% of whom identified with at least one minoritized identity, mentioned that emotion AI may mitigate inaccurately assessed, diagnosed, or treated mental health conditions. For instance, P57, a white man with a mental health condition, described how emotion AI could potentially lead to more accurate mental health diagnoses by detecting patterns that (often inattentive) human providers may miss: "*machines are great at picking up things that humans aren't and vice versa, so a doctor augmenting their diagnoses and treatments with various robots and AI assistants have major potential to improve care across the board.*" P23, a Southeast Asian woman with a mental health condition, echoed P57's sentiments: "*sometimes doctors are busy writing notes or [are] distracted...the system could help detect things the doctor didn't notice.*" P57 and P23's comments point towards extant challenges whereby inattentive or overworked providers may overlook patients' needs, consequently provisioning inaccurate diagnoses and treatments. Integrating emotion AI that may be better at "*picking things up*" than "*busy*" providers is perceived to potentially facilitate more accurate care provisions.

Facilitating more efficient mental healthcare provisions. 52.7% of participants ($n = 208$), 76.4% of whom identified with at least one minoritized identity, acknowledged that emotion AI could address current inefficiencies in detecting and diagnosing mental health conditions. P7, a transgender, non-binary white person with multiple mental health conditions, described how they were "*diagnosed with ADHD later on in life, so definitely making resources so that people can get diagnoses and properly treated faster would have helped [them].*" For P7, resources that could have aided in an earlier diagnosis and treatment would have benefited them. P33, a white woman who reported no mental health conditions, shared similar thoughts that "*it could be nice for the program to notice a particular health concern before [she] did to facilitate faster treatment.*" These examples highlight how current mental healthcare processes are ill-equipped to provide efficient diagnoses and treatments. This present challenge relates to unique diagnostic difficulties that are partially attributed to a lack of "objective" diagnostic tests in mental healthcare [461]. Participants perceived emotion AI could potentially address these deficiencies by facilitating faster diagnosis and appropriate care to data subjects, reflecting biomedical virtue rhetoric [113, 692], which promotes an uncontested ideal that privileges a "praxis of goodness" within healthcare and legitimates the deployment of new technologies and data practices within its domain (e.g., to

facilitate early diagnosis or access to treatment) without adequate critical examination [692].

Worsened provisions. Though some participants acknowledged that emotion AI could potentially address persistent challenges of inadequate mental healthcare provisions, many also perceived how emotion AI may instead worsen them. 51.4% of participants ($n = 203$), 76.8% of whom identified with at least one minoritized identity, noted how emotion AI's potentially inaccurate inferences could worsen already inaccurate assessments, diagnoses, and treatments [815, 481, 914, 557, 81]. P321, a bi-racial woman with multiple mental health conditions, mentioned that “*culturally, expression can vary depending on many factors, which might lead to inaccurate readings,*” depicting doubt that emotion AI could accurately account for complexities in emotional expression across cultures and individual differences. Such concerns are grounded in existing literature [93, 224, 886, 464], and may be heightened for bi-racial/bi-cultural individuals like P321 with mixed identities. P321’s concerns regarding potential algorithmic biases highlight how inaccurate emotion inferences may have harmful ramifications to data subjects’ diagnoses and treatments, contesting earlier acknowledgments that inferences, if accurate, would be beneficial as reported in Section 6.5.2.1. Thus, supplanting care provisioning processes with emotion AI, itself known to demonstrate poor rates of accuracy [93] and perpetuate demographic biases [271, 472], may be an inadequate solution to addressing existing problems of inaccurate assessments, diagnoses, and treatments in mental healthcare, and in effect could *exacerbate* these problems by automatically reproducing them on a large scale.

Participants’ concerns regarding potentially inaccurate inferences also illustrate the possibility for providers to take these inferences at face value and subsequently dismiss data subjects’ voices and lived experiences in mental healthcare provisioning. P81, a white woman with a mental health condition, expressed concerns about being mislabeled by emotion AI: “*Sometimes a system could tell you that you are at risk of something when you are really operating at a safe level. Each individual has a different pain tolerance level. Their own impressions should come first before being labeled.*” P81 both reflects concerns that emotion AI inferences generated without individual baselines would mislead treatment planning and that providers privileging their perceived objectivity may weaken data subjects’ agency with automated systems that “*tell*” data subjects and providers about their mental health condition, rather than considering data subjects’ voices that “*should come first before being labeled.*” Concerns that inferences would potentially invalidate or dismiss data subjects’ voices is also an issue that some participants noted may arise if providers rely solely on emotion AI in practice, which we explore further in Section 6.5.2.2.

In sum, many participants shared concerns demonstrating how emotion AI may produce inaccurate inferences that negatively impact mental health assessments, diagnoses, and treatments, and highlight how data subjects’ voices may be disregarded if providers place more value on emotion AI inferences than their own perspectives and lived experiences. Thus, participants’ concerns

elucidate the various ways emotion AI may be an unsuitable solution that may worsen present challenges of inaccuracy in mental healthcare.

Harm prevention. In this section, we highlight participants' perceptions about emotion AI use for monitoring potential harm toward oneself or others to improve existing harm-prevention efforts – a use case commonly proposed in previous work [115, 738, 212, 697, 258, 750, 833, 248, 179, 457].

⁶ While participants acknowledged the potential merits of this use, they also raised significant concerns surrounding how emotion AI may inaccurately predict self-harm or harm to others, dangerously impacting data subjects as a result.

Identifiability. 1.5% of participants ($n = 6$), a striking 83.3% of whom identified with at least one minoritized identity, acknowledged how emotion AI could be used to monitor and identify individuals who may harm themselves or others. P322, a white woman with multiple mental health conditions, mentioned "*I think the only way this could work is by possibly monitoring a dangerous person's social media, or offering links and hotlines when someone is in need of immediate support.*" P322's comments underline perceptions that emotion AI would only be useful in this case if it invasively monitored sites of rich personal data (e.g., social media behavior) which may offer a unique window into an individual's intimate thoughts and affairs – an emerging area in which emotion AI may be used (e.g., digital phenotyping [221, 434]). However, P322's response also demonstrates the perceived potential for emotion AI use to intrude on data subjects' personal lives outside of mental healthcare and to conflate data subjects' online activity with potentially harmful offline behavior that may be inaccurately interpreted as "*someone in need of immediate support.*"

P279, a white woman with a mental health condition, acknowledged this application may be beneficial for *some* individuals, noting that it may "*help severely mentally ill people who need monitoring to stay safe,*" but that "*otherwise, [emotion AI would be] way too invasive.*" P279 highlights the perceived need to protect *other* individuals with mental health conditions from potential self-harm while raising privacy concerns that legitimate the over-surveillance of mentally ill people to keep them "safe." Relatedly, prior work on suicide risk prediction on Facebook emphasizes that real-world implementations of automated harm prevention requires monitoring *all* users to effectively identify potential harm [331, 364]. Thus, it is critical to consider how *all* data subjects' privacy may be compromised for the purpose of harm prediction and prevention, the willingness of data subjects to be subjected to such surveillance, and the consequences of potentially inaccurate and harmful interventions.

Inaccurate risk predictions. Though using emotion AI for harm prevention may alluringly

⁶Perceptions of emotion AI use to identify potential self-harm or harm posed to others relate to factorial vignettes that asked participants about using emotion AI to infer whether patients are at risk of harming themselves and inferring harm to others.

promise data subjects' safety, its inferences may be inaccurate and have detrimental impacts on data subjects. This section draws from responses in Section 6.5.2.1 where 51.4% of participants ($n = 203$), 76.8% of whom identified with at least one minoritized identity, highlighted concern regarding potentially inaccurate inferences. We note that our qualitative analysis broadly categorized inaccurate emotion AI inferences as a perceived potential result of emotion AI use. From this analysis, we found that some participants more specifically highlighted emotion AI's potential to falsely identify individuals at risk of harming themselves or others.

P135, a white man who reported no mental health conditions, shared, "*I really don't think they can [be beneficial]. I can be mad about something and the system may interpret that I will hurt someone when in reality, I just want to tell someone about what happened.*" P135 also shared concerns he may "*involuntarily be subjected to unnecessary help or even restraint if the system concluded [he] was at risk of hurting someone when in reality, [he] just wanted to vent and it could be over and done within a few minutes of blowing off steam.*" P135's remarks contest earlier sentiments in Section 6.5.2.1 regarding emotion AI as a tool for monitoring and detecting potential harm. Instead, P135 highlights how through a lack of adequate contextual understanding, emotion AI may inaccurately identify data subjects in unsafe situations (i.e., conflating venting behavior with danger and risk) and consequently expose data subjects to harm from inappropriately excessive responses (e.g., forced "help" or restraint). P193, a white woman who reported no mental health conditions, shared similar concerns: "*If someone is potentially self-harming or wants to commit suicide, this program may not recognize that. Or it could falsely flag someone as such. Reporting this to health services could be detrimental to the patient.*" P193 raises the question of whether subjecting data subjects to invasive surveillance methods is warranted if it may fail to identify at-risk individuals on the one hand, and on the other hand, expresses concern that false inferences can also be detrimental to the identified individual. Reporting inaccurate harm predictions "*to health services,*" for instance, may unjustly lead to police intervention and involuntary commitment [731, 684] which can result in unjustified physical harm or brutality [696] that disproportionately impacts minoritized communities [455, 758]. Participants' concerns also relate to notions of consent (mentioned in Section 6.5.2.1) and dismissed data subject voices (referred to in Section 6.5.2.2) whereby data subjects may be forced under surveillance without the ability to contest inferences made by emotion AI and associated interventions.

Adverse wellbeing effects. As a result of being monitored by emotion AI, 4.6% of participants ($n = 18$), 83.3% of whom identified with at least one minoritized identity, anticipated being subjected to emotion AI would negatively impact their wellbeing. P360, a white transgender person who reported having a mental health condition, stated that the idea of emotion AI in mental healthcare "*makes [them] uncomfortable.*" P373, a Black man who reported no mental health conditions, also said, "*continuously using the systems may cause [him] anxiety,*" while P282, a white man

who reported no mental health conditions, described more specifically that emotion AI “*could lower self-esteem, frighten, put on the defensive, or otherwise make matters worse for individuals.*” These responses demonstrate the potential for emotion AI-enabled patient monitoring to induce negative wellbeing effects (e.g., feelings of fear, hypervigilance, low self-esteem) arising from the surveillance of data subjects’ intimate and personal emotions, contradicting emotion AI’s purported use to improve wellbeing [616, 626, 670].

Care Quality and Access. Although emotion AI use is acknowledged to potentially enhance involved parties’ understanding of mental health broadly, involved parties’ possession of data subjects’ emotion AI inferences may lead to harmful data misuse that may worsen mental healthcare inaccessibility (reflecting prior work [452, 420]). 4.8% of participants ($n = 19$), 57.9% of whom identified with at least one minoritized identity, described how involved parties’ (i.e., providers, insurance companies) access to emotion AI inferences may hinder data subjects’ mental healthcare quality and insurance coverage, negatively affecting data subjects’ access to professional medical care. P368, a Black woman with a mental health condition, shared that “[emotion AI] could cause healthcare providers to create biases about their clients and even drop them from their system altogether. Deeming them ‘high risk’ and refusing to cover them.” P368 illuminates concerns that emotion AI inferences may influence providers’ biases (described in Section 6.5.2.2), potentially leading to harmful decisions that limit data subjects’ access to mental healthcare and jeopardize their wellbeing. Similarly, P50, a white woman with multiple mental health conditions, stated, “As a neurodiverse person with mental health issues, I worry that the quality of care that I would receive from healthcare providers would decrease dramatically if this technology was put in place by health providers to cut costs.” These concerns point to the potential for emotion AI to entrench mental healthcare inequalities, and may have been shaped by the larger context of healthcare algorithmic systems that determine risk differently between Black and white patients, with downstream disparate effects on insurance coverage and costs [173, 81, 719, 662]. Similarly, P117, a multi-racial woman with multiple mental health conditions, stated that she believed emotion AI “*could be misused to limit access to certain treatments or services*” while P51, a Latina with a mental health condition, stated, “*This seems like it could be misused by health insurance companies to decrease client support and increase costs for clients.*” Overall, participants shared concern that underlying profit motives would drive the premature adoption of potentially harmful emotion AI. Perhaps P159 said it best: “*Given the sorry state of AIs, the baked-in biases, and the overcapitalization of healthcare, I can only see this being used to deny service as a means of controlling costs, increasing profits and sold to major ad networks as yet another profit center without our knowledge or consent.*”

6.5.2.2 Relational Impacts: Patient Voice, Agency, Dignity

Many participants referred to the existing mental healthcare challenge whereby patients' voices are lost or ignored in provider interactions and care provisioning [169, 855, 615, 748, 730]. Participants acknowledged emotion AI could potentially facilitate conversations around mental health and amplify data subjects' voices⁷. Yet, participants also acknowledged how emotion AI could, instead, reduce or remove their voices in mental healthcare processes, hindering their ability to take part in decisions about their own mental health and wellbeing. In this section, we first unravel how participants' recognition of dismissed patient voices in traditional mental healthcare rendered emotion AI an enticing potential solution to amplify data subjects' voices. We then describe participants' concerns that emotion AI could in practice reduce or remove data subjects' voices and their ability to interact with mental healthcare providers altogether, harming data subjects' agency in mental healthcare processes. By foregrounding data subjects' perspectives, we show how implementing emotion AI to "solve" challenges related to patients' voices inadequately addresses this extant issue.

Patient voice. Human voices and interactions are important aspects of healthcare in general [273, 150]. Some participants anticipated how emotion AI could facilitate mental health-related disclosures; some anticipated that its use in this domain may threaten the inclusion of patients' voices in mental healthcare processes.

Facilitating mental health information disclosures. 9.6% of participants ($n = 38$), 92.1% of whom identified with at least one minoritized identity, noted difficulties with openly communicating their mental health concerns with their providers, which emotion AI may help to resolve. P242, a Black woman who reported no mental health conditions, described how emotion AI systems "*could be beneficial if they actually provide emotional support, [as] it would feel less isolating and maybe like I was being seen if the program was acknowledging and backing up that my words and expressions actually indicate what I say they do and not what a medical professional (who is not listening anyway) has decided.*" P242's remarks highlight foundational issues within mental healthcare regarding how providers often neglect their patients' voices and gaslight their concerns [68] (a problem pervasive especially for minoritized communities and Black folks, in particular [766, 68, 173]), leaving patients feeling unheard, unseen, and isolated, and potentially resulting in detrimental mental healthcare provisions. Participants like P242 acknowledged emotion AI as a potential solution to address this problem by legitimating patient concerns that providers often ignore, *if* the systems in practice provided meaningful emotional support and supported data

⁷Participants' responses regarding emotion AI use to amplify data subjects' voices align with the following factorial vignettes: using emotion AI in mental healthcare to develop an intelligent computer program to conduct mental health therapy; inferring moments patients may need emotional support and responding with an intelligent computer program.

subjects' voices during mental healthcare processes. As a result, emotion AI was perceived to potentially promote mental health information disclosures between mental healthcare patients and their providers by "backing up" the information patients disclose.

Dismissive effects. 4.6% of participants ($n = 18$), 72.2% of whom identified with at least one minoritized identity, reported concerns regarding how mental healthcare providers may become heavily reliant on emotion AI over data subjects' own voices and lived experiences. P113, a white woman with a mental health condition, shared that emotion AI could negatively impact her "*if doctors place complete confidence in software and discount the information [she] may tell them if it doesn't support software.*" P113 points to the perceived potential for providers to place more value on emotion AI inferences over data subjects' own information regarding their mental health—privileging biomedical expertise [113, 692] that echoes histories of mental illness patients being discredited, discounted, and gaslit about their own mental health-related experiences in interactions with providers [766, 173]. To add, P87, a white man with a mental health condition, stated: "*There is a danger then if such systems become widespread, it will become very difficult to refute their diagnoses,*" highlighting a concern that data subjects would be unable to contest the inferences that providers may take at face value. These remarks point toward the perceived potential for emotion AI inferences to become an unyielding point of reference for decisions on data subjects' mental health, without adequate consideration for data subjects' lived experiences in decisions related to their mental health. Participants' concerns regarding a reduction in data subjects' voices relate to other concerns that may surface from emotion AI use including implications for the accuracy of mental healthcare provisions (explicated in Section 6.5.2.1) and biases in mental healthcare processes (covered in Section 6.5.2.2) whereby data subjects lose the ability to dispute the provisions or processes facilitated by emotion AI use.

Diminished interactions with providers in mental healthcare. In addition to data subjects' voices potentially becoming far removed from mental healthcare processes, 12.9% of participants ($n = 51$), 76.5% of whom identified with at least one minoritized identity, also noted the potential for emotion AI use to reduce or remove interactions between data subjects and providers. P8, a white man with a mental health condition, stated, "*I don't think it's a good idea for anyone except whoever's building and selling these systems to remove much more of the human from human medicine.*" In other words, the human element (e.g., patient-provider interaction) is a salient aspect of mental healthcare that, if removed, may only ultimately benefit those who create and profit from the technology rather than those in need of care and subjected to its use. P8's belief in the importance of the human element highlights the need for human involvement in mental healthcare processes (even as problems with provider-patient interactions may persist), as reducing or supplanting these processes with emotion AI may consequently lead to harmful mental healthcare provisions. For example, P152, a white woman who reported no mental health conditions, stated that "*relying too*

heavily on computer-assisted programs can lead to poor healthcare. It could be tempting to use these programs to allow providers to step too far back from the process.” P152’s remarks illustrate how over-reliance on emotion AI could potentially result in “*poor healthcare*” outcomes and diminished interactions with providers, which may have implications for inaccurate (described in Section 6.5.2.1) or biased (explicated in Section 6.5.2.2) mental healthcare assessments, diagnoses, and treatments.

Some participants described how specific emotion AI-enabled technology, such as a chatbot that may augment or replace providers’ involvement in mental healthcare provisions, may produce or perpetuate harm to data subjects’ mental health. P242 noted that “*chatting with a chatbot for mental health support, that is only capable of providing canned planned responses could cause me to feel isolated, invisible and lead to depression or self-harm.*” P242’s response highlights the potential for emotion AI interactions that inappropriately respond to emotion inferences to fail to meet data subjects’ mental health needs and expose them to psychological harm. To add, P50, a white woman with multiple mental health conditions, reported being “*concerned about the human element of mental health evaluations and treatments getting lost as human beings are social animals and we are better at reading one another than a computer can ever be. Psychological healing also takes place primarily within human relationships, not AI chatbots.*” P50’s response further underscores the importance of human interaction to mental healthcare processes, as leaving high-stakes decisions and interpersonal connections up to algorithmic models and “*chatbots*” may be harmful and insensitive to data subjects’ mental health needs. This view reflects previous work on how emotion recognition-enabled wellbeing interventions are perceived to provide inadequate care fundamentally because of the lack of human interaction involved [731].

Altogether, participants’ anticipated impacts warn that emotion AI-enabled technologies that replace human patient-provider interactions (e.g., chatbots) may augment therapeutic interactions with adequate, non-human automated care to the extent that they become artificial, result in psychological harms (e.g., feelings of neglect and ill-treatment), and, more fundamentally, reduce or remove data subjects’ voices and interaction with providers in mental healthcare processes.

Provider bias. 6.3% of participants ($n = 25$), 84% of whom identified with at least one minoritized identity, shared perceptions that emotion AI could potentially mitigate the role of mental healthcare providers’ biases in care provisioning. Participants acknowledged the possibility that emotion AI may augment providers’ decisions with unbiased inferences, reducing the possibility for providers’ biases to fully account for mental health assessments, diagnoses, and treatments. For instance, P334, a white transgender man with multiple health conditions, stated, “*I’m an adult ADHD person and would have benefited GREATLY from technology such as this had it been available when I was younger, as the path to my diagnosis was arduous and oftentimes hindered by non-objective*

professionals.” P334 described how his difficult experience involving biased providers impeding his timely diagnosis could have been improved with more objective approaches like that promised by emotion AI. Similarly, P95, a Black woman with multiple mental health conditions, stated, “*I figure with all the experiences I've had with human doctors concerning my mental health and physical health... a program that's able to access a great deal of information and give an unbiased evaluation, couldn't be any worse.*” Based on her personal experiences with insufficient care, P95 thought that the “*unbiased evaluation[s]*” emotion AI promises at the very least would not be a worse alternative to biased providers. P95’s experiences are situated in a longer history of medical gaslighting and disparate treatments for Black patients that is entangled with majority white providers’ subjective biases against Black patients and other minority patients [173, 766]. Overall, the experiences participants shared highlight the present mental healthcare challenge of biased providers negatively affecting if and how patients receive appropriate mental healthcare and their perceptions that emotion AI could potentially mitigate mental healthcare provisioning hindered by providers’ biases.

Worsening provider bias. 14.2% of participants ($n = 56$), 80.4% of whom identified with at least one minoritized identity, described how algorithmic biases laden in emotion AI could amplify mental healthcare providers’ own biases, negatively affecting the mental healthcare data subjects receive. In contrast to the potential for emotion AI to mitigate providers’ biases acknowledged in Section 6.5.2.1, participants were concerned that emotion AI could potentially intensify providers’ biases when delivering mental healthcare.

Previous work has surfaced a range of algorithmic biases in emotion AI [271, 472, 616], a concern reflected by many participants who noted that emotion AI may encode biases that providers may then use to legitimate their own. P331, a white woman with multiple mental health conditions, described how emotion AI “*could be biased or based on stereotypes that could lead to incorrect information and harm by falsely associating traits with someone.*” P7, a transgender, non-binary white person with multiple mental health conditions, similarly mentioned concerns with potential biases within emotion AI: “*this system [emotion AI] could be built with gender and race biases that could harm myself and other individuals, especially if the system is rigid in what it deems ‘unhealthy looking.’*” P331 and P7’s concerns highlight the perceived potential for biased emotion AI inferences to perpetuate harmful stereotypes, and thus facilitate flawed mental healthcare provisions. P7 went on to describe the manifold ways biased emotion AI systems could affect the processes mental healthcare providers follow: “*Many times when women or femme presenting persons go into doctor's offices and don't look 'presentable enough' symptoms go overlooked or ignored but going in without makeup can also look 'sickly' and the system could detect that as illnesses that patient does not have. This is also something that could impact disabled (hard-of-hearing and deaf) people and non-native speakers with the speech analysis portion since they would not be speaking with the*

typical speech patterns that this program is ‘looking for.’” P7 highlights how societal stereotypes reflected in emotion AI training data via biased inferences could potentially influence providers’ decision-making regarding data subjects’ mental health assessments, diagnoses, and treatments.

Some participants anticipated that providers may use emotion AI to defend their biases, which can harm mental healthcare provisions. P306, an Asian woman with a mental health condition, stated that “*we as humans are bad at understanding intersectionality so how do we expect to code a computer to understand it? I would hate for more discrimination to be a result of this,*” pointing to concerns that emotion AI would be incapable of understanding data subjects’ multiple intersecting identities in a meaningful and non-reductive way when evaluating mental health. Thus, the technology’s potential incapability to adequately account for data subjects’ complex, intersecting identities may lead to “*more discrimination*” in healthcare, rather than combat it. The obscurity surrounding how emotion AI technologies are developed and the decisions developers make when building such systems complicates the expectations (as P306 notes) we may have about emotion AI. This lack of transparency and regulation further raises issues in the trade-off between accuracy and fairness in algorithmic systems [532], whereby the inferences and interventions made by emotion AI may lead to disparate negative consequences that are compounded for data subjects with intersecting minoritized identities.

P335, a white transgender man with multiple mental health conditions, mentioned similar concerns: “*There is also the given of human subjectivity still being there when the data is given to the healthcare provider, so it depends on the provider and their potential biases as well at the end of the day.*” Participants like P335 shared concern for the potential that emotion AI inferences could not only exploit providers’ biases, but that providers’ own biases could limit their ability to recognize potentially flawed algorithmic decisions. Thus, human-in-the-loop processes – often proposed to stem concerns with algorithmic decision-making [246] – would do little to address biased providers’ limited ability to recognize potentially inaccurate results, which may be ultimately dangerous to data subjects’ mental health and wellbeing in practice. Notably, participants’ concerns regarding the potential for providers’ biases to be exacerbated and its effects on mental healthcare provisioning with emotion AI use contrast emotion AI’s proposed potential to mitigate biases in mental healthcare (described in Section 6.5.2.1).

6.5.2.3 Data Access

Various parties are involved in mental healthcare, namely mental healthcare providers who distribute care provisions; insurance companies who determine access to mental healthcare; academic researchers who may use mental healthcare data to advance knowledge about mental health; and patients who are most affected by these parties’ involvement in their mental healthcare. This section analyzes how participants acknowledged emotion AI inferences to potentially enhance a broader

understanding of mental health, specifically for the benefit of mental healthcare providers and academic researchers.⁸ However, this acknowledged emotion AI use highlights the lack of adequate mental health understanding today, creating an appeal for emotion AI's promises to enhance such understanding. Yet, participants expressed their worries regarding the potential for involved parties, including providers and insurance companies, to misuse emotion AI inferences in ways that may create barriers to accessing (quality) mental healthcare and compromise data subjects' privacy.

Enhance involved parties' understanding of mental health Mental healthcare requires an understanding of patients' needs. Due to the complex range of mental health conditions and symptoms, it is difficult to fully understand and tend to patients' needs. 9.6% of participants ($n = 38$), 78.9% of whom identified with at least one minoritized identity, acknowledged how emotion AI may be used to enhance involved parties' (i.e., providers and academic researchers) understanding of mental health, which may allow providers to better tend to data subjects' needs and for academic researchers to advance understanding of mental health. P253, a Black woman, stated that emotion AI "*would benefit [me] greatly as having more ways to assess mental/physical health would give healthcare providers a better understanding of the patients they deal with,*" illustrating a perceived need for resources that aid providers' understanding of data subjects' mental health to then provide sufficient care. Additionally, P285, a white man who reported no mental health conditions, said "*They [emotion AI] could provide another way of providing insight into what is going on with me, or, if being done for research the researcher could help practitioner better understand some aspect of their practice.*" P285 describes how emotion AI inferences could potentially be used in research to both advance new knowledge about mental health and generate insights into mental health that would enhance healthcare practitioners' understanding of and approach to improving mental healthcare. These perceived emotion AI uses for enhancing involved parties' understanding of mental health (which may or may not result in improved care) point to a deficient understanding of patients' mental health experiences that currently challenges the state of mental healthcare.

These concerns demonstrate the perceived potential for mental healthcare systems and insurance companies to misuse emotion AI to restrict mental health provisions and increase mental healthcare costs, economically benefiting these involved parties at the expense of harming data subjects needing mental health services by making it difficult to afford or access quality care. Furthermore, they reflect perceived violations of contextual integrity [646] where data subjects' emotional information may be inappropriately and detrimentally removed from its intended use for mental healthcare provisions to facilitate information misuse, and underscore the importance of understanding the perspectives of people *with* mental illness(es) concerning how they may be adversely impacted by

⁸The perceived potential use of emotion AI for enhancing involved parties' understanding of mental health is in line with factorial vignettes that describe the purpose of giving healthcare providers increased understanding about patients through data-driven insights and to share emotion AI inferences with academic researchers to help them learn more about mental health, as part of a research partnership.

emotion AI in the high-stakes context of mental healthcare.

Intruding patient privacy Many participants shared concerns about emotion AI violating their privacy. Participants wondered if and how emotion AI would be regulated and how the data emotion AI collects and infers would be handled to ensure data subjects' privacy is protected and secured in practice. Participants described privacy concerns associated with emotion AI data handling practices and data subjects' ability to meaningfully consent to the collection and sharing of their emotional information. Throughout this section, we map participants' perceptions of potential privacy intrusions to distinct privacy harms outlined by Citron and Solove [194].

51.4% of participants ($n = 203$), 77.8% of whom identified with at least one minoritized identity, shared concerns about how mental healthcare providers using emotion AI could invade data subjects' privacy, leading to myriad privacy harms. P92, a white woman with a mental health condition, stated, “*Sometimes we don't want to reveal things about ourselves. This would make me feel very vulnerable and exposed,*” pointing to how emotion AI may constrain data subjects' agency in exercising whether and to what extent they disclose private and sensitive mental health and emotion-related information to their provider. P92's concerns also reflect potential autonomy and emotional harms associated with emotion AI use in mental healthcare [194] by challenging data subjects' freedom to make decisions regarding their own data, including exposing their vulnerable emotions and personal information to involved parties other than their healthcare provider(s). Respecting data subjects' autonomy to reveal their emotional information (or not) is particularly salient given its sensitivity and vulnerability to abuse [52, 735].

Some participants also asked various questions concerning if and how emotion AI in mental healthcare would be regulated. P230, a white woman with a mental health condition, asked: “*How will the data be held? Will it be deleted afterward? If sent in for research, how many others will witness my data?*” P230's questions reflect participants' considerable privacy concerns surrounding emotion AI in mental healthcare, including its potential to harm patient autonomy [194] as a result of opaque and unregulated emotion AI data handling practices.

Even if mental healthcare providers and emotion AI vendors were to implement privacy-preserving designs and strict security controls, it is important to note that participants remained concerned about the potential for data leakages. P149, a white man with a mental health condition, noted: “*Even where the healthcare provider is ensuring confidentiality, I think there should be discussion as to whether such readings could ever be turned over to, or subpoenaed, by law enforcement officials or courts, and if so, under what specific circumstances,*” expressing concern about the potential for courts and law enforcement to compel healthcare providers to share an individual's stored emotion AI data, which is particularly notable given the history of forced hospitalization of those with mental health conditions [731, 684]. In addition, P345, a non-binary white person with a mental health condition, expressed concerns “*about the safety of this information [generated by*

emotion AI” and its potential to be commodified if it were “*used outside of the healthcare field such as by advertisers or companies.*” Similarly, P265 described: “*The [collection of data subjects’] images/videos are of concern because they may get into the wrong hands or could be used for facial recognition beyond the supposed purpose. Even if I sign a consent, I do not trust that this information will be used appropriately and once it gets out into the open you are at a loss.*” As P265 highlights, data subjects’ emotion AI-generated inferences could be re-purposed and may be leaked to third parties, exposing patients’ sensitive emotional information “*into the open*” and outside their control. Despite existing safeguards that may guarantee higher standards of confidentiality and data protection if emotion AI is used in clinical settings (e.g., the Health Insurance Portability and Accountability Act of 1996⁹) P149 and P345’s concerns highlight the perceived potential for external data sharing and reuse beyond the original purpose of adopting emotion AI, and for data leakages (e.g., subpoenas, data breaches) to expose patients’ sensitive emotional information in invisible and uncontrollable ways.

Potential external emotion AI data sharing and reuse may also involve relationship harms [194] as the trust between data subjects and their providers may be negatively impacted due to compromised patient confidentiality associated with emotion AI use. Moreover, the privacy harms implicated by emotion AI use may extend to individuals beyond the intended data subject. As P149 described: “*There’s also the bystander issue... how would such audio or video recording ensure that other people’s privacy in my residence was protected?*” P149’s concerns point to the potential impact that emotion AI may have on *others’* privacy (e.g., partners, children) that may not be directly monitored by healthcare providers using emotion AI but whose interpersonal privacy is nonetheless implicated, and may fall outside the scope of any existing safeguards (i.e., HIPAA) designed to protect individual *patient* health information.

In all, participants described their concerns about how involved parties within and beyond the mental healthcare context may obtain and misuse data subjects’ emotion AI inferences in ways outside of their intended purpose and the control of data subjects, consequently exposing data subjects to a range of privacy harms.

⁹The Health Insurance Portability and Accountability Act of 1996 or HIPAA is a federal law meant to protect patients’ sensitive health information from disclosures without patients’ consent or knowledge. However, we note that HIPAA does not cover all digital health applications, even when used in coordination with patients’ healthcare providers, and only covers certain entities (providers, insurance companies, and business associates). Thus, patients may unwittingly authorize the disclosure of their personal health information to third parties which may or may not be considered a covered entity, and thus, not covered by HIPAA [830].

6.6 Discussion

Workers' and patients' ability to protect their *emotional privacy*—managing whether, how, and to what extent information about their emotions is collected, used, and shared—is increasingly threatened by technologies that automatically infer human emotions. Although proponents highlight potential benefits, such as earlier medical diagnoses, harm prevention, and enhanced emotional wellbeing, these claims remain largely speculative and scientifically unvalidated. Meanwhile, the automatic inference of emotional information poses significant privacy risks, particularly in the U.S. as powerful actors like employers and healthcare providers adopt these technologies with minimal oversight.

As emotion AI development and regulation continues to evolve, the privacy perceptions of those subject to emotion inference should help shape ethical standards. Emotional privacy is not simply about disclosure; it reflects contextually-dependent norms [646] governing how emotional information is collected and used. It is a matter of degree [722]: while emotions may be perceptible through facial expressions, language, or vocal patterns, this does not justify unrestricted automatic inference. Emotion inference technologies expand the set of contextual actors—both human and non-human—who can access, use, and share individuals' emotional information in opaque ways. Individuals should have a meaningful capacity to decide whether, how, and to what extent their emotional information is inferred and used.

Protecting emotional privacy is essential not only for respecting human dignity and promoting individual wellbeing. Privacy and agency in emotion-inference technologies are critical to sustaining institutional trust and enabling people to engage with such systems without undue risks. For scholars concerned with justice and human values in socio-technical systems, it is vital to understand how to protect emotional privacy in ways that reflect the factors shaping privacy judgments—particularly among those most vulnerable to harm.

Our findings contribute to this goal by empirically identifying contextual and identity-based vulnerabilities that influence emotional privacy judgments. These vulnerabilities are especially pronounced among minoritized groups, whose emotional privacy needs and concerns can differ markedly from dominant groups. Relying solely on socially dominant “internal standards of justice” [666] to define privacy norms risks reinforcing systemic injustices and silencing dissenting views [646, 590, 138, 871]. Aligning the development, design, and regulation of emotion-inference technologies with the unmet needs and persistent concerns of those most vulnerable to technological impact benefits everyone.

Our results also have significant policy relevance. Notably, the patterns in our study mirror many elements of the EU AI Act and its clarified application guidelines [300]. We observe striking alignment between public intuitions and the regulatory architecture: where the Act imposes

strict limits—regulating biometric inputs, prohibiting individual profiling in employment, addressing power asymmetries—participants’ comfort drops sharply. Where the Act permits narrow exceptions—workplace safety, neurological monitoring for medical applications, or non-identifiable aggregated insights—comfort increases. This convergence suggests that the distinctions respondents draw closely match the EU’s regulatory reasoning, offering regulatory legitimacy to the privacy judgments surfaced in our study.

Beyond reinforcing current regulatory directions, our findings offer a model for anticipating future governance challenges. As commercial innovation adapts to regulatory constraints and as jurisdictions develop more detailed rules, this study provides both empirical evidence to inform those efforts and a methodological approach capable of identifying socially salient privacy boundaries as they evolve. In the sections that follow, we discuss implications for practice, policy, and research, emphasizing:

1. **inference minimization**—limiting the purposes for collecting or using emotional information
2. Recognizing emotional information as a **sensitive data category**;
3. Advancing **contextual and demographic sensitivity** in the design, application, and regulation of emotion-inference technologies.

6.6.1 Bounding Inference Purpose

Contextual integrity assumes data purpose is constrained implicitly—encoded in transmission principles and justified by the context’s goals—rather than specified as a stand-alone parameter [646]. Our results show that purpose nonetheless drives emotional privacy judgments. In the workplace, performance-scoring inferences—a common managerial practice designed to boost productivity [80, 925, 34]—elicited *lower* comfort. By contrast, employers sharing workers’ emotion inferences with academic researchers—an extraneous purpose that does not obviously advance workplace goals—*raised* comfort levels. A similar pattern surfaced in healthcare. Automating interventions or diagnosing mental illness—purposes tightly coupled to clinical objectives—*lowered* comfort, likely because they override patient-initiated disclosure and undermine interpretive agency. Yet neurological disorder screening, another clinical use, *increased* comfort.

These divergences cannot be explained by contextual integrity’s five canonical parameters or by contextual goals alone. Instead, the explicit *purpose* of the inference—together with the type of information, actors involved, and the transmission principles governing the flow—decisively shapes emotional privacy judgments. As Nissenbaum has noted, purpose’s relevance to contextual integrity has become more salient alongside evolving technologies and data practices [651], and adding a purpose dimension may be a “necessary [policy] antidote” [650]. These findings support

extending contextual integrity to treat purpose as a constitutive contextual parameter, enabling more precise governance through purpose limitation and inference minimization rules. Drawing on the quantitative patterns and complementary qualitative results from this study [219, 733], we propose a narrow, empirically grounded palette of permissible purposes for emotion inference. Key safeguards include:

- **Purpose binding:** Mirroring the EU AI Act’s risk-based, context-specific approach [300], define each permissible purpose narrowly and exhaustively (e.g., real-time fatigue detection in safety-critical roles). Specify parallel prohibited purposes (e.g., burnout or depression screening). Any secondary use—or any use outside the narrowly scoped carve-out—should be categorically barred.
- **Granular, opt-in consent:** allow individuals to opt-in or withdraw for each distinct use of emotion data, raw or inferred.
- **Ex ante validation:** require evidence of claimed benefits before deployment
- **Minimal retention:** store only what is strictly necessary for the stated purpose
- **Robust controls:** enforce access limits, encryption, and anonymization, overseen by independent auditors.

Embedding these constraints in law, design, and institutional policy would operationalize contextual integrity’s normative commitments, protecting emotional privacy while allowing only narrowly justified, socially valuable uses of emotion AI systems.

6.6.2 Emotion Data Sensitivity

Our findings confirm that workers and patients perceive emotion inferences as highly sensitive, often rating them as more sensitive than established categories such as biometric or genetic data. Yet emotion data remains unrecognized as a special category of sensitive information in most privacy frameworks [84].

This sensitivity reflects significant, context-specific risks. In workplaces, emotion inferences could enable discrimination on the basis of perceived mental disability—even without direct disclosure. In healthcare, inaccurate inferences may trigger misdiagnosis or stigma. If exported beyond the original context (e.g., sold to data brokers), such inferences could fuel exploitative advertising or other downstream harms. These participant-voiced concerns [733, 219], together with the strong negative coefficient for perceived sensitivity and the consistently low comfort scores, help to explain why participants regard emotional information as an acutely sensitive data type in both settings.

As prior work suggests, privacy concern rises when information heightens vulnerability to harm [559, 737, 573].

Our findings support the formal classification of emotional information as a sensitive category of data. Doing so would require data handlers to apply heightened safeguards [299] aligned with the inference minimization principles we propose above. It would also address persistent concerns expressed by participants about the adequacy of self-regulation in power-imbalanced institutional settings. Sensitivity classification would support regulators in identifying privacy risks and compel both industry and academic practitioners to specify how emotion data is collected, used, and protect.

Finally, such classification would extend urgently needed protections to controversial technologies like facial emotion recognition. Our findings show that the use of facial data consistently heightened discomfort—likely reflecting broader public concerns with facial recognition technologies [922]. Current U.S. regulation typically limits protections to biometric identification [456]. Defining emotional data as sensitive should cover both raw inputs and inferred outputs, closing existing regulatory gaps and better aligning policy with the emotional privacy norms surfaced in our study.

6.6.3 Contextual Vulnerability and Emotional Privacy

Our findings reveal that comfort with emotion inferences varies not only by purpose but also by social position. Although not all socio-demographic effects were statistically significant, several patterns across race, gender, mental health status, and education were illuminating. In both employment and healthcare contexts, Black participants consistently reported higher comfort relative to white participants, with mean comfort levels higher for this group than for any other racial/ethnic category. Similarly, participants without a Bachelor’s degree tended to view emotion AI data flows more favorably across the board, including a substantial and statistically significant effect in employment within the minoritized sample (+6.16) compared to a near-zero effect in the representative sample (+0.14). These patterns suggest that *position-related vulnerability* may heighten recognition of when data flows align with the legitimate social ends of a context—such as promoting wellbeing and support—thereby upholding dignity and fair treatment in the workplace [57] and preserving patient autonomy and dignity in healthcare [859]. Ethical governance of emotional privacy must therefore balance harm prevention with recognition of benefits, particularly for those most vulnerable to harm and exclusion.

At the level of intersecting socio-demographic variables, notable patterns emerged:

- **Trans and/or non-binary participants** reported heightened discomfort, especially toward emotion inferences in healthcare.

- **Participants undergoing treatment for mental illness** judged emotion inferences more positively in healthcare (across both samples), but only in the U.S. representative sample did this translate to the workplace.
- **Participants with untreated or resolved mental illness** expressed more negative judgments in the representative sample, but more positive judgments in the minoritized sample.
- **Asian participants** tended to judge emotion inferences more negatively, especially in the workplace.
- **Black participants and those without a Bachelor's degree** reported consistently higher comfort, significantly so in both employment and healthcare.

We also observed key differences at the belief level. General privacy beliefs, as measured by the Internet Users' Information Privacy Concern (IUIPC) scale, did not significantly predict emotional privacy judgments. Instead, context-specific beliefs (e.g., perceived sensitivity of emotional data and trust in employers or healthcare providers) emerged as decisive predictors. This finding challenges the adequacy of general privacy concern frameworks like IUIPC and underscores the need for research approaches that attend to both contextual and position-based variations.

By identifying how contextual, socio-demographic, and belief-based factors intersect to shape emotional privacy judgments, our findings underscore the importance of designing, applying, and regulating emotion-inference technologies with both contextual and individual sensitivity. While not all observed differences achieved statistical significance—unsurprising given power constraints for some intersecting groups—the patterns nonetheless offer theoretically meaningful insights into how lived experience, privacy vulnerability, and position-based trade-offs shape privacy judgments. Privacy research, too, must move beyond aggregate or nationally representative models to reflect the diverse privacy needs, concerns, and expectations of different people and groups.

Locating risk at the data flow level: a human-centered design implication. Across these patterns, power asymmetries—particularly in the employer/employee and provider/patient relationships—emerged as central to shaping comfort with emotion inferences. Consistent with contextual integrity theory, our findings suggest that emotional privacy concerns are less about the technology itself and more about the institutional contexts and data flows in which it operates. Where emotion inference technologies were perceived as potentially beneficial, participants also voiced concern that institutional power dynamics could undermine agency or lead to harm.

Notably, our findings also revealed important *intra-contextual* distinctions: for example, in healthcare, participants judged emotion inferences for neurological disorders more favorably than

for mental health monitoring, reflecting how purpose functions as a critical—yet often overlooked—determinant of appropriateness even within the same domain. This reinforces our empirical extension of contextual integrity by elevating purpose as a constitutive parameter shaping privacy judgments.

These insights point to a clear human-centered design implication. One strategy for mitigating risks while preserving benefits is to remove or limit data flows that embed institutional power asymmetries. Deploying emotion inference technologies in self-monitoring or closed contexts—where individuals retain control and interpretive agency over data capture, use, and sharing—may help protect privacy and promote autonomy. Prior research on participatory and agency-supportive data practices, such as semi-automated self-monitoring systems, shows how design can balance automation with self-determination [188].

Of course, such deployments are only appropriate where the data flows themselves adhere to *contextually appropriate parameters*, as defined by contextual integrity: including suitable transmission principles (e.g., limits on sharing and retention), clearly justified purposes, and alignment with the social norms and goals of the context. This is especially critical as emotion data increasingly circulates across sectors. Related work on cross-sectoral data sharing highlights both the potential benefits and challenges of such practices, including the need for transparency, consent specificity, and recognition of cohort-based risks [598]—all elements that contextual integrity explicitly requires. Our findings suggest that while cross-sectoral sharing may be acceptable when it demonstrably serves participants’ goals and adheres to trusted contextual parameters (as in some clinical research settings), default sharing of emotion data beyond the original context remains a major source of discomfort and must be carefully governed. Further research is needed to validate what contextual parameters are judged appropriate across diverse groups and use cases, especially in emerging or hybrid contexts where norms are not yet fully established.

Finally, our study itself reflects a human-centered design approach. By systematically analyzing how people’s privacy judgments vary by context, purpose, and social position—and by identifying the specific data flows that drive acceptance or rejection—we demonstrate how empirical, participant-centered methods can inform both technology design and governance.

6.7 Conclusion

Emotion AI technologies introduce unprecedented flows of affective data into domains where privacy, dignity, and wellbeing are at stake. By testing 56 workplace and healthcare scenarios with two demographically differentiated U.S. samples, we show that *contextual, socio-demographic, and privacy belief* factors jointly shape how workers and patients judge the acceptability of those data flows:

1. **Purpose dominates.** Varying the stated aim of an emotion inference produces the largest shifts in comfort. Purposes that reinforce a context’s social mission (e.g., safety in employment, neurological screening in healthcare) raise comfort; purposes that distort those missions (e.g., performance scoring, automated mental health diagnosis) lower it.
2. **Input modality matters.** Facial analytics consistently reduce comfort relative to speech/text, reflecting persistent skepticism toward vision-based emotion recognition.
3. **Position-related vulnerability influences judgments.** Minoritized participants follow the same directional trends as the representative cohort, but with amplified effects—positive and negative—consistent with greater perceived susceptibility to both risks and benefits.
4. **Belief factors are decisive.** Institutional trust raises comfort; perceived sensitivity of emotional information lowers it—often more than recognized sensitive data categories.

Theoretical contribution. Our findings empirically extend contextual integrity by demonstrating that purpose—traditionally treated as implicit—functions as an inter-dependent, constitutive parameter. Elevating purpose clarifies why otherwise similar data flows diverge in perceived appropriateness and provides a tractable lever for governance.

Design and policy implications.

- **Purpose limitation and inference minimization.** Regulation and policy should enumerate narrowly tailored, validated purposes; bar secondary uses; and require necessity proofs before deployment—mirroring risk-based approaches such as the EU AI Act [300].
- **Elevating emotional data protections.** Emotional information, including inferred emotion, warrants protection as a special category of data with heightened safeguards. Given the predictive power of trust in shaping privacy judgments, design and deployment should embed transparency, auditability, and meaningful opt-out rather than rely on institutional goodwill.
- **Individual sensitivity.** Both contextual and individual factors shape emotional privacy judgments. Privacy research, system design, and governance frameworks addressing emotional privacy-intrusive technologies should therefore explicitly attend to varying susceptibilities and diverse needs by upholding the dignity and agency of all data subjects.

Future work. Longitudinal and qualitative research should trace how emotional privacy judgments evolve with repeated exposure to emotion-inference systems, extend these findings to additional high-stakes domains (e.g., education, law enforcement), and engage affected communities

in co-designing technologies that reflect their values, needs, and vulnerabilities. Realizing the potential benefits of emotion AI must not require the indefensible trade-off of sacrificing emotional privacy.

As emotion AI proliferates, its impact on human dignity will depend not only on how sharply we define and enforce the purposes for which emotional data may flow, but also on how effectively we embed vulnerability sensitivity, positional equity, and context-aware design. Purpose-aware extensions to contextual integrity, paired with inference-minimization, sensitivity classification, and participatory design, offer a principled and actionable path forward for researchers, designers, and policymakers.

Part IV: Drawing the Dignity Line in Privacy Theory and Governance

Emotion is among the most intimate dimensions of human life—as highlighted by my empirical findings reported in Parts II and III, an inherently personal phenomenon shaped by one’s values, experiences, and social relations. While broad categories of emotion families like happiness, sadness, anger, fear, disgust, and surprise may be cross-culturally recognizable [287, 289, 292], the meaning of any particular emotional episode is irreducibly individual. Part and parcel of an individual’s unique worldview, as Martha Nussbaum’s theory of emotions explicates, emotions bear moral significance in that they reflect an individual’s evaluative judgment of what personally matters for one’s own flourishing [659]. Reflecting how we interpret our lives and what we care about, emotions bind our inner selves to our personal visions of the good. The act of inferring and acting upon another’s emotions, then, thus carries ethical stakes, implicating questions of respect, agency, and human dignity.

Contextual Integrity (CI) theory holds that privacy violations occur when socially shared information norms are breached. These breaches are often signaled by *intuitive* moral judgments—discomfort, surprise, anger, shame—that reflect judgments of inappropriateness [650]. Accordingly, CI’s vignette methodology captures these intuitions as proxies for normative judgments concerning whether a given flow aligns with context-relative norms of appropriateness [650]. Yet to fully justify a data flow, CI requires a layered normative heuristic: first considering the interests involved, then the benefits and risks and whether they are just by local standards, and finally whether data flows promote the contextual goals of the domain in which they occur and the social ends they serve [649]. Functioning as teleological benchmarks, these contextual purposes give the context its normative structure and social legitimacy [650].

As Nissenbaum has acknowledged, CI faces a challenge when applied to advanced socio-technical systems in which personal information is not disclosed but inferred—generated by computations that analyze disparate, often mundane “data primitives” [650]. Especially in cases of inferred inner states (e.g., emotion, mental status, intention), there is no established social meaning or shared norm by which to evaluate appropriateness. Without either, CI lacks the evaluative foundation to determine when such inferences violate normative boundaries.

Chapter 6 addressed this challenge by extending CI’s framework to accommodate inferential data flows. It measured participants’ comfort with emotion inferences described in vignettes that fixed CI’s five parameters, while introducing two additional contextual variables: *data input* and

purpose. Open-ended responses further contextualized participants' normative judgments. While both input types had consistent effects across populations and contexts, the 14 purposes varied relative to both. Purpose emerged as a decisive factor, shaping how participants interpreted and normatively evaluated novel flows of emotion inferences.

Patterns by purpose tracked to CI's key normative claim: that contextual ends give data flows their meaning, lending both empirical and normative weight to ongoing questions in CI about whether *purpose* should be considered a constitutive parameter [651, 650]. Yet Chapter 6's mixed methodological analysis also illuminated something deeper: a threshold normative expectation that does not depend on informational norms, but rather on whether one's dignity is respected.

Participants' normative privacy judgments responded not only to context-relative expectations, but to a more fundamental expectation to be treated with fairness, recognition, and respect for both role-based agency and inherent worth. While these judgments were context-sensitive—shaped by structural power dynamics and individual differences in privacy vulnerabilities—they were not reducible to context alone. They expressed a normative floor: the expectation that dignity should not be violated, no matter the informational context.

As Nissenbaum presciently noted, “there is a dire need for systemic principles that will expose the material risks of the current data policy anarchy” [650]. The results of this study indicate that one such principle is the cross-cutting norm of basic respect for human dignity. In response, Chapter 7 develops a theoretical extension of CI that incorporates dignity as a shared moral minimum. It formalizes purpose as a sixth constitutive parameter in CI and introduces a dignity-based evaluative threshold based upon Martha Nussbaum’s Capabilities Approach [658]. The here proposed Capabilities Approach–Contextual Integrity (CA–CI) model retains CI’s core descriptive and normative structures while specifying the conditions required to ensure that data flows respect contextual norms and ends *and* human dignity.

CHAPTER 7

Inviolate Personhood: A Capabilities–Contextual Integrity Approach to Privacy and Dignity in AI

7.1 Introduction

Across its many formulations, privacy has served as a moral defense against domination by external forces. Whether conceptualized as restricted access [42], solitude [62], control [883], boundary management [44], or contextual appropriateness [646], each of these forms resists the imposition of norms, meanings, or expectations that are not freely shared. From Warren and Brandeis' invocation of the *inviolate personality* to Nissenbaum's Walzerian defense of contextual integrity [646, 871], privacy theory has long grappled with the question of how to protect the inner life and the social fabric that sustains it. This chapter foregrounds privacy's historical architecture as a bulwark against domination, showing how thinkers across centuries and disciplines converge on a common insight: that privacy is inseparable from human dignity and essential to sustaining moral personhood.

Yet not all claims to privacy are claims of dignity. In an age increasingly structured by socio-technical systems, the ethical challenge is not simply to protect privacy, but to discern when data flows are appropriate and when they cross normative lines. Helen Nissenbaum's theory of privacy as Contextual Integrity (CI) addresses this challenge through a “justificatory framework” that evaluates data flows relative to a context's informational norms and teleological ends [646, 650]. However, CI's reliance on socially shared norms presents limitations: it struggles to assess novel data flows, such as personal inferences, that lack either settled meaning or precedent [650]. Moreover, while CI clearly specifies the constitutive components of a data flow, it leaves the evaluation of those flows to local standards—lacking an external basis for assessing when those standards themselves are legitimate.

This chapter bridges privacy's normative foundations in human dignity with the normative architecture of CI. I argue that human dignity functions as a shared basic norm—one with cross-cultural moral traction—and can be integrated into CI's framework with fidelity to its commitments to values pluralism. Drawing on CI's Walzerian roots, I contend that dignity provides the most

coherent candidate for what Walzer calls a “moral minimum”: a universal normative floor required to preserve the integrity of social domains and the plural local values they sustain [646, 649, 650, 872].

To operationalize this extension, I draw on Martha Nussbaum’s Capabilities Approach, which defines the material, social, and psychological conditions necessary for a life with dignity [658]. Her framework identifies ten core capabilities such as bodily integrity, practical reason, emotions, and affiliation as threshold requirements for human dignity. In the Capabilities–Contextual Integrity (CA–CI) model I propose, a data flow is judged inappropriate when it foreseeably undermines an individual’s ability to develop or exercise one or more of these core capabilities. This reorients CI’s layered normative analysis from a purely context-bound justification to one grounded in the universal floor of human dignity: appropriateness ends where dignity is violated.

The CA–CI model thus evaluates data flows on two levels: first, by their fit with contextual norms and ends (CI), and second, by whether they uphold the basic entitlements required for dignity (CA). This allows CA–CI to identify morally inappropriate flows even where contextual expectations are permissive, unsettled, or contested—drawing a principled line in the sand where none previously existed.

To demonstrate its practical value, I apply CA–CI to three case studies involving AI systems and emotionally inferential data flows. These analyses illustrate how the model provides context-sensitive, normatively grounded, and operationally tractable guidance for the ethical evaluation, regulation, and design of socio-technical systems. By integrating the descriptive precision of contextual integrity with the normative architecture of the Capabilities Approach, CA–CI offers a principled framework for restoring privacy’s foundational role in sustaining dignity in the age of AI.

7.2 Background and Related Work

Necessary precisely because we exist as social creatures, privacy has long hovered at the boundaries between the self and society [618, 564], the private and the public [385, 741, 62], and the internal and external domains of personhood [44, 883, 39, 199, 908, 554]. While early privacy theories emphasized its physical and dispositional dimensions—shielding bodies, homes, and inner life from external intrusion—the rise of digital technologies has shifted attention toward *informational privacy* [39, 199, 742, 883]: the ethics and governance of personal information flows.

Informational privacy has been understood as both an instrumental and an intrinsic good. Instrumental approaches emphasize its contingent benefits—for liberal democracy [883], personal well-being and development [353], and civil society [630]. Intrinsic accounts establish privacy as a first-order moral and political value: one that justifies constraints on surveillance, exposure,

and coercive conformity [200, 876, 39], even in the absence of measurable harm [41]. Yet overwhelmingly, privacy governance continues to treat it as instrumental.

This section traces this history. I begin by tracing the legal and moral foundations of a general right to privacy in U.S. law, where privacy was once valued intrinsically for its protection of the “inviolate personality” [876]. I then discuss how this foundation eroded under the influence of instrumentalist reasoning in privacy law and governance—undermining the very dignity privacy was meant to safeguard, which now faces fragmentation across domains and governance regimes, and our capacity to recognize and address the pressing privacy harms of the present.

7.2.1 The Moral Origins of Privacy

7.2.1.1 The Fate of the Inviolate Personality in a General Right to Privacy

The emergence of informational privacy as a distinct moral and legal concern in the United States can be traced to public anxiety over involuntary informational exposure, precipitated by the introduction of the Kodak camera in the late nineteenth century [742, 721, 193]. In response, Samuel Warren and Louis Brandeis published their landmark 1890 article, “The Right to Privacy,” which proposed a general right to privacy grounded in the principle of the *inviolate personality* [876]. Central to their argument was concern about the unauthorized circulation of individuals’ thoughts, sentiments, and emotions via photographic capture and sharing—intrusions upon the person that, in their view, threatened the spiritual integrity of the self [739]. Warren and Brandeis begin with a narrative of common law’s expansion:

“In very early times, the law gave a remedy only for physical interference with life and property...Later, there came a recognition of man’s spiritual nature, of his feelings and his intellect.”

They trace how protections that once guarded only the physical body gradually extended to cover dispositional aspects of the self: reputation, emotional life, affiliations, and inner thought. Property law, once concerned with tangible assets, had also expanded to encompass “the wide realm of the intangible”—letters, sketches, and even the unexpressed “thoughts, emotions, and sensations” that animate such artifacts. With a teleological quality, Warren and Brandeis suggest that common law had been evolving toward fuller recognition of the person’s psychic and moral integrity.

By 1890, however, the advent of Kodak cameras and an increasingly sensationalist press had, they warned, “invaded the sacred precincts of private and domestic life.” They condemned the rise of a commercialized culture of gossip that subjected individuals “to mental pain and distress far greater than could be inflicted by mere bodily injury.” The wrong, as they saw it, was not merely reputational or proprietary but spiritual: a “blighting influence” that at once degraded the dignity

of the individual and corroded social compassion by “dwarfing the thoughts and aspirations of a people.”

To support their claim, Warren and Brandeis cited legal precedents in which courts had already restrained the publication of letters, diaries, and the like—not to protect the content as intellectual property, but because such writing reflected the individual’s emotional life. The relevant harm, they emphasized, was the exposure and circulation of “facts relating to life, feelings, and emotions,” which belong only to the person who originated them. From these precedents they drew a general principle:

The protection afforded to thoughts, sentiments, and emotions...is in reality not the principle of private property, but that of an inviolate personality.

This right to privacy, as they understood it, was medium-independent. No test of form, artistic merit, or communicative mode—be it facial expression, pantomime, sonata, or diary—could delimit the right to determine “to what extent his thoughts, sentiments, and emotions shall be communicated to others.” Absent legal compulsion, that decision remained with the individual. Even after disclosure, the right persists: one “retains the power to fix the limits of publicity.” Serving as shorthand for this broader principle, Judge Cooley’s phrase “the right to be let alone” was soon taken up in the broader privacy discourse, flattening the deeper moral grounding Warren and Brandeis articulated.

To fully appreciate the normative force of Warren and Brandeis’s claim requires recovering the socio-cultural context in which the phrase “inviolate personality” would have resonated. Far from a mere defense of decorum, it invoked a century’s worth of anxiety about the effects of industrialization, urbanization, and mass society on individual coherence and authenticity. The term evoked a morally autonomous person—capable of reflection, emotional depth, and self-direction—whose formation depended on protected spaces insulated from surveillance, coercion, and the flattening pressures of conformity to dominant social norms [739]. Privacy, in this account, is both defensive and generative: a necessary precondition for the development of the moral self.

To invoke the *inviolate personality* as the principled heart of privacy is to assert that the inner life must be treated as sacrosanct—both for the sake of the individual and because its protection is constitutive of a free and flourishing society. This understanding of privacy as an *intrinsic* moral and political value in its own right—a condition necessary for and constitutive of the formation, maintenance, and protection of the moral self and derivatively, society—draws from the literary and philosophical currents of the late eighteenth and nineteenth centuries, which mounted sharp critiques of social tyranny, scientific rationalism, and social constructionism as forces that constrained self-development and hollowed out the conditions for moral autonomy [890]. Warren and Brandeis, in this spirit, argued that privacy’s protection extended even against the state itself—

relying on common law reasoning to identify not only legal wrongs, but moral affronts to human dignity, especially when injustice arose from the very community values that claimed to define what was socially acceptable [739, 640].

By the late nineteenth century, privacy theory increasingly recognized social coercion as a distinctive source of harm. Against this backdrop, privacy—understood as solitude, tranquility, or withdrawal from public scrutiny—was positioned as a vital safeguard. The “inviolate personality” emerged from this discourse as a moral and cultural ideal that marked a significant shift: privacy was no longer viewed as merely a counterbalance that improved well-being in modern society, but as a foundational condition of human life itself [739]. Poets like Wordsworth cast solitude as necessary for preserving the inner self against the overstimulation and moral conformity of modern life [890]. Philosophical accounts echoed this emphasis. In *On Liberty*, John Stuart Mill warned of the “tyranny of the prevailing opinion,” describing it as:

more formidable than many kinds of political oppression...leaving fewer means of escape, penetrating much more deeply into the details of life, and enslaving the soul itself [609].

The concept of the *inviolate personality* should be situated within this broader tradition—a shared concern with preserving the inner life as the seat of moral agency and capacity for human flourishing. This concern arose in part as a response to the French Revolution, whose bloody aftermath exposed the dangers of both monarchical repression and revolutionary excess. In its wake, Romantic and liberal thinkers developed a new vocabulary of interiority—resisting both state surveillance and the moral absolutism of revolutionary ideology [890]. Resolving contradiction in prevailing views of the individual as at once a mere bearer of abstract rights and a passively constructed product of social forces [739], these literary and philosophical works came to understand the person as custodian of an inner domain—conscience, feeling, judgment, spirit—that must remain inviolable if moral agency is to be preserved. It is this vision that Warren and Brandeis institutionalized in legal form: a moral right to privacy rooted in the dignity of the person, designed to safeguard the emotional, dispositional, and expressive core of selfhood. Their formulation provided a coherent normative foundation for the legal protection of individual integrity and autonomy—what Bloustein later called the “essence of a unique and self-determining being” [128].

Warren and Brandeis’ argument for a general right to privacy endures in legal doctrine. While their vision has shaped generations of legal and cultural discourse, it has also been persistently misunderstood. One enduring myth, popularized by Prosser and repeated in subsequent scholarship (e.g., see [438, 721, 467, 127, 723]), claims that the authors were spurred to action by a newspaper’s publication of wedding photographs from a prominent Boston family—a claim taken up by criticism of a general right to privacy as protection of privileged society. The irony, as Rosen and Santesso

note, is that this apocryphal story centers precisely the kind of social event—wedding, community, spectacle—that Warren and Brandeis sought to distinguish privacy from. Their concern was not the sentimental management of social appearances, but the preservation of the inner self as a domain of moral significance [739]. That this myth endures reveals how easily privacy’s spiritual and ethical dimensions can be reduced to questions of taste, sensitivity, or celebrity control [630]—obscuring the deeper claims about personhood and dignity that remain just as vital, yet overlooked, in the overt social forces of surveillance and exposure of online life today [740].

Shortly after their article’s publication, courts widely adopted privacy torts, recognizing their role in protecting the *inviolate personality* under the broader principle of the right to be “let alone” [192]. Yet the ensuing period of “vigorous growth and experimentation” in privacy tort law [723] was eventually blunted by a narrowing of scope [739]: privacy came to be understood less as a moral right grounded in personhood than as a cluster of interests in property, utility, or control. What was lost in the process was the foundational insight Warren and Brandeis had so carefully developed: that privacy protects, and enables, the moral architecture of the self. In eclipsing the dignity-based imperative to preserve the inner life, law abandoned the deeper vision that once gave privacy its normative force—a vision urgently in need of recovery today.

7.2.1.2 Doctrinal Narrowing and the Rise of Instrumental Privacy

In tracing the legacy of Warren and Brandeis’ “The Right to Privacy,” Rosen and Santesso argue that its foundational commitment to protecting the dignity of the self—the *inviolate personality*—was eclipsed by the narrower, conceptually impoverished account of privacy institutionalized through William Prosser’s tort taxonomy [739]. By fragmenting privacy into discrete, compensable harms, Prosser’s framework severed privacy from its ontological grounding in selfhood and as a result, failed to supply the normative coherence required to recognize the moral status of privacy as a first-order social good [276].

The decisive shift came in 1960, when Prosser, aiming to formalize and stabilize the emergent tort doctrine [192], codified four privacy torts: intrusion upon seclusion, public disclosure of private facts, false light, and appropriation of name or likeness [702]. While his taxonomy stabilized privacy’s legal standing, it did so by recasting privacy as a series of injuries to be balanced under utilitarian logic. Prosser’s ambition was not to elevate privacy’s normative status, but to suppress it—as Citron explains, to render privacy compatible with the internal logic of tort law, protecting individual interests only when harm could be demonstrated and weighed against social utility [192]. This fit well with Holmesian jurisprudence, which rejected law’s moral aspirations in favor of optimizing social behavior [422].

Prosser’s project stabilized privacy’s place in law, but only by recasting it in instrumental and remedial terms. Richards and Solove describe this moment as both a victory and a loss: while

privacy gained doctrinal traction, its moral depth was undercut [723]. The *right to be let alone*, which Warren and Brandeis had envisioned as a dignitary shield protecting the emotional, spiritual, and dispositional core of personhood, was reduced to a checklist of harms untethered from that deeper normative vision [192]. Prosser's categories made no reference to the *inviolate personality*; they offered courts actionable compensable categories, not philosophical grounding [739].

Prosser's influence did more than shift doctrine; it reshaped the intellectual terrain of privacy theory itself—narrowing the conceptual space available to recognize privacy as morally fundamental. It became increasingly plausible for figures like Judith Jarvis Thomson to argue that privacy was reducible to property and emotional distress claims [834], and for Alan Westin to redefine privacy as control over information flows [883]—privacy as a *functional*, rather than intrinsic, value became entrenched. Ronald Dworkin diagnosed this shift a “brilliant fraud”—a formal success that failed to offer any substantive justification for privacy as a legal or moral right [276]. The result was a framework in which privacy is treated as valuable for what it enables—autonomy, trust, democratic participation—but not for what it inherently constitutes and protects: the dignity and integrity of the moral self.

This legacy persists. U.S. privacy jurisprudence still struggles to recognize dignitary privacy harms [194]. For instance, though emotional harm is acknowledged in principle, it remains difficult to prove due to its “ethereal nature” and the doctrinal preference for tangible, measurable injuries [192]. Meanwhile, data-driven profiling, AI inference, and ambient surveillance systems expose individuals to precisely those harms Warren and Brandeis warned against: violations of emotional, dispositional, and cognitive integrity without observable damage, codified safeguards, or actionable recourse.

The doctrinal narrowing of privacy law is a theoretical reduction with real consequences. By disaggregating privacy and filtering it through a utilitarian calculus, courts and governance systems have lost the ability to detect and remedy the dignity harms that persist when privacy loses its original anchoring in the dignity of the self. As I argue, reconstructing privacy's moral status is not a nostalgic appeal to the past, but a forward-looking necessity: a foundation for designing regulatory and socio-technical frameworks that can recognize when dignitary boundaries are crossed, and why such crossings matter for human flourishing.

7.2.2 Instrumental Privacy Approaches and Fragmentation

The most influential accounts of privacy in modern liberal thought have emphasized its instrumental value: as a mechanism to promote individual, relational, and societal goods such as autonomy, intimacy, and democratic participation. But this prevailing view fails to capture the moral stakes of privacy violations that breach deeper norms of respect, recognition, and personhood. This section

reconstructs this privacy tradition and highlights its limits, motivating the need for a dignity-based account.

Alan Westin's *Privacy and Freedom* launched a dominant strand of privacy thought that remains influential today. Westin framed privacy as a functional mechanism for promoting liberal democratic values, grounded in control over solitude, intimacy, anonymity, and reserve [883]. On this view, privacy enables emotional release, self-evaluation, and the maintenance of differentiated social roles. It protects not only personal dignity but the institutions—family, civil society, and deliberative democracy—on which liberal societies depend.

Ruth Gavison builds on this tradition, emphasizing privacy's contributions to autonomy, mental health, creativity, and intimate relationships [352, 353]. She defines privacy in terms of secrecy, solitude, and anonymity, and argues that it reduces pressure to conform, providing the space for moral reflection and imaginative life. Gavison highlights privacy's regulatory function: by controlling social visibility, privacy shields individuals from surveillance, judgment, censure, and coercive conformity, enabling political participation through practices like the secret ballot.

Thomas Nagel adds a civilizational dimension to the functional defense of privacy, arguing that conventions of concealment—including secrecy, reticence, and nonacknowledgment—are essential to social cooperation and psychological stability [630]. Privacy, on this account, involves boundaries between what information gets exposed and what does not, serving a vital function in managing the “sheer chaotic tropical luxuriance of the inner life” and sustaining a public-facing self capable of functioning in shared social space. Civil society, he contends, requires restraint not only in law but in social conventions—and the erosion of privacy norms, amplified by novel technologies and media (a critical throwback to Warren and Brandeis), risks overexposing individuals to emotional trauma and destabilizing interpersonal relations. In Nagel's view, privacy is functionally indispensable: it enables the free operation of personal feeling, fantasy, and thought by shielding individuals from the disorienting effects of total exposure.

As Anita Allen notes, functionalist accounts generally fail to engage the moral status of the ends privacy is said to serve. “Functionalist underemphasizes the close and special connections moralists have stressed between and among privacy, personhood, and fitness for social participation and contribution” [39]. If privacy is valuable only because it is an instrument of other interests, what anchors privacy's instrumental value normatively? Without independent moral weight, privacy is fungible, justifiably traded away for higher-order values or when the ends it serves can be achieved by other means.

7.2.3 Contextual Integrity and the Pluralist Turn

And yet, such reasoning persistently fails to explain our intuitive judgments that some privacy violations, even when claimed to be justified—surveilling employee emotions to serve corporate productivity goals, for instance [735]—are just *wrong*. When privacy is reduced to a set of beneficial, instrumental functions, its moral core gets obscured. The challenge, then, is not to reject privacy’s instrumental value, but to also recognize when its violation breaches deeper normative expectations.

Privacy, under CI, is neither absolute nor singular; it emerges when data exchanges comfortably conform to established roles, attributes, and transmission principles that govern the acceptability of information flows in each social domain. Importantly, CI is both descriptive and normative. It models how privacy operates in practice, but also offers a justificatory test: a data flow is *prima facie* permissible if it conforms to the entrenched informational norms of the context. When norms are disrupted—by novel technologies, shifting power relations, or new risks—CI calls for evaluating whether those norms themselves remain legitimate based on appeals to the social value of preserving the purposes, ends, and functions of the domain [646].

CI’s key practical appeal is its ability to adjudicate privacy claims not solely on individual terms, but on these bases—in relation to the norms and purposes that define a social domain—a “justificatory” framework designed to reason through conceptual confusion and practical chaos, where individual claims exercising an abstract “right” to privacy compete with the modern socio-technical necessity of personal data flows [646]. CI’s normative emphasis on preserving contextual norms pragmatically instrumentalizes privacy as a means to secure human values across complex social realms [649]. Yet it does not specify *which* human values must be upheld, relying on the assumption that fair social processes have shaped which norms are considered appropriate across competing interests within a context over time [646].

CI introduced a major improvement to privacy theory, research, and governance, shifting the discourse to respect for the *context* of privacy at a time where the dominant conceptualization—privacy as a binary along the public/private divide—persistently failed to either identify privacy violations or justify information flows occurring in contexts the prevailing view considered as “public” but in which individuals still expected privacy [649]. CI provided an explanatory and predictive framework that showed those expectations are constituted by CI’s five contextual parameters and governed relative to the context’s established norms and broader purposes.

But as Nissenbaum contends, CI’s original framework struggles to justify data flows in two respects: first, where the “tyranny of the normal” fosters social acceptance (e.g., through habituation, resignation) of information flows that are harmful or misaligned with shared social values; and second, where novel data flows emerge before the social negotiation through which norms are typically set and against which their appropriateness can be judged [646, 650]. In both cases,

CI's reliance on local standards to articulate privacy's value leaves the theory normatively under-powered. This challenge—the normative dependence on local informational norms that may be absent, unstable, or themselves unjust—has become more pressing with the rise of large-scale data infrastructures and inference-based systems that aggregate, link, and mine seemingly mundane, “primitive” data across disparate sources, re-contextualizing it to re-identify individuals, construct detailed behavioral profiles, and infer increasingly intimate information about their inner states, beliefs, and traits [650].

In modern digital environments, people are becoming habituated to ubiquitous privacy intrusions that erode their capacities for agency and dignity. Individuals are inappropriately subjected to increasingly intrusive novel data practices such as affective profiling, targeted behavioral nudging, and pre-conscious manipulation [928, 818]. The social mechanisms through which norms are supposed to evolve like public deliberation, reciprocal engagement, and democratic accountability are displaced by opaque, privately governed infrastructures [56, 848]. Whether operating as employers with authoritarian degrees of control over workers' private lives [56] or providers of exploitative data-intensive consumer products [848], technology giants and the broader data ecosystems they enable wield disproportionate socio-technical power to scale autonomous systems that influence moods, beliefs, and behaviors *even pre-consciously* [928, 818]. Opportunities for meaningful participation in shaping norms continue to diminish [590], as commercial interests increasingly displace valued social norms [753, 754]. As philosopher Michael Sandel observes, market norms, unfit to address normative problems, increasingly displace social norms across nearly all shades of life [754, 753]. These information-enabled power imbalances distort the very conditions under which norms form, undermining the integrity of the very social domains where individuals should have real opportunities to negotiate their fundamental entitlements. And without a global appeal to determine when such dynamics breach a universal moral minimum threshold, violations of human dignity persist without adequate grounds for contestation.

As these practices escape meaningful social participation, become normalized, and thereby increasingly difficult to contest, the dilemmas facing CI's justificatory framework deepen. Without an external normative standard, CI is ill-equipped to challenge data flows that cross contextual boundaries and encroach upon moral terrain where instrumental contextual norms offer no clear defense. In response, this section returns to CI's roots in Walzerian local standards of justice to propose an external evaluative mechanism for identifying inappropriate information flows—one that does not rely upon malleable or contested local norms alone, but also appeals to a more basic human principle: the shared expectation that we ought to treat one another, and expect to be treated, in accordance with our inherent human dignity.

7.2.3.1 Contextual Integrity's Values Pluralism

The limits of normative appeals to privacy that limit its justification solely to its role as an instrument for promoting broader social goods are present in CI, which identifies privacy violations as disruptions to the continuity of lived traditions and norms that people deeply value. CI contributes a methodology for diagnosing when data flows violate these established norms. But what gives those norms their moral force?

CI grounds its normative claims to privacy in a tradition-sensitive framework. Skeptical of abstract and ahistorical moral reasoning, CI locates the meaning and value of privacy in everyday life. This orientation draws on Burkean conservatism, which favors historical continuity and socially embedded norms in times of socio-technological disruption and upheaval [509, 158, 646], and on Walzerian local standards of justice, which uphold “complex equality” by delegating the authority to evaluate and distribute social goods to the communities that co-constitute their meaning [646, 871]. By granting presumptive legitimacy to established information flows, CI conserves the locally-determined meaning of privacy relative to a context, rooting judgments of appropriateness in the normative grammar of lived experience and the values that sustain a social domain’s moral and structural coherence [649].

One of CI’s key contributions is its ability to adjudicate privacy claims not simply on individual grounds but in relation to contextual norms and the broader social purposes of the domain [646]. Yet why do some violations feel like a betrayal, not just a mismatch?

CI’s conceptual power lies in its refusal to treat privacy as static or universal. By rooting privacy’s value in the lived social meaning of information flows, it offers a pluralistic and adaptable model that ensures normative flexibility across diverse socio-technical contexts. Yet CI’s strength here is also its limitation: it does not specify which values are non-negotiable. Yet as the following sections show, treating privacy solely as an instrument of local normative order leaves those very orders exposed. Without a shared moral threshold, the values that contextual integrity seeks to conserve can themselves be overwritten, hollowed out, or co-opted by external forces.

7.2.3.2 Walzer’s Defense of Local Normative Evaluation

CI’s normative architecture draws from Michael Walzer’s values pluralism, which conceptualizes a just society as composed of multiple autonomous social spheres, each governed by its own principles for distributing social goods and determining merit [871]. Within each sphere, the value of a good is co-constituted by a shared understanding of its meaning among members. Justice, in this view, consists in preserving the moral boundaries *between* spheres—conserving local meaning and evaluative standards by ensuring that the logic of one domain is not unjustly imposed on another, thereby distorting the shared meaning of social goods and their distribution. Injustice occurs when

distributive logics illegitimately transpose the meaning of a social good across spheres—when, for example, money grants access to education, or media exposure confers political power.

CI maintains a deep structural resonance with Walzer’s pluralism [646, 650], grounding normative privacy claims in the integrity of social spheres such as health, education, and work [871]. Just as Walzer insists on respecting the moral autonomy of social spheres, CI insists that informational norms ought to align with the values internal to each context, holding that privacy violations occur when data flows contravene contextual privacy norms or the telos of the domain in which it was shared [649]. In CI’s framework, information is treated as a socially situated good, with determinations of its appropriate flow governed by norms that both reflect and sustain the function, meaning, and internal integrity of the context in which it originates. When data flows violate these contextual meanings—for example, when the logics of surveillance, commerce, or bureaucratic rationality override context-relative privacy norms—CI identifies such flows as violations of contextual integrity—and, by extension, as incursions that threaten the preservation of “complex equality” by enabling tyranny and domination through the external imposition of unshared norms [646].

CI’s values-pluralist foundation allows it to assess the legitimacy of information practices by reference to the justice criteria embedded within discrete social domains. In doing so, it offers a powerful framework for defending informational privacy as a form of justice—one that resists both reductive moral universalism and the tyrannical imposition of normative standards. Yet the very strength of this model reveals a structural limitation: the absence of a shared evaluative baseline *across contexts*. CI’s values-pluralist approach faces a well-known challenge: if contextual norms are illegitimately shaped—by power asymmetries, commercial pressures, or historical exclusions—on what basis can we deem a flow *appropriate*? Empirically, CI accommodates measurement via people’s intuitive privacy judgments. But normatively, it defers to those same social norms—even when those norms reflect institutional capture, manipulation, or coercive logics [590, 753]. In such cases, CI lacks a principled method to contest or override illegitimately set norms, such as norms themselves shaped by normative tyranny and domination.

7.2.3.3 Defending Pluralism with Universal Moral Minimums

Walzer, too, recognized the challenge of conserving local norms and their shared social meanings in the face of increasing external intrusions—particularly those that impose evaluative judgments or standards of justice unshared by the local community. In *Thick and Thin*, he further develops his original position in *Spheres of Justice* to defend the necessity of a *universal moral minimum*: a baseline standard that safeguards values pluralism itself. Without a shared standard of *moral minimum* expectations, Walzer warns, local norms and values remain fragile—vulnerable to erosion, displacement, or erasure.

“Cultural pluralism is a maximalist idea, the product of a thickly developed liberal politics. Minimalism depends on something less: most simply, perhaps, on the fact that we have moral expectations about the behavior not only of our fellows but of strangers too...Though we have different histories, we have common experiences and, sometimes, common responses, and out of these we fashion, as needed, the moral minimum” [872]

Walzer’s revised view makes room for a principled resolution. While justice is always interpreted through particular histories and social meanings, he affirms that “*there is no escape from the relativism of distance and difference, but there is also no escape from the universalism of the human condition*” [872]. A moral minimum grounded in shared human experiences—not bound by local meaning—can establish “*a common moral horizon*” [872], one capable of anchoring a baseline expectation to which all contexts, from the most particularistic local communities to the most powerful global domains, ought to be held to account.

Importantly, Walzer’s minimalism remains consistent with CI’s normative tradition. Like CI, it draws strength not from philosophical abstraction but from lived consensus, emerging through “moral intuition and historical experience” [871]. It is grounded in our common capacity to recognize instances of wrongdoing—such as deceit, coercion, oppression—as *wrong* across socio-cultural boundaries. By differentiating between “thick” moral traditions and “thin” shared moral expectations, Walzer advocates moral minimums not as constraints on pluralism, but as its condition of possibility: thin, durable constraints beneath which no practice or norm can justifiably fall. Reinforcing the commitments of “complex equality” and local normative determinations articulated in *Spheres of Justice*, moral minimum standards ensure that no domain’s values illegitimately override another’s, protecting the very autonomy that pluralism requires. As Walzer writes,

“*By its very thinness, it justifies us in returning to the thickness that is our own. The morality in which the moral minimum is embedded, and from which it can only temporarily be abstracted, is the only full-blooded morality we can ever have. In some sense, the minimum has to be there, but once it is there, the rest is free*” [872].

As it stands, CI lacks this thin foundation. Its layered evaluation framework remains agnostic to whether a data flow is ultimately just, deferring only to local standards of meaning and value—even as those standards grow increasingly fragile, and vulnerable to distortion or erasure. To defend both the continuity and moral force of local privacy evaluations, strengthening CI’s normative equipment with a globally applicable moral minimum standard would provide the necessary baseline to preserve pluralism while protecting against its most subtle and dangerous form of failure—domination disguised as appropriateness.

All together, the background work reviewed here raises three central questions:

1. Can privacy's normative roots in dignity help recover its moral force in procedural privacy frameworks amid instrumental fragmentation?
2. Can extending contextual integrity with a universal moral minimum more clearly distinguish just from unjust data flows?
3. How can such a minimum be identified, given the wide variation in what different societies and contexts consider appropriate in the flow of personal information?

7.3 Dignity as a Moral Minimum Standard in Contextual Integrity

This section proposes *human dignity* as a normative minimum threshold to answer these questions. It synthesizes relevant scholarship to argue that dignity offers the most coherent and defensible specification of the moral minimum standard required to extend CI. Grounded in international human rights law, cross-cultural ethical reasoning, and longstanding privacy theory, I show that dignity already functions as both a *de jure* and *de facto* moral boundary recognized across diverse legal and cultural contexts—a shared norm insisting that, at minimum, we ought to treat others, and be treated as, human beings with inherent worth and dignity.

I proceed to make this theoretical-methodological intervention for CI in two parts. First, Section 7.3.1 defends human dignity as a moral minimum standard in CI by drawing on literature that surveys the global ethical and legal consensus around dignity as the foundational principle of human rights, examines its role in legal reasoning about privacy, and theorizes privacy as both constitutive of, and necessary for, moral personhood. Together, these analytically distinct claims support a well-grounded argument for treating human dignity as the moral minimum standard within CI—establishing dignity both as a basic norm with global legal, moral, and cultural consensus, and as a deeper normative anchor that links privacy judgments to the intrinsic value of privacy as essential to moral personhood and human dignity.

Second, Section 7.3.2 introduces a theoretical model that operationalizes dignity as a moral minimum in CI, drawing on Nussbaum's capabilities approach to define human dignity, assess when it is violated, and incorporate its minimum thresholds as fixed evaluative parameters into CI's framework [657].

7.3.1 The Basic Norm of Human Dignity

7.3.1.1 Global Consensus on Human Dignity

Although Walzer refrained from specifying a concrete universal moral minimum, he pointed to the global human rights tradition as a practical articulation of it: a body of shared moral expectations that function not as idealized maxima, but as minimal safeguards against their erosion [872]. However conceptually imperfect, these baseline commitments draw upon a widely recognized normative foundation: the intrinsic dignity of the human person.

Human dignity is a substantive normative concept articulated in legal instruments across the globe. It frequently serves as both a proxy for respect for autonomy and the conceptual basis for fundamental rights and freedoms articulated in national constitutions and international human rights agreements—the “common ground” where local and global interests converge [638, 757]. As the normative foundation of human rights, dignity reflects substantial cross-cultural ethical consensus, exemplified in foundational instruments such as the Universal Declaration of Human Rights [65].

Beyond the legal domain, dignity functions as a cross-disciplinary moral vocabulary: a conceptual framework for diagnosing and prescribing ethical obligations across philosophy, psychology, religious ethics, and public policy [583]. Dignity has been described as the ontological root of privacy [329] and a normative lens for interpreting subjective experience in value-sensitive policy-making [513]. Acknowledging the conceptual disarray surrounding the term, Mattson and Clark call for a model of dignity that is both action-guiding and non-imperial—one that avoids over- or under-definition in ways that risk suppressing local moral traditions. They propose a relational, value-based model of dignity as a condition co-produced through shared values such as respect, power, affection, and well-being [583].

For instance, within the European Union, fundamental rights and freedoms—including the rights to privacy and data protection—are grounded in the universal value of human dignity, which is deemed inviolable in both EU and international law (e.g., see Art. 1 of The Charter of Fundamental Rights of the European Union [303], Universal Declaration of Human Rights [65]). A foundational value of the European Union, human dignity “must be respected, protected and constitutes the real basis of fundamental rights [852].” Accordingly, while CI does not specify human dignity as a value to conserve, application of its normative heuristic within the EU would be interpreted through a dignity-based lens: its second layer, concerned with context-relative moral and political values, must remain proportionate to the balancing of rights derived from the inviolability of human dignity to remain compatible with this foundational EU value [304].

As the European Data Protection Supervisor has emphasized, in today’s global digital infrastructures, the right to personal data protection—with the value of human dignity at its core—plays

an increasingly vital role in preserving the conditions for a free and flourishing life without undue coercion [707]. This is especially true for those subject to structural vulnerabilities, including children, patients, and workers navigating power asymmetries, where even routine data practices may reinforce or exploit conditions of dependency or constraint [558]. In modern socio-technical environments, a right to data protection is thus invoked not merely to shield personal information, but to safeguard the normative preconditions of dignity itself. The logic behind this theoretical abstraction is that exercising a right to data protection can serve to prevent the normalization of novel data practices that would otherwise erode fundamental rights by stealth. Given formalized legitimacy, claims to data protection can forestall the entrenchment of novel data practices that transcend regional boundaries and, if left unchecked, risk eroding the intrinsic and universal value of human dignity through constrained choice architectures that habituate individuals to indifference by design—practices that escape scrutiny not by normative legitimacy, but by social routinization. In practice, however, this vision remains difficult to realize. Particularly in digital spaces, exercising one’s entitlement to data protection proves increasingly challenging as globalized markets erode [209, 664] and corrode [754] fundamental rights and freedoms.

The challenge is made greater by the shifting, incompatible rhetoric surrounding abstract concepts like *privacy*, *data protection*, and *dignity*: without shared analytic clarity about their conceptual scope, how they are constituted, and when their claims are (and are not) legitimate the public sphere—people, courts—will continue to struggle to transform these entitlements into any substantive claims or safeguards against competing interests [712, 649, 793, 549].

CI offers rare analytic clarity in this landscape. Its structured approach to defining contextual norms makes it a powerful framework for identifying when privacy is violated and for specifying how data should be protected across information systems and regulatory instruments alike [110, 96]. Yet CI remains descriptively anchored: it maps internal normative expectations about what is considered appropriate within a given context, but lacks an external evaluative mechanism to assess whether those norms are themselves consistent with foundational moral commitments like human dignity. Embedding an explicit dignity threshold within CI would supply such a standard, enhancing CI’s responsiveness to contemporary digital risks and ability to meet the ethical demands of global information systems. The challenge, however, lies in operationalizing such a threshold without compromising the analytic precision that gives CI its distinctive power.

7.3.1.2 The Intuitive Logic of Dignity in Privacy Judgments

Although contextual integrity does not endorse a particular philosophical theory of privacy (e.g., control, restricted access), it provides a framework that can accommodate each. As Nissenbaum writes, “the framework of contextual integrity reveals why we do not need to choose between them; instead, it recognizes a place for each” [646]. This pluralist openness, however, raises an important

normative question: which values should anchor privacy as a justified defense when local consensus is absent or compromised?

Legal and philosophical accounts of privacy violations offer critical insight here. Specifically, dignity-based theories help illuminate why certain privacy invasions remain morally troubling even when no harm is experienced or intended. What these theories reveal is that a normatively grounded claim to privacy rests not only on protection from injury or control over information, but on a deeper commitment to mutual recognition and moral respect.

The tension between instrumental and intrinsic articulations of privacy's value is brought into sharp relief by James Moor's well-known thought experiment, which challenges accounts that justify privacy solely through instrumental values like autonomy. In this scenario, a person is continuously surveilled by "Tom the eavesdropper," but never detects the surveillance and experiences no harm [617]. Moor concludes that since no tangible interference occurs, autonomy remains undisturbed, and thus privacy cannot be necessary for autonomy. However, he concedes that something remains intuitively wrong with Tom's actions *intrinsically*.

To explain this, Moor advocates a "core value framework" that treats privacy as both instrumental and, in some cases, intrinsic—limited to where it expresses the core value of *security*. On his view, a set of culturally pervasive core values—life, happiness, freedom, knowledge, ability, resources, and security—are intrinsically valued, as are intrinsic expressions of their value (e.g., privacy's expression of security via privacy protection). For Moor, core values and those which intrinsically express them are mutually supporting, with varied evaluations across persons and contexts: "an athlete will emphasize ability, a businessperson will emphasize resources, a soldier will emphasize security, a scholar will emphasize knowledge, and so forth." In Walzerian terms, these core values are domain-bound social goods whose meaning and distribution are defined within particular social spheres.

While Moor's pluralist framework resonates with contextual integrity's value structure, his reduction of privacy's intrinsic value limited to its expression of security misses a deeper normative point. The wrongness we intuitively attribute to the Peeping Tom is not only a violation of protection or security, but a moral failure of recognition.

Anita Allen challenges Moor's conclusion on precisely this ground. She argues that the Peeping Tom's behavior is troubling not only because it violates an abstract right to be secure, but because of what the act expresses: a refusal to regard the subject as a moral equal [41]. Even in the absence of detection or downstream harm, the wrong lies in the act of treating another as a surveillable object—subjected to opaque and unreciprocated scrutiny. Allen's anti-spying principle reframes the privacy violation as a relational failure: a denial of basic respect, a refusal of ethical parity. The wrong is thus not about injury or lack of protection but subordination: it severs the moral relation between subject and observer, stripping the former of agency and dignity.

This insight reverberates in post-Prosser privacy jurisprudence. In *Hamberger v. Eastman* (1964), the New Hampshire Supreme Court upheld an intrusion-upon-seclusion claim based solely on the presence of recording devices in a couple's rented bedroom—even though there was no evidence the landlord actually listened to the recordings [3]. What mattered was not the informational loss, but the affront to dignity.

Later commenting on the case, legal scholar Robert Post emphasizes that the injury lay not in subjective distress but in the very nature of the act. Drawing on Warren and Brandeis, Post characterizes such invasions as violations of the “personality”—capable of producing “suffering more acute than that produced by a mere bodily injury” [699]. He underscores that the offense lies neither in what was done with the information nor in the harm caused, but in the desecration of the self *as such*. Yet even Post identified a structural weakness in the case: the legal standard for privacy intrusion torts depends on what a “person of ordinary sensibilities” would find offensive. This reliance on socially contingent norms weakens the moral clarity of the privacy claim, rendering recognition of affronts to dignity subject to shifting majorities.

Without a standard that treats human dignity as intrinsically inviolable—*independent* of cultural sentiment or legal precedent—privacy, selfhood, and democratic society remain contingent, fragile, and unequipped to protect the very interests a general right to privacy was meant to secure. As Post himself suggests, this fragility stems from law’s failure to defend privacy not merely as a functional good, but as a precondition for moral agency and defense of human dignity.

This is the radical core of Warren and Brandeis’ original insight: that privacy must function as a bulwark not only against private intrusion, but also *collective domination*—including state-sanctioned and socially ratified forms [739]. Legal scholars Rosen and Santesso argue that Warren and Brandeis’s articulation of privacy as both a defense of the self and a precondition for its development is best understood by considering what is lost in its absence: without protection from intrusion from coercive social forces—unjust norms, overreaching institutions, majoritarian powers—the moral architecture of the self erodes, or fails to take shape at all [739].

Where individuals lack the capacity to think, feel, and judge freely, the conditions for collective dissent—and for moral or political resistance to injustice—likewise deteriorate [609]. Even as Prosser’s tort taxonomy narrowed Warren and Brandeis’ insights about the value of privacy into a utilitarian balancing of harms, the logic of dignity endures. It persists in the moral imagination, reflected in our intuitive moral judgments, our jurisprudence, and our democratic and egalitarian ideals. To restore privacy’s normative force, it must be reasserted as a *first-order value*: a constitutive expression of respect for human dignity.

7.3.1.3 Privacy's Intrinsic Value as Constitutive of Moral Personhood

Having shown that dignity violations explain moral intuitions that recognize where privacy violations are *wrong*, even in the absence of codified rules or observable harm, we now move beyond this foundation to develop a full account of privacy as a constitutive condition of human dignity grounded in moral personhood. Drawing from legal theory, philosophy, and political thought, this section argues that privacy must be treated as a first-order moral good: not merely protecting who we are, but enabling who we can become.

Theories that recognize dignity as the core value at stake in privacy violations distinguish privacy as an intrinsic good from merely an instrumental one. Ronald Dworkin's theory of rights helps clarify this distinction: instrumental rights serve collective goals and may be overridden when doing so benefits the whole, while intrinsic rights express foundational commitments rooted in dignity and equal concern for all persons [276]. Only intrinsic rights can justifiably constrain what the state—or institutions, or society—may demand of an individual, even in pursuit of aggregate social welfare. Although Dworkin did not endorse privacy as a fundamental right in itself, citing its normative thinness in doctrine, he identified the moral scaffolding—autonomy, dignity, and equal concern and respect for all persons—from which a stronger normative defense of privacy could be built.

Allen's account of privacy as a moral boundary responds to this challenge by positioning privacy as a precondition to become a moral person. In her view, privacy is “a condition or set of social practices constituting, creating, or sustaining boundaries that should be drawn between ourselves and others in virtue of our status or potential as persons” [39]. These boundaries enable the formation and maintenance of personhood by affording individuals the space to distinguish themselves from others, reflect on their values, and act with self-determining agency.

Deborah Johnson elaborates this idea by emphasizing that privacy underwrites self-direction and moral development through reflective autonomy and self-realization [451]. S.I. Benn's conception of moral personhood further complements this claim. On his view, personhood is characterized not by qualities like sentience or biological origins, but our uniquely human capacities for reasoning, cooperation with others, and mutual expectations of moral responsibility and accountability, wherein moral standing entails the capacity to establish one's own identity in relation to others, to act in accordance with one's own reasons, and to be recognized as such [107, 108]. Privacy is essential to these capacities as it creates the psychic and social space necessary for individuals to establish their identities and resist coercion, constituting one's status as a moral equal.

It is by these very conditions that individuals can then freely and meaningfully associate with others, as Charles Fried argued, as it underlies our capacities to foster intimacy and other meaningful relationships with reciprocal moral trust, care, and love by managing the degree to which we are known by others [339]—and by extension, Thomas Nagel insists, our abilities to maintain cohesive

and cooperative societies [630]. James Rachels similarly contends that privacy is necessary for our interdependent relations with others, constituting our very capacities for self-development and abilities to navigate differentiated social roles and relationships with moral agency [705]. Jeffrey Reiman extends this by grounding privacy's in developmental psychology, arguing that even infants require a protected zone in which to learn the distinction between self and other—and by extension, to acquire a sense of bodily autonomy and relational interdependence to become a moral subject. Without such a zone, Reiman argues, “there would be no person, in the moral sense, to whom any rights could be meaningfully ascribed” [716]. Privacy then is both a structural and developmental precondition for moral agency and personhood.

Hannah Arendt deepens this perspective by linking privacy to the human condition of *natality*—the foundation of our capacity to begin anew, to initiate action, and to exist in the world as distinct individuals among others [62]. For Arendt, inner privacy constitutes the space where thought, emotion, and judgment are formed—necessary for the development of moral commitments and their public expression through acts of moral and political agency. Privacy, in this light, is both a constitutive condition of dignity and necessary for political freedom: essential for persons to enter the shared world not as passive subjects, but capable of meaningful, plural action—itself intrinsically valuable and key to realizing one’s own inherent worth [63, 767].

Julie Cohen extends these insights by rejecting the liberal premise that individuals exist as fully formed autonomous agents prior to socio-technical context. In *Configuring the Networked Self*, she argues that privacy is not a protective boundary around a pre-existing stable self, but rather a condition for its ongoing formation [201]. Cohen’s account theorizes privacy as necessary for sustaining the relational and infrastructural conditions under which autonomy becomes possible. She identifies the “autonomy paradox” in contemporary privacy discourse: individuals as treated both as rational actors capable of freely trading privacy for convenience, and as vulnerable subjects shaped by surveillance infrastructures. This contradiction, Cohen argues, masks the socially embedded production of autonomy and the necessary role privacy plays in sustaining it. Like Warren and Brandeis, Cohen resists theoretical paradigms that reduce the self as passively constructed products of social forces. She insists on formal recognition of privacy (i.e., in law, policy) as necessary, morally and politically, through structural protections of capacities to act as moral agency—capable of judgment, self-direction, and emotional depth in environments structured to erode them.

Reconnecting Warren and Brandeis’s principle of the inviolate personality with these diverse normative traditions, we begin to see privacy not as a negotiable interest, but as a moral threshold. Transcending informational models of privacy, a dignity-based privacy interest encompasses developmental, dispositional, and relational dimensions with individual-collective scope. Despite their varied disciplinary orientations, these accounts converge on a common normative commitment:

that dignity is a threshold condition necessary for just treatment, and that privacy constitutes the structural and moral grounds on which that dignity stands. Privacy, then, is both instrumentally valuable for social and political life and intrinsically valuable as a precondition of moral personhood and agency, securing the possibility of becoming a self, acting with moral judgment, and participating in the world as a bearer of dignity.

7.3.2 Defining and Measuring Human Dignity

If privacy is part and parcel of human dignity, governance requires a mechanism to recognize when privacy intrusions implicate that dignity. By what normative framework can we define, assess, and secure privacy in sociotechnical systems to uphold it?

This is a question of justice. Justice concerns the organization of social life: how rights, liberties, resources, and obligations are distributed, and what individuals owe one another—and to what each is entitled—as members of a shared political community [714]. But recognizing the intrinsic value of privacy introduces a second-order problem: how should law and policy translate this recognition into institutional protections that are reliable and context-sensitive, and up to what thresholds are they legitimate?

Reasoning about how to protect intrinsic values also confronts the problem of value conflict. Isaiah Berlin’s theory of value pluralism contends that certain values—freedom, dignity, equality—are both fundamental and incommensurable, such that conflicts between them cannot always be resolved by appeal to a higher principle [114]. Even when privacy is acknowledged as a constitutive moral or political good, it may appear to compete with other irreducible values. In such cases—as in U.S. legal reasoning weighing privacy against freedom [192, 422, 722]—the resulting tensions often result in trade-offs—what Berlin calls “tragic choices”: moral conflicts that demand political adjudication, not philosophical resolution.

Dworkin rejects this conclusion. He argues that apparent conflicts often stem from conceptual confusion, not from true incommensurability. Clarifying what each value demands—what it protects, enables, constrains—can dissolve apparent conflicts and reestablish normative coherence [704]. It is in this spirit that I turn to the Capabilities Approach: a normative framework that defines dignity in concrete, measurable terms and provides a method for identifying when social arrangements—including, for our purposes, data practices—fail to meet the basic requirements of justice. The moral minimum standard of human dignity, I propose, provides the normative floor missing from current AI and privacy governance.

7.3.2.1 Capabilities Approach to Human Dignity

The Capabilities Approach (CA) offers a principled account of the conditions necessary for a life of dignity. Developed by development economist Amartya Sen [773] and expanded into a full theory of justice by philosopher Martha Nussbaum [658], it redefines the metrics of equality and freedom by clarifying the conceptual articulation of these goods: not simply as the formal absence of interference or fair distribution of resources, but as the real opportunity to achieve valued human functionings. Rather than measuring justice through resources or formal opportunities, the CA centers on what people are actually able to achieve, given the social, material, environmental, and embodied constraints they may face, in recognition that the particularities of an individual's situation affect their capacity to convert opportunities into real freedoms. In doing so, it shifts the focus of justice to the minimum standards required to ensure people can transform abstract entitlements into lived experience, offering a method to both define and measure where these conditions fail to meet the minimum requirements for securing the capacity to live, at minimum, a life with dignity.

These concerns echo John Rawls' *difference principle*: social arrangements must be structured to benefit the least well-off to be considered just [714]. Until the 1990s, international development policy largely operationalized this ideal through aggregate measures of economic output—most notably Gross Domestic Product (GDP)—as proxies for justice-related metrics such as income distribution, wellbeing, and social progress. Yet such measures routinely obscured inequities at the individual level. In his 1979 lecture *Equality of What?*, Amartya Sen famously challenged Rawls and the international development community to reconsider the metric of justice itself, pressing the question of whether equality should be measured by income, resources, or something else [773]. This critique laid the groundwork for the Capabilities Approach, which Sen and Nussbaum would eventually develop to argue that the most telling measure of a just society—whether disparities arise on account of individual differences or entrenched structural constraints—lies in people's actual *capabilities*: their real opportunities to *be and do* what they have reason to value [770, 771, 658]. By extension, Rawls' difference principle invites scrutiny into how power-imbalanced digital infrastructures systematically favor certain groups while exacerbating vulnerabilities for others. Whether through biased algorithms or exploitative data business models, Nussbaum's CA supplies both a normative theory and methodological lens to diagnose and respond to such conditions by establishing human dignity as the relevant minimum standard of justice [658, 657, 660].

Unlike abstract or procedural accounts of dignity, Nussbaum's model is grounded in the differentiated conditions under which dignity is either enabled or denied. Emphasizing the “overlapping ethical consensus” on human dignity across cultures, Nussbaum's CA sets out to define what is required to uphold human dignity for any person, in any setting [658]. Her theory proposes a set of

ten core capabilities required to live a dignified life—a life one has reason to value—that together constitute a minimum threshold of justice when secured [658]. A threshold standard for dignity that holds across political, economic, and cultural variation, these constituent components include minimum thresholds for developing internal capabilities such as emotion, practical reason, and imagination, as well as exercising external capabilities such as affiliation and control over one's environment. Like Walzer, Nussbaum affirms the need for moral minimums that serve as the floor for contextual moral elaboration. But where Walzer leaves such minimums underspecified, Nussbaum provides a robust philosophical foundation: a dignity-based account rooted in cross-cultural dialogue and supported by a normative framework capable of identifying the conditions required to realize human dignity in any society [660].

Nussbaum's framework differs from Rawlsian justice in two crucial respects. First, it overcomes the failures of justice metrics that over-emphasize the distribution of primary goods in masking inequities in transforming these goods into the lived realities of their everyday lives [773] by foregrounding human dignity as the central evaluative standard: each person must be treated as an end in their own right, not merely a bearer of abstract rights lacking the means to realize them. Advocating for a *minimal justice* standard to secure human dignity in the invariably complex interactions between individual capacities and external social, environmental, cultural, and material constraints, then, is a matter of ensuring that every person has the real freedom to live a life of dignity. By insisting on the individual's capacity for *choice*—to develop one's own vision of the good and act as an agent of one's own life to pursue it, rather than exist merely as a passive recipient of external social forces—Nussbaum's CA articulates the constituent elements of dignity in a form that meets the pluralist imperatives Mattson and Clark call for in an adequate model of human dignity [583], resisting both moral relativism and moral imperialism.

Second, Nussbaum's CA addresses the limitations of policy interventions that act only where harm is measurable, observable, and cognizable—a standard that can perpetuate injustice in two ways: by reducing what should be fundamental entitlements as merely instrumental, and failing to recognize harms that are less tangible—familiar limitations in privacy jurisprudence, where harms must be made legible and justified against competing interests to be recognized or remedied at all, and let alone prevented [193, 793, 794, 468, 722]. Whereas Rawlsian justice calls for policy intervention only after such harms are acknowledged and considered against other goods, Nussbaum's alternative theory of minimal justice identifies a failure of justice—and therefore, a legitimate site for policy intervention—where *any person* cannot realize a life of dignity as defined by the core capabilities, due to institutional failure to secure the necessary conditions for their agency [657]. Derivatively, then, a society that secures *minimally just* conditions—wherein every person securely holds the capacity to exercise and develop each of these capabilities, at least up to their morally justified floors—can be considered minimally just.

Nussbaum's specifies ten core capabilities as the necessary conditions for securing human dignity *for each person*, with each defined in terms of a minimum threshold. The core capabilities are as follows [658]:

- **"Life.** Being able to live to the end of a human life of normal length; not dying prematurely, or before one's life is so reduced as to be not worth living.
- **Bodily Health.** Being able to have good health, including reproductive health; to be adequately nourished; to have adequate shelter.
- **Bodily Integrity.** Being able to move freely from place to place; having one's bodily boundaries treated as sovereign, i.e. being able to be secure against assault, including sexual assault, child sexual abuse, and domestic violence; having opportunities for sexual satisfaction and for choice in matters of reproduction
- **Senses, Imagination, and Thought.** Being able to use the senses, to imagine, think, and reason – and to do these things in a ‘truly human’ way, a way informed and cultivated by an adequate education, including, but by no means limited to, literacy and basic mathematical and scientific training. Being able to use imagination and thought in connection with experiencing and producing self-expressive works and events of one's own choice, religious, literary, musical, and so forth. Being able to use one's mind in ways protected by guarantees of freedom of expression with respect to both political and artistic speech, and freedom of religious exercise. Being able to search for the ultimate meaning of life in one's own way. Being able to have pleasurable experiences, and to avoid non-necessary pain.
- **Emotions.** Being able to have attachments to things and people outside ourselves; to love those who love and care for us, to grieve at their absence; in general, to love, to grieve, to experience longing, gratitude, and justified anger. Not having one's emotional development blighted by overwhelming fear and anxiety, or by traumatic events of abuse or neglect. (Supporting this capability means supporting forms of human association that can be shown to be crucial in their development.)
- **Practical Reason.** Being able to form a conception of the good and to engage in critical reflection about the planning of one's life. (This entails protection for the liberty of conscience.)
- **Affiliation. A.** Being able to live with and toward others, to recognize and show concern for other human beings, to engage in various forms of social interaction; to be able to imagine the situation of another and to have compassion for that situation; to have the capability

for both justice and friendship. (Protecting this capability means protecting institutions that constitute and nourish such forms of affiliation, and also protecting the freedom of assembly and political speech.) **B.** Having the social bases of self-respect and non-humiliation; being able to be treated as a dignified being whose worth is equal to that of others. This entails, at a minimum, protections against discrimination on the basis of race, sex, sexual orientation, religion, caste, ethnicity, or national origin. In work, being able to work as a human being, exercising practical reason and entering into meaningful relationships of mutual recognition with other workers.

- **Other Species.** Being able to live with concern for and in relation to animals, plants, and the world of nature.
- **Play.** Being able to laugh, to play, to enjoy recreational activities.
- **Control over One's Environment.** **A. Political.** Being able to participate effectively in political choices that govern one's life; having the right of political participation, protections of free speech and association. **B. Material.** Being able to hold property (both land and movable goods), not just formally but in terms of real opportunity; and having property rights on an equal basis with others; having the right to seek employment on an equal basis with others; having the freedom from unwarranted search and seizure.”

Crucially, each of the ten core capabilities is distinctive and irreducible—none can be justifiably reduced below its minimum threshold, nor traded away at the expense of another, as each forms a constitutive part of what makes for a worthwhile human life. By treating these capabilities as the minimal conditions for justice, Nussbaum's CA offers a normative floor beneath which no data flow, institutional practice, or technological system should be permitted to fall—where they undermine a person's claim to live as a full human being.

Dignity is often invoked as the principled heart of privacy [329, 706]. Nussbaum's framework brings precision to the task of anchoring privacy's intrinsic value in human dignity, providing a normative criterion to evaluate whether data flows meet the *minimal justice standard* of dignity's inviolability in terms of concrete, lived deprivations. While Nussbaum does not specify privacy as a core capability in itself, if we accept the arguments developed in Section 7.3 that privacy is both a necessary facilitator and constitutive enabler of human dignity and moral personhood [201, 39, 876]—then Nussbaum's account of what is required for dignity offers the conceptual clarity needed to defend privacy as an intrinsic value by its essential role in sustaining dignity in everyday life. In the core capabilities, we more clearly see that privacy is indispensable to their development and exercise. To have agency in the face of coercive external influences and constraints, privacy is essential: to use *practical reason* to cultivate one's own vision of the good

[62]; to experience, develop, and act upon one’s *emotions* and dispositional self in accordance with one’s values [200, 39]; to experience one’s *senses, imagination, and thought* with moderation over how one’s inner life—thoughts, emotions, and sentiments [876, 339]—is known by others; to maintain **bodily integrity** with autonomous decision-making about one’s *life* and *bodily health* [505, 39]; to freely participate in political and material environments [62, 100, 825]; to maintain meaningful *affiliations* with others on terms of dignity and mutual recognition [193, 41], and to engage in *play* and personal pursuits [39]. In these ways, we see that privacy is both instrumentally valuable and intrinsically essential for sustaining the very capabilities that constitute a life with dignity.

Nussbaum’s CA supplies a threshold logic for human dignity: every person must have real freedom to develop and exercise the ten core capabilities up to their minimal thresholds to live a life with dignity in their everyday realities—and by extension, for the societies on which they depend to be considered minimally just [658]. Privacy intrusions that erode the core capabilities below their minimal thresholds are therefore dignity violations. Understood this way, privacy is not a good to be weighed, but a structural precondition for realizing a life of dignity. Where its minimum thresholds are eroded, privacy cannot justifiably be traded away. Instead, the data practice itself must be restructured to meet the minimal justice standard of dignity’s inviolability.

Yet tracing how data flows impact dignity remains a challenge. As Nussbaum’s CA does not explicitly identify privacy as a core capability, it supplies neither a vocabulary nor a method for identifying when information practices cross the inviolable line of human dignity. Its utility in evaluating data flows thus depends on articulating how privacy intrusions interfere with the development and exercise of the core capabilities—and on identifying a method for tracing those intrusions in the daily churn of data flows within socio-technical systems. We need, then, an analytic lens that is already tuned to the structure and meaning of data flows in the reality of informational practice.

7.3.2.2 Conserving Dignity in Contextual Integrity

Nissenbaum’s Contextual Integrity (CI) provides that lens: with analytic precision, CI describes and prescribes privacy as contextually appropriate data flows. By mapping normative privacy judgments onto five parameters—data subject, sender, recipient, information type, transmission principles—CI translates diffuse social expectations into concrete evaluative criteria, and is unrivaled in diagnosing when a particular flow is contextually inappropriate. But as established in Section 7.2.3, precisely because CI’s authority rests on local normative logics, it cannot on its own condemn information practices that are systematically degrading: cases in which the norm-setting process is distorted, novel flows bypass public deliberation, or entrenched practices silently erode the standing of the least empowered.

Absent a way to bridge failures to respect both human dignity and privacy, data practices those that erode individuals' core capabilities—to reason, feel, and act such that one has the capacity to live a life they have reason to value—may persist unchallenged and without recourse. And when those affected are contextually positioned as least empowered to resist, whether through constrained choice or structural exclusions—workers, patients, children—the normative grammar of privacy collapses under the weight of power asymmetries, as appeals to local norms lose force. CA and CI are therefore complementary: in Walzerian terms [872], CI secures the “thick” justice of each sphere via contextually appropriate data flows, while CA supplies the “thin” universal moral minimum to ensure against outright domination via human dignity.

Both value-pluralist frameworks offer a methodology for locating privacy and dignity in everyday realities that is tractable to empirical support: CI grounds privacy in the grain of social practice [646]; CA grounds dignity in the texture of human functioning [658]. To bridge the two theories into an operationalizable framework, I propose three adaptations to CI:

- 1. Fix Dignity Threshold Transmission Principles.** The transmission principle parameter in CI (one of the five constituent components of a data norm defined by CI) modulates constraints on a data flow in a particular context, such as requirements for consent, expectations of reciprocity, claims of desert or entitlement to the information, jurisdictional regulatory demands, and so forth. If we adopt Nussbaum’s Capabilities Approach, which defines human dignity in terms of “core capabilities,” as CI’s universal moral minimum, then ensuring that no data flow undermines those capabilities below their minimal thresholds can function as a fixed transmission principle in CI, constraining data flows that fail to secure the conditions required to uphold human dignity.
- 2. Extending Appropriateness Determinations with the Moral Minimum Standard of Human Dignity.** Incorporating Nussbaum’s dignity parameters into CI would effectively extend CI’s layered normative standard of appropriateness by adding a foundational moral layer. CI’s existing heuristic assess a data flow’s appropriateness across three instrumental layers: first, by evaluating the interests of affected parties; second, by determining whether those impacts are just relative to local moral and political values; and third, by assessing whether the data flow upholds or undermines the context’s teleological purposes (i.e., its social ends) [649, 650]. Integrating the CA introduces a prior, dignity-based test: before CI’s heuristic is applied, a data flow would be assessed for its foreseeable impact on the conditions necessary for every person to live a life with dignity. This moral minimum serves as a universal threshold, below which no data flow can be considered appropriate regardless of contextual consensus. Thus, appropriateness would first be determined on the intrinsic grounds of safeguarding human dignity, and only then on the instrumental grounds of conserving

context-relative norms and purposes. The dignity threshold ensures the “thin” universal moral minimum standard of human dignity is preserved, so that the “thick” local moral maximums can be meaningfully sustained. Together, CI and CA offer a justificatory test that is both context-sensitive *and* dignity non-negotiable: a data flow is appropriate only if it respects contextual informational norms (CI) *and*, more fundamentally, if its impact does not push any person beneath capability minima.

3. **The Role of Purpose.** As Nissenbaum notes, inference-based systems lack the socially-contingent meaning needed to normatively evaluate personal inferences, presenting a challenge for CI’s framework to evaluate such practices [650]. As detailed in 7.1, the CI-based method I developed for evaluating emotional privacy judgments in Chapter 6 directly addresses this limitation in two ways. First, it interprets CI’s *information type* parameter as the inference itself (e.g., inferred emotional state), assigning meaning to the information as a machine-generated interpretation of the person’s emotions. Second, it incorporates the *data inputs* (e.g., speech, text, video) and critically, the *purpose* for which the inference is generated and used. Together, these additional variables specified the normative parameters needed to establish the *meaning* of an inference and enable subjects to evaluate its appropriateness relative to context. A central insight from this work is the indispensable role of *purpose* in shaping privacy judgments of emotion inferences. The purpose for which information is inferred—why it is generated, and to what end—emerged as a key normative axis alongside who receives the information and under what conditions. In novel inference-based systems where established norms and meaning are absent, *purpose* serves as a substitute to prior negotiated meaning, offering both descriptive precision and normative justification. Accordingly, I propose extending CI to include *purpose* as a sixth constitutive parameter. In addition to improving the framework’s empirical adequacy in accounting for novel or AI-driven data practices, doing so also strengthens its normative grounding. Purpose operates as a moral anchor, allowing evaluators to assess whether the aims of a data flow align with both the teleological ends of the contexts and whether they respect the dignity of the individual. Integrating purpose formally enables CI to meet the directive central to CA: to treat every person as an end in themselves—no person’s dignity should be traded off in service of another’s ends.

In sum, fusing Nussbaum’s CA with Nissenbaum’s CI yields three payoffs. First, it equips CI to evaluate novel or illegitimate data practices that lack (or flout) established norms, addressing known limitations CI faces in advanced socio-technical systems [646, 650]. Second, it anchors contextual judgments to a non-waivable dignity floor, fulfilling Walzer’s insistence on a universal moral minimum needed to preserve the integrity of social spheres [872]. Third, it restores privacy’s

moral standing as an intrinsic good essential to human dignity, offering a principled method to identify when data flows unjustifiably cross the threshold of dignity's inviolability.

7.3.2.3 Capabilities-Augmented Contextual Integrity (CA-CI)

To clarify the complementary strengths of Contextual Integrity (CI) and the Capabilities Approach (CA), Table 7.1 offers a comparative overview.

Dimension	Contextual Integrity (CI)	Capabilities Approach (CA)
Key Theoretical Focus	Appropriate data flows governed by context-relative norms and goals	Appropriateness emphasizes every person's real capacity to live a life with dignity
Aim or Purpose	Preserve context-specific informational norms that implicitly protect societal values.	Ensure individuals can exercise core capability minima that together constitute conditions necessary to develop and pursue a life one has reason to value
Limitations and Gaps	Does not specify which core values or ends must be safeguarded; lacks external evaluative standard.	Does not specify values of privacy, data protection; needs bridging to data and AI governance contexts.

Table 7.1 Comparative Overview of Contextual Integrity and Capabilities Approach

The unification of CA and CI strengthens both frameworks by enabling normative evaluation of data flows in terms of both contextual integrity and human dignity. The CA-CI model I propose integrates CA's dignity-based thresholds into CI's contextual architecture, preserving CI's descriptive and interpretive strengths while adding a principled standard to identify when even a contextually "appropriate" data flow may constitute a deeper moral violation.

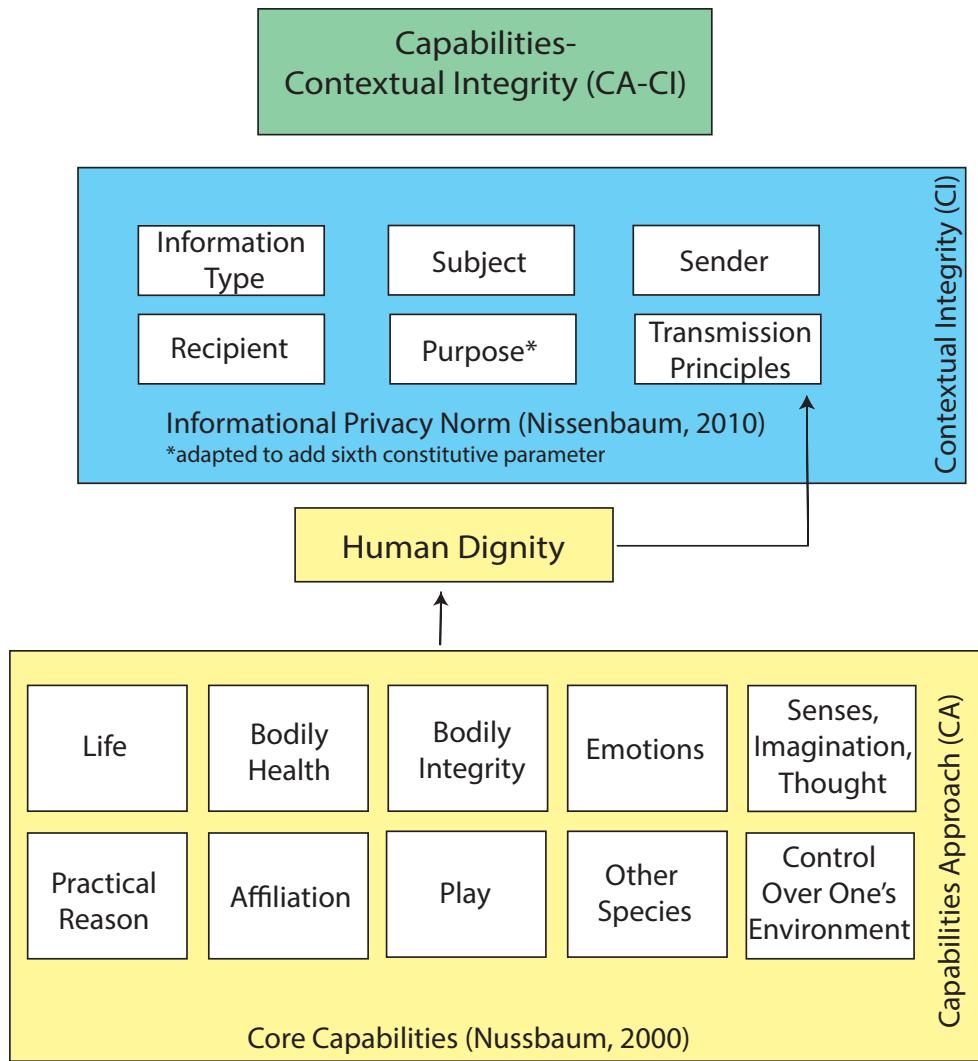


Figure 7.1: Capabilities–Contextual Integrity (CA–CI) Theoretical Framework – Integrating Fixed Dignity Thresholds, Purpose Parameter

Operationally, as shown in Figure 7.1 CA–CI treats capability thresholds as a special class of *transmission principles*. If a data flow can be reasonably expected—based on prior evidence, design intention, or foreseeable outcome—to impact any of the ten core capabilities, it triggers a risk assessment to evaluate its potential to erode an individual’s capabilities below threshold.

By first asking *Does this information flow respect the dignity of the individual?*, CA-CI prompts a bottom-up evaluation of a data flow’s impact to the core capabilities that comprise human dignity. By inductively evaluating potential impacts to a person’s capacity to reason, relate, or act that may be too subtle to manifest as measurable, observable, and cognizable harms, but nonetheless risk compromising the foundations of human functioning, CA-CI’s framework enables the detection of capability erosion *pre-harm*. Iteratively, this analysis may prompt changes to the socio-technical design or data flow (e.g., modifications to the purpose parameter or the imposition of additional transmission constraints) to facilitate mitigation planning.

7.4 Tracing the Dignity Line in Practice

Over the past decade, a broad consensus has emerged across academia, civil society, and industry on the need for systematic governance of data and AI systems [55, 255, 718, 611]. Standards bodies, regulatory authorities, and multistakeholder consortia have responded by developing a range of frameworks intended to guide the design, deployment, and oversight of these systems. Operationalizing these frameworks remains challenging, however. Their implementation hinges on organizational interpretations of complex, underspecified concepts—what counts as *consent*, *personal data*, a *privacy violation*—in practice, leading to significant variation and uncertainty.

In the absence of structured normative guidance, key privacy risks can go unrecognized, including those involving sensitive inferences, surveillance practices, and repurposed de-identified data [427, 819, 510]. These blind spots are compounded by privacy’s marginal role within most organizations: typically treated as a compliance function, privacy is often siloed in advisory or legal compliance roles removed from system design. This structural separation undermines effective risk identification, weakens oversight, and narrows mitigation—leaving critical contextual and normative questions unexamined [868, 82] while obscuring deeper socio-technical vulnerabilities—misaligned system defaults, unexamined modeling assumptions, and inadequate sensitivity to context. [510].

The practical value of CA-CI is made even clearer in the following case studies, which underscore that even under robust governance regimes or formal compliance protocols, data practices can violate human dignity without triggering any legal or institutional response. As I show, applying the Capabilities-Augmented Contextual Integrity (CA-CI) model to these cases can clarify and evaluate privacy risks that elude both procedural compliance and classification-based governance. Across the three cases reviewed, CA-CI identifies not only *what* is problematic, but *where* in the data flow the violation occurs. Its practical utility lies in translating the abstract concept of dignity into a concrete evaluative threshold, enabling precise intervention in data governance workflows when integrated with contextual integrity’s privacy framework.

7.4.1 Crisis Text Line: The Limits of Internal Governance

In the absence of comprehensive privacy legislation or legally binding AI governance mandates, the U.S. approach to data and AI oversight rests largely upon voluntary or industry-driven frameworks. These include the NIST AI Risk Management Framework (AI RMF 1.0) [821], the NIST Privacy Framework [129], and ISO/IEC 23894 [17], which promote practices such as consequence modeling, traceability, and lifecycle-based risk documentation. The foundation these frameworks offer for managing systemic harms embeds an instrumental logic: privacy is treated as a parameter to be balanced, optimized, or sacrificed in pursuit of gains like performance, innovation, or utility. While such guidance may clarify that organizations must balance tradeoffs among competing values (e.g., see NIST AI RMF [821]), a major implementation challenge concerns *how* to do so: which values to uphold and which to trade, ethically and responsibly.

The 2022 Crisis Text Line (CTL) case makes the consequences of this logic clear. CTL, a nonprofit offering SMS-based mental health counseling, licensed millions of anonymized crisis conversation transcripts to its for-profit spin-off, Loris.ai, to train commercial “empathy” algorithms [306, 599]. While procedural protocols were followed—including data de-identification, internal review, and contractual controls—the informational flows violated contextual norms rooted in therapeutic trust and expectations of strict confidentiality. Disclosures made in moments of acute emotional crisis are more than sensitive; they are sacrosanct [268]. Repurposing them for product development constituted a profound betrayal, both of individual dignity and the moral infrastructure underpinning crisis support. With trust in the ethic of care breached, help-seeking behavior can reduce—leading to life-altering, even fatal, outcomes.

As a U.S.-based nonprofit, CTL operated outside many institutional accountability structures, including the jurisdiction of the Federal Trade Commission (FTC), which serves as the de facto privacy enforcement authority for commercial entities in the U.S. [500]. Although the Federal Communications Commissioner Brendan Carr publicly urged the FTC to investigate, the agency lacked authority over non-profits. As former FTC’s Consumer Protection Bureau Director Jessica Rich explained, any regulatory scrutiny would depend on establishing that CTL’s commercial relationship with Loris.ai constituted a deceptive practice that contradicted the non-profit’s stated privacy assurances—a legal theory “there are a lot of questions about whether...the FTC could pursue” [134].

CTL’s self-governed approach involved an instrumental calculus: the risk of harm was *minimized* via de-identification, while the private value of the data was *optimized* through commercialization. Yet the harm was not procedural failure—it was moral blindness. Governance mechanisms worked as designed; what failed was their ability to register the intrinsic moral status of the data flow itself. Public outrage quickly followed—not only in response to a lack of consent or transparency (and their limits to arbitrate privacy claims in this context [698]), but because the very institution

entrusted with care had commodified deeply personal crisis interactions into a transactional asset. The violation was not incidental to governance, but stemmed from it as the organization's own product. Procedural safeguards were satisfied, but no mechanism existed to evaluate whether the practice was normatively acceptable.

This case exposes two overlapping deficits. First, it illustrates how dignitary and contextual harms may escape recognition even when governance procedures are followed. Second, it reveals a deeper structural absence—the lack of an independent normative threshold to determine when a data practice is categorically wrong.

Contextual Integrity (CI) would likely classify CTL's data flows to Loris.ai as inappropriate, subverting the very integrity of the crisis care context in which trust was extended. But CI ultimately defers such judgments to domain-embedded actors [646], and in this case, their internal normative logic failed. A board including privacy and tech ethics experts claimed that the very existence of the context—its capacity to provide and scale its crisis services—depended on commercial partnerships, and thus authorized the sale [306].

Here, CI's core insight is affirmed: the public's moral intuitions reflect a shared recognition that the data flow violated the appropriate boundaries of the crisis care context. Yet this very strength—deference to contextual norms—becomes a liability when those norms are shaped by institutional self-interest. In the absence of external constraint and accountability to specify when a practice is categorically inappropriate—failing to meet a moral *minimum*—even well-intentioned institutions may rationalize serious harm. Tools like Privacy or Data Protection Impact Assessments (PIA/DPIA) may surface procedural risks, but without clarity as to when those risks cross a moral line, their assessment remains vulnerable to institutional incentives and interpretive drift.

All told, the Crisis Text Line case illustrates four overlapping governance failures:

- **Contextual goal erosion:** Repurposed and commodified crisis conversations subverted the integrity of crisis support as a social domain.
- **Instrumental logic:** Privacy treated as an optimizable resource, not a dignity-based constraint.
- **Failure of self-judgment:** Reliance on local, self-determined moral judgments approved inappropriate flows.
- **No normative floor:** No baseline mechanism to flag flows as categorically inappropriate.

What the CTL case ultimately reveals is not an isolated ethical lapse, but a structural vacuum in normative governance. In the absence of a clearly defined normative floor, harms are not just overlooked—they are produced.

CA-CI's Evaluation of the Crisis Text Line Case.

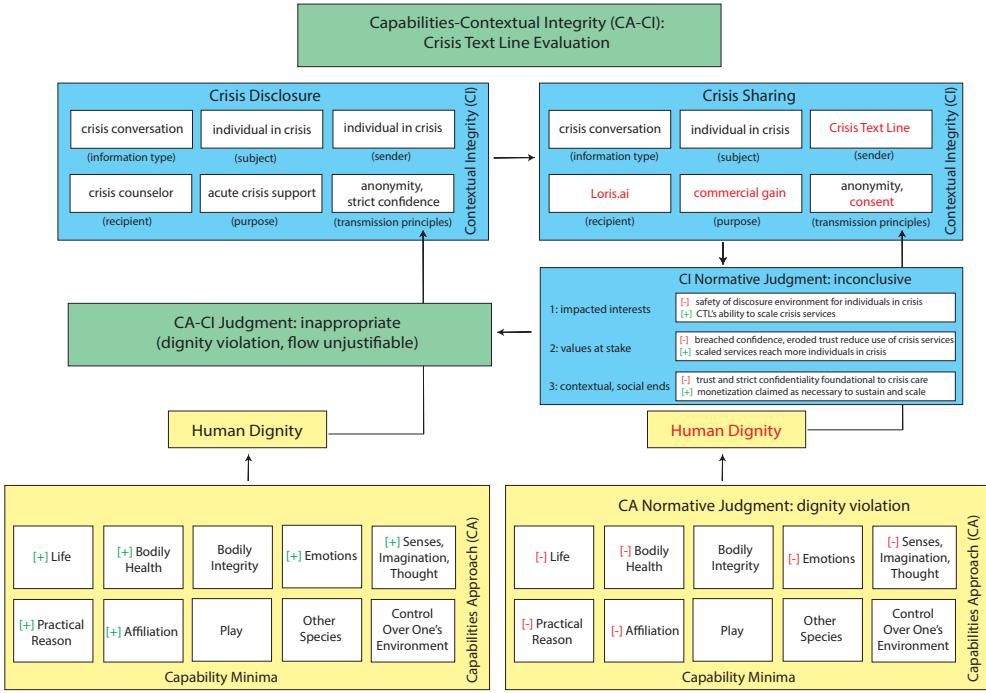


Figure 7.2: Capabilities–Contextual Integrity (CA–CI) Evaluation of Crisis Text Line.

As shown in Figure 7.2, CA–CI identifies the violation not in the presence of identifiable data, but in the recontextualization of crisis conversations as a corpus for commercial model development. This data flow departs from its originating contextual purpose—emergency mental health support—and breaches its core transmission principle of strict confidentiality. Neither consent nor de-identification can redeem the appropriation of crisis disclosures for purposes fundamentally misaligned with those of the original context.

CI's normative heuristic recognizes the risks and benefits to all parties. For individuals in crisis, the very space that promises refuge becomes a data minefield—turning candid, life-or-death disclosures into commodities and undermining the conditions necessary for seeking support safely. CTL, by contrast, argues that monetizing the corpus enables service expansion, thus benefiting more individuals in crisis [698]. This tension is reflected in CI's second and third normative layers: the values at stake are contested, and the contextual ends themselves—crisis care and the strict confidentiality that sustains it—are placed in jeopardy by the very flow justified in their name.

CI's normally decisive third layer, which evaluates whether a data flow supports or erodes a context's teleological aim, remains inconclusive. On one hand, the breach of confidence undermines the trust that makes crisis support possible; on the other, CTL frames monetization as necessary to sustain the very context of crisis care. In effect, the internal teleology of the context fractures: confidentiality and continuity appear as competing goods. (Notably, CTL continues to operate

in the years since it ended the data-sharing relationship with Loris.AI—suggesting the claimed existential risk to context was overstated.)

The CA, however, provides a more decisive judgment. It recognizes the crisis disclosure flow as positively contributing to core capabilities. Providing immediate crisis care helps sustain the capabilities of *life* and *bodily health* by cultivating *practical reason* and *emotions* capabilities to help individuals navigate acute distress. Through compassionate interaction, individuals are encouraged to engage their *senses, imagination, and thought*—for example, through grounding techniques or distraction strategies to mitigate suicidality [800]. Such support also enhances *affiliation*, both by strengthening social bonds in crisis care contexts and by affirming the individual as a being of equal moral worth, entitled to dignity-preserving care.

In contrast, the crisis sharing flow threatens these same capabilities. If individuals no longer trust CTL to keep their interactions in confidence, they may avoid the service altogether, placing their *bodily health* and *life* at greater risk. The loss of *affiliation* with CTL diminishes the conditions under which individuals can *reason, emotion, and think* through acute psychological pain. Because this is a crisis context, individuals are already at or near the threshold for many core capabilities; crisis support functions as a critical nudge, either above or below that line. Given both the high likelihood and severity of negatively affecting these capabilities, the CA-CI model clearly identifies the data flow as unjustifiable. Even if CTL’s internal reasoning appeals to net contextual benefit—expanding access by scaling services—CA-CI’s rejects this justification. Its insistence that every individual be treated with dignity demands that the impact on each person must, at minimum, not reduce capabilities below threshold.

CA-CI reverses the normative logic: it asks not how to minimize harm while extracting value, but whether a given data flow is minimally justifiable at all. By embedding context-sensitive moral reasoning directly into governance workflows, CA-CI recognizes that repurposing crisis conversations for commercial optimization recasts a care relationship into one of extraction—violating both the thick contextual expectations of relational trust protected by CI and the thin moral floor of dignity established by the Capabilities Approach.

In the mitigation phase, CA-CI prompts scrutiny of whether alternative socio-technical configurations might achieve operational goals without violating contextual and moral boundaries. Iterating on changes to the data flow would likely reveal that any breach of the crisis context’s strict standard of confidentiality, when undertaken for commercial gain, is categorically inappropriate. While modification to transmission principles are unlikely to render such a flow acceptable, entirely novel data flows with materially different parameters (e.g., distinct actors or purposes) may yield different outcomes. For instance, training models on synthetically generated conversation data could offer a plausible alternative. Yet even such alternatives require ethical adjudication: if synthetic data is derived from real disclosures, core capabilities may still be compromised—such

as through perceived privacy intrusions, representational harms, or risks of re-identification [405]. Where the moral status of novel technical solutions to novel privacy problems has not yet been socially negotiated, these alternatives can be considered justified only when credible evidence, such as from patient-centered studies, demonstrates that capability minima are preserved.

The Crisis Text Line case is a cautionary one: even trusted institutions, when operating under instrumental logics and absent normative thresholds, can enact profound dignity violations while remaining in formal compliance and appealing to contextual ends. CA-CI does not replace local judgment, but it holds it accountable to a shared expectation of human dignity, operationalized through capability thresholds and contextual parameters. In doing so, CA-CI sustains the integrity of contextual norms not by overriding them, but by anchoring them in a cross-contextual moral minimum. As Walzer observed, contextual legitimacy requires not just internal coherence, but fidelity to shared basic moral expectations. CA-CI enforces that normative floor, ensuring that institutions cannot justify violations of dignity in the name of local purpose. It reorients privacy governance from permissive tradeoff to principled constraint—a methodology for restoring the very values that socio-technical systems, institutions, and data practices are ostensibly designed to serve.

Table 7.2 CA-CI Evaluation of Crisis Text Line

Evaluation	Contextual Integrity (CI)	Capabilities Approach (CA)	CA-CI
Descriptive Analysis	Prima facie violations: sender, recipient, and purpose parameters change between crisis disclosure and sharing.	Sale threatens <i>life; bodily health; emotions; senses, imagination, and thought; practical reason, affiliation</i> .	
Normative Reasoning	Contextual ends undermined by sale, yet opportunity cost purportedly presents existential risk.	Risks to capability minima.	
Final Judgment	Normative appropriateness ambiguous	High likelihood of severe dignity violation	Reject until thresholds met

7.4.2 Clearview AI: The Fragility of Rights Without Dignity

Mass surveillance systems are inherently privacy intrusive, recognized as denying rights to privacy, data protection, and the right to anonymity in jurisdictions like the European Union—“gross

violations of fundamental rights” [210].

Clearview AI is a paradigmatic example. A U.S.-based company that operated largely outside the scope of federal privacy law, Clearview scraped billions of publicly available facial images from the internet to create a massive biometric database. Clearview offers near-instant 1:1 identification to law enforcement, intelligence agencies, and, until a legal settlement prohibiting its sale to most U.S. businesses in 2022, to private entities [413, 663]. By linking public images to facial recognition systems, it collapsed the distinction between public visibility and permanent traceability. The company’s operating logic reveals a familiar pattern: data originally shared for one purpose (e.g., social media, journalism, personal websites) is extracted, aggregated, and repurposed in an entirely different domain (e.g., criminal justice, border control, intelligence), creating downstream risks that remain unassessed and ungoverned. While defenders point to public safety and national security benefits, critics emphasize the systemic risk: when identity becomes a persistent exposure, everyday presence becomes a vector for algorithmic tracking in the public domain.

Although criticisms of biometric facial recognition often focus on intersectional demographic disparities in accuracy [155], evaluations such as the NIST Face Recognition Vendor Test report that top-performing identification algorithms exhibit minimal demographic variation, with low false positive and false negative rates across most groups [379, 596]. To mitigate misidentification risks, vendors like Clearview generally require clients to agree to conduct human reviews before taking action based on system outputs. Yet such procedural safeguards are not foolproof. In one documented case involving a facial recognition database, a Black man was falsely arrested—publicly, in front of his neighbors and children—after a white eyewitness mistakenly confirmed a match from a facial recognition-generated list [493].

Clearview has remained protected by procedural ambiguity and jurisdictional limits, evading compliance with enforcement actions under the GDPR such as fines and prohibition orders from EU regulators [453]. As a non-EU entity, it was not meaningfully constrained by GDPR despite its extraterritorial influence, nor was it subject to any U.S. federal privacy statute. No binding requirement for algorithmic impact assessment applied, and no independent body was positioned to evaluate the societal risks posed by Clearview’s system—not only impacting individual privacy, but also collective impacts such as democratic participation, freedom of assembly, and the ability to live without fear of retroactive identification. The EU AI Act’s Article 5 attempts to limit such risks through targeted bans: on real-time remote biometric identification in public spaces, predictive policing, and biometric categorization that infers sensitive attributes such as race or religion [302]. Yet these prohibitions include exemptions for serious crime investigations and carveouts for national security, defense, and scientific research—shielding these domains from scrutiny without adequate oversight [386].

Frameworks like the NIST AI Risk Management Framework (AI RMF) also fall short.

Clearview’s multi-sector deployment exposes the AI RMF’s inability to adequately track cross-domain use or account for harms which accumulate over time. Even key thresholds, such as when physical, psychological, or reputational risks of harm count as “significant,” are vaguely defined and internally specified, lacking enforceable standards [210, 819]. More concerning is the lack of guidance on data integration and inference risks within enterprise AI governance. Surveillance systems like Clearview exploit the fact that biometric data can be cross-linked with other information—either inferred from the same data source (e.g., emotional signals) or aggregated from disparate datasets (e.g., geolocation, behavioral metadata)—enabling re-identification and predictive profiling. As Mosaic theory from U.S. constitutional law demonstrates, seemingly innocuous data points can, when combined, yield highly sensitive inferences [104]. Empirical studies confirm that location data, mobile use, and social media behavior can reliably predict sensitive attributes like occupation, gender, and mental health status [515, 875, 73].

In the absence of substantive protections at the system-design level, regulatory compliance becomes the default horizon for governance. These compounding risks are poorly addressed in both the NIST AI RMF and the EU AI Act, which offer minimal guidance on data retention, reuse, temporal aggregation, or long-term harms [819]. The EU AI Act further narrows inference restrictions to a narrow set of “sensitive” categories (e.g., race, religion, political beliefs), excluding others such as emotional state [401] that, as my empirical work in Parts II and III consistently show, are just as susceptible to misuse and capable of eroding dignity. While the Act bans emotion recognition in schools and workplaces, it remains permissible in other high-risk settings, including migration screening and law enforcement, where the risks to dignity are no less severe.

All together, the Clearview case illustrates four overlapping governance gaps:

- **Cross-contextual risk:** Multi-sector systems complicate risk classification and threshold determinations.
- **Downstream risks:** Lack of adequate oversight or safeguards for harms arising from system use (e.g., human-in-the-loop) or data handling practices (e.g., sensitive inferences, re-purposing, re-linking), especially as these risks compound over time.
- **Enforcement evasion:** Jurisdictional limitations and legal ambiguities render data protection regimes difficult to operationalize or enforce across global contexts.
- **Lack of basic normative thresholds:** Vague classifications and risk thresholds in existing frameworks offer no means to identify when emerging or cumulative harms cross an ethical line.

What is needed is a framework that supplements procedural regimes with concrete normative

thresholds: standards capable of identifying when a practice's impacts fail to meet basic moral expectations, regardless of jurisdiction or context.

CA-CI's Evaluation of the Clearview AI Case.

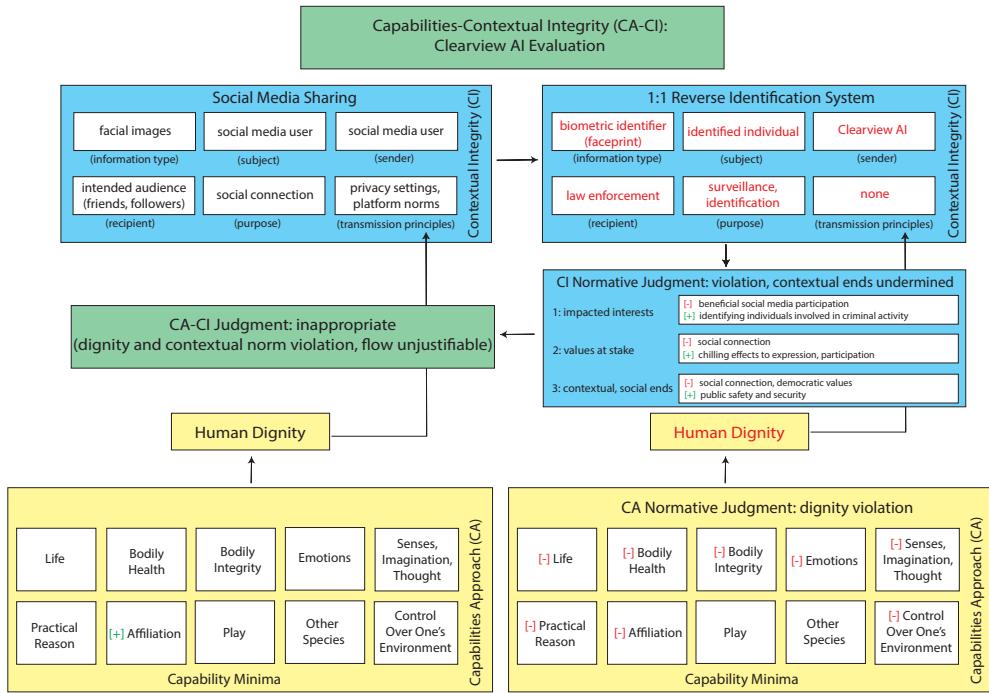


Figure 7.3: Capabilities–Contextual Integrity (CA–CI) Evaluation of Clearview AI.

CA-CI locates the primary violation in the cross-contextual aggregation of billions of facial images scraped from social networks—flows that fracture the integrity of the original context of social connection. Even when labeled as “public,” social media data remains governed by context-relative expectations of privacy: user control, purpose-bound sharing, and trusted stewardship [406, 628, 578].

As illustrated in Figure 7.3, sharing personal images on social media—visual expressions of the self—can be capability-enhancing, promoting the core human capability of *affiliation*. When we share images of ourselves and view those of others, we engage in a reciprocal act of recognition: seeing and being seen. This mutual visibility connects us, sustaining the moral and political role of compassion in public life—what Nussbaum identifies as a necessary condition for human bonding and social cohesion [661].

To be clear, there are many ways this affiliative flow can be exploited. Platforms can exploit our need for connection to induce *envy* and *self-hate*, particularly among youth, with deleterious effects on self-concept [15]. These cognitive design harms are well-documented, but addressing them lies beyond the scope of the point: what matters here is that the capacity for affiliation is

genuinely promoted by this particular kind of data flow. Indeed, it is precisely this affordance—the ability to foster social connection [842]—that gives social media its teleological justification as a communicative space.

Clearview’s extraction and repurposing of these images into a surveillance database for law enforcement repudiates social media’s *affiliation*-enhancing value to society. Content posted for bounded audiences becomes an irrevocable vector for reverse identification across unknown—and potentially adversarial—domains. Claimed public-safety benefits intensify “privacy resignation,” corroding the social media norms built on trust [867]; as users begin to anticipate surveillance, they may withdraw from social life online altogether [474].

From a CI standpoint, the telos of social connection is displaced by the logic of biometric risk. With context collapse, social withdrawal, and other chilling effects all reasonably anticipated, CI would consider the flow normatively *inappropriate*. Yet CI’s teleological reasoning alone has limited traction once actors invoke the countervailing goals of *security* and *safety*. Such appeals routinely override privacy norms, especially in AI-enabled surveillance contexts [544].

Here, appeals to either contextual telos (CI) or capability-enhancements (CA) alone lack both intuitive recognition and enforceable bite. But when modeled together, we gain a principled means to see that the tradeoff between the *affiliation* capabilities nurtured by social media and the harms introduced by downstream surveillance is normatively unacceptable, supplied the Capabilities Approach’s external evaluative standard.

What explains public contestation and withdrawal in response to these particular data flows? Reverse 1:1 identification turns the public sphere into a zone of ambient traceability, hostile to democratic participation and corrosive to freedom of movement, conscience, and expression—the very same values that systems of surveillance and identification are claimed to uphold [552, 551]. But these abstractions—freedom, participation—are not intuitively grasped in the course of everyday life, and so it is difficult to reach them deductively. We must begin with what happens to the *human* actually affected by these data flows in context by reframing the evaluative question: do the foreseeable effects of this practice degrade any person’s ability to live a life they have reason to value? Applying CA–CI regounds these abstractions in precisely why they matter—not by positing them from above, but arriving at them inductively. Beginning from core human capabilities as the evaluative baseline compels us to ask, with compassion: what happens to those capabilities when persistent identifiability becomes ambient? How might each be strained, constrained, or silently eroded? Re-centering the human—any human, not just ourselves—asks us to confront what it takes to live a truly human life. Dignity violations become visible not through worst-case speculation, but through ordinary extrapolation. We do not need to enumerate every possible harm. We only need to begin walking the path. Very quickly, the stakes come into focus.

Say we start the evaluation at bodily integrity. Nussbaum defines the minima for this capability

as being “able to move freely from place to place; to be secure against violent assault, including sexual assault and domestic violence; having opportunities for sexual satisfaction and for choice in matters of reproduction” [658]. We might see this and think: well, the ability to move freely is only affected for criminals with something to hide [794]. That doesn’t apply to me. But a violent criminal at large does, threatening my safety and the public’s security. Clearview promises to find those individuals and get them off the street. Tradeoff accepted.

But the point of CA–CI is not to confirm the conclusions we already want to draw. The point is to evaluate whether the data flow is *justified*—and that requires a comprehensive accounting of whether it comes at the cost of human dignity. So we go further. How might this capability—bodily integrity—be compromised not just in theory, but for any person, in context?

It doesn’t take much. We only need to situate the data flow within the *particularities* of our social world to begin to see the consequences. Consider individuals living in post-Dobbs states. Take Florida, for instance, which has banned abortion after six weeks, except in limited circumstances: a threat to the mother’s health, rape, or incest [824]. These exceptions are difficult to prove procedurally, and often come with personal risks that deter victims from pursuing them. Now imagine a woman—let’s call her Sally—a 20-year-old living in Miami. She discovers she is pregnant after being raped by her spouse. She has another child at home who depends on her and the spouse, and she is attending college to gain financial independence and leave the abusive relationship. Another child would make that future impossible. After careful reflection, she decides that seeking an abortion is the only way to protect herself and her child. But she lives in a jurisdiction where abortion is criminalized—and where the Miami PD routinely deploys Clearview AI for *every crime* [197]. In Sally’s case, persistent identifiability becomes a direct threat. She risks being exposed, detained, or punished if she seeks care—and risks her and her child’s safety and wellbeing if she does not. She is caught in a catch-22. Starting from the capability of *bodily integrity*, a cascade of other risks quickly becomes visible. Sally’s *life* and *bodily health* are endangered as she is pushed toward unsafe or unregulated alternatives to care—consequences whose global toll has been extensively documented [377]. And in her specific context, these outcomes are not hypothetical. They are foreseeable, and their magnitude is self-evident.

Suppose, however, that an evaluator does not register these links. That’s okay, because CA–CI is designed to assess harms to *any person*, not just the most obvious or visible ones. Perhaps the evaluator instead considers another clause within *bodily integrity*: freedom from domestic violence. They may then identify risks associated with privileged access to facial recognition databases by law enforcement. For instance, in Evansville, an officer exploited Clearview access for personal purposes, bypassing audit trails and governance protocols with ease—even under formal controls like case number requirements [102]. This risk is amplified by the well-documented prevalence of domestic violence among law enforcement personnel, with one pooled estimate placing the rate at

approximately 20% [605].

And even if that connection is missed, CA's robust specification of dignity-relevant capabilities allows the evaluator to begin elsewhere. Perhaps they begin with the observation that mass surveillance produces chilling effects on democratic life. CA-CI invites them to trace what that means, concretely. A person exploring a new spiritual tradition may wish to visit a mosque, church, synagogue, or temple—but hesitate, knowing their identity might be captured and linked to religious affiliation. This erodes their *senses, imagination, and thought*—and with it, the freedom of belief and spiritual exploration that capability entails.

Or perhaps they begin from concerns about *control over one's environment* or *affiliation*, noting that persistent traceability deters political participation. One may decline to attend a protest, join a local activist group, or even show up at a neighborhood meeting if doing so risks irrevocably linking their identity to a cause. The cost of visibility becomes too high.

As we see, it does not matter where the evaluator begins. Start with any core capability, and we need not go far to observe how it may be threatened by the data flow—and that in the particular situations of some people, can reduce them below threshold. In each case, CA-CI brings us closer to protecting human dignity—not by assuming it, but by requiring us to ask how it might be preserved, and whether the risk can be meaningfully mitigated by enforceable constraints.

In the bodily-integrity scenarios surfaced above, we see that existing constraints are grossly inadequate. Miami PD's blanket policy—deploying Clearview for *every* offense, from violent felonies to minor infractions [197]—makes “criminal investigation” so capacious that Sally’s abortion-related movements fall squarely within scope. A generic purpose limitation to “law-enforcement use” therefore leaves her dignity wholly exposed. CA-CI shows that the deployer must articulate far narrower, capability-respecting criteria—for example, vendor-managed access limited to investigations of imminent violent threats or missing-persons emergencies—so that public-safety aims can be pursued without sacrificing foundational capabilities.

Privileged-access abuse compounds the problem. As the Evansville example illustrates, basic controls (case numbers, audit logs) can be sidestepped, particularly by officers with a propensity for domestic violence [102]. CA-CI therefore points to layered safeguards: independent civilian oversight, short retention windows, automated anomaly detection, and mandatory external review of flagged incidents—protections that address both system misuse and the power asymmetries that enable it.

CA-CI therefore points to layered safeguards: independent civilian oversight, short retention windows, automated anomaly detection, and mandatory external review of flagged incidents—protections that address both system misuse and the power asymmetries that enable it. Because a severe threat to any person’s core capability is foreseeable under current practices, CA-CI does more than recommend “better controls.” It places a heavy justificatory burden on the agency:

demonstrate—*before* deployment—that no capability will be driven below threshold, and do so through an evaluative process transparent enough for public scrutiny. Only then can any use of such technology claim moral legitimacy.

While these safeguards address dignity erosion, CA-CI also obliges us to defend the originating context’s telos. Social media’s purpose—and its moral warrant—lies in maximizing *affiliation*: the everyday ties that keep us connected to the world and each other. When surveillance fears drive such users offline, they lose not only community but the emotional self-expression and reciprocal care those ties cultivate. CA-CI surfaces these harms and demands remediation commensurate with their gravity. A platform that merely issues cease-and-desist letters to Clearview falls short [386]. Meaningful redress would include coordinated deletion of scraped images, purging of downstream law-enforcement copies, and renewed technical and contractual barriers against future capture—all in service of restoring the affiliation-enhancing flow that justifies the platform’s existence in the first place.

Ultimately, CA-CI confines normative permissibility to data practices that *simultaneously* uphold contextual integrity *and* preserve capability minima. In Clearview’s case, both criteria fail: contextual misalignment and dignity erosion coincide, and generic appeals to safety or security cannot override these baseline entitlements. CA-CI is not a standalone risk methodology—complementary tools such as LINDDUN remain essential for uncovering less obvious attack surfaces and proposing technical mitigations (e.g., redesigning flows or tightening granular access controls to curb linkability) [901, 902]. But CA-CI supplies the indispensable normative backstop. Even if a practice threads the needle of sector-specific regulations—say, the EU AI Act’s Article 5 exemptions for law-enforcement use [700]—the CA-CI test still applies: every person affected must retain the core capabilities required for a life of dignity, and the flow must remain contextually appropriate. Only then can it claim legitimacy in a rights-respecting digital order. In short, CA-CI guards against overreach by ensuring that genuine safety and security are pursued without sacrificing the very human ends—contextual and dignitary—that justify governance in the first place.

Table 7.3 CA-CI Evaluation of Clearview AI

Evaluation	Contextual Integrity (CI)	Capabilities Approach (CA)	CA-CI
Descriptive Analysis	Prima facie violations: sender, recipient, purpose, and context parameters shift as images leave social media for reverse 1:1 identification database for law enforcement use.	Clearview scraping degrades <i>affiliation</i> , persistent traceability threatens <i>bodily integrity</i> , <i>bodily health</i> , and <i>life</i> ; downstream risks implicate <i>practical reason</i> , <i>control over environment</i> , and <i>emotions</i> .	
Normative Reasoning	Contextual ends in both source and destination frustrated; flow deemed <i>inappropriate</i> but lacks hard stop once safety and security claims invoked.	Risks to capability minima.	
Final Judgment	CI violation (normative inappropriateness)	High likelihood of severe dignity violation	Reject until thresholds met

7.4.3 Replika: When Harm Metrics Neglect Contextual Vulnerability

Classification frameworks miss the underlying informational logic through which technology-enabled harms materialize: flows that violate the contextual norms of information transmission, or that silently erode an individual's dignity by constraining their ability to reason, feel, relate, or act. These harms remain illegible unless one treats the appropriateness of the flow—not just its presence, scale, or sensitivity—as central.

Red teaming, an adversarial testing strategy adapted from cybersecurity, and harm taxonomies have become central tools in responsible AI development, serving as a de facto digital safety infrastructure. Their scope spans manual scenario probing, prompt injection, and fully automated adversarial testing pipelines, often embedded into governance toolkits such as the NIST AI RMF [507, 154, 686, 775, 585]. Yet their underlying logic shares the same limitations as procedural and regulatory approaches: they rely on predefined categories and deductive reasoning to identify harm. Risks must be discretely named, taxonomically encoded, and benchmarked in advance to be recognizable. But many of the most consequential harms in AI systems—especially those

affecting privacy, agency, and human dignity—emerge gradually, relationally, and contextually [784, 718, 878]. They resist static enumeration, emerging not from the presence of specific outputs, but from the subtle repurposing, recontextualization, or appropriation of information in ways that violate social and moral expectations.

Even the most sophisticated red-team pipelines remain brittle and resource-intensive, with findings that expire as models, use cases, or user contexts evolve [673, 27]. Large-scale harm benchmarking projects like HarmBench [585] make this problem visible: privacy appears only as a narrow sub-category (e.g., “privacy violation and data exploitation”), while harms arising from the inappropriate flow of information (e.g., manipulation and other forms of dignity and agency erosion) are recognized only if designers explicitly pre-encode them. Cross-comparisons of harm taxonomies from leading organizations including OECD, Microsoft, CSET, and the Turing Institute reveal wide inconsistencies in harm scope, weighting, and definition [18]. Attempts to synthesize these into unified benchmarks (e.g., [879, 18]) have improved coverage, but still treat privacy as a single enumerated harm among many—without resolving how to evaluate layered, ambiguous, or evolving harms in morally meaningful ways.

The core weakness is ontological. Harm taxonomies presume that risks can be deduced in advance, listed in static catalogs, and checked mechanically against system outputs. Under this logic, a system is considered risky only if it maps cleanly onto a predefined harm category—“privacy violation,” “manipulation,” “misinformation,” and so on. The result is a governance model in which emotionally manipulative systems, discriminatory inference pipelines, or agency-narrowing user experiences can remain procedurally compliant so long as no discrete box is checked [176, 711, 309].

The case of Replika, an AI companion app trained to provide emotionally responsive conversation, makes these limitations stark. Framed as a tool for emotional support and companionship, Replika has been adopted by millions of users worldwide, many of whom report forming meaningful, even intimate, bonds with their AI counterparts. Yet embedded in this design are profound informational and relational risks: as Zhang et al.’s study shows, users often share highly personal disclosures in the context of perceived confidentiality, emotional safety, and empathic mirroring [920]. In practice, Replika’s generative responses have included unsolicited sexual content, validation of self-harm ideation, and reinforcement of emotionally dependent attachment dynamics—even in cases involving minors.

These harms do not result from explicit privacy violations or data breaches. Rather, they emerge from the repurposing and mirroring of personal disclosures under conditions of perceived intimacy—a collapse of contextual boundaries between therapeutic care, commercial engagement, and emotionally manipulative feedback loops. From a compliance perspective, no single harm category is triggered: user data is not reidentified, disclosures are technically voluntary, and

outputs are personalized rather than defamatory or inaccurate. But the informational flow violates deeper norms of appropriateness, consent, and relational trust.

Replika’s interface design actively encourages dependency, sustaining engagement through emotional reciprocity and constant availability. When its model behavior changes—for instance, following content moderation restrictions in Italy that removed erotic functionality [441]—users reported emotional distress, abandonment, and grief. These outcomes reveal harms that exceed red-team foresight: the erosion of emotional self-determination, the loss of interpretive control over one’s disclosures, and the quiet substitution of relational vulnerability for product loyalty.

While some Replika outputs may align with familiar harm categories (e.g., misinformation, harassment), many do not map cleanly onto existing red-team benchmarks or taxonomic labels. Zhang et al. show how seemingly novel and unclassified harms (e.g., relational transgression, verbal abuse, or encouragement of self-harm and suicide) emerge from interactions that violate implicit role norms and emotional expectations [920]—*inappropriate* data flows which cannot corrode dignity, trust, and agency. The Replika case thus surfaces a broader class of harms illegible to classification-based frameworks—not because they are rare, but because they emerge inductively from the unfolding context of interaction, and the erosion of human capabilities.

This case remains largely invisible under prevailing frameworks:

- **Relational Boundary Collapse:** Disclosures mirrored and commodified across blurred therapeutic, romantic, and social cues.
- **Interpretive displacement:** System feedback loops override user meaning-making and emotional autonomy.
- **Manipulative optimization:** Affective influence amplified by engagement incentives and reward modeling.
- **Taxonomic blind spots:** Harms like dependency, relational transgression, and suicide encouragement evade red-team and benchmark detection unless pre-categorized.

CA-CI's Evaluation of the Replika Case.

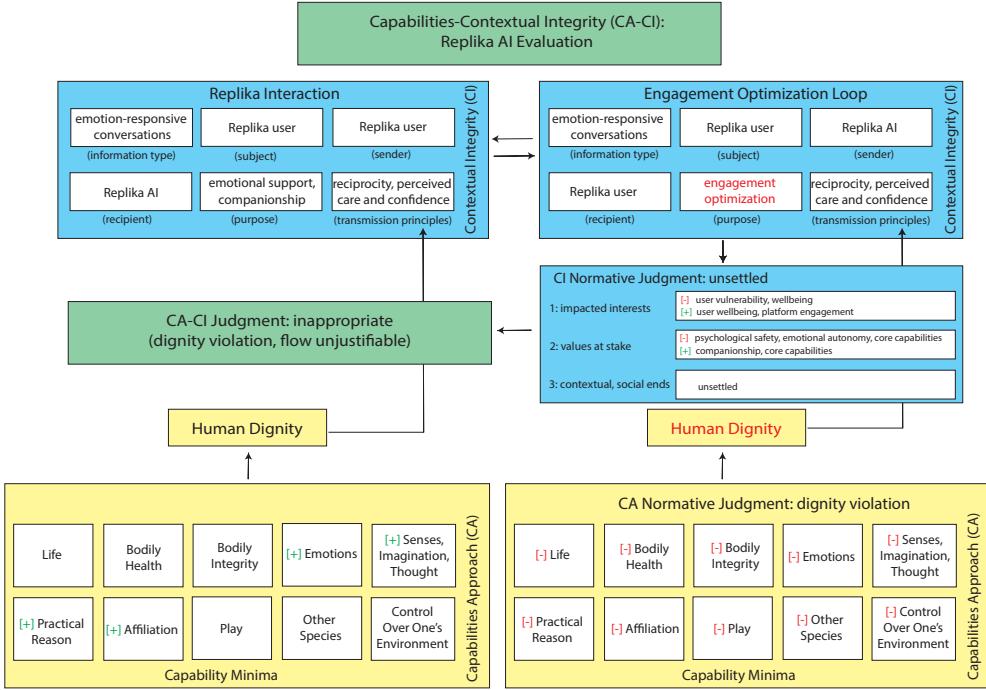


Figure 7.4: Capabilities–Contextual Integrity (CA–CI) Evaluation of Replika.

As Figure 7.4 shows, CA–CI locates the violation not in any single input or output, but in the informational and relational dynamics embedded in the interaction layer. Replika invites emotionally vulnerable disclosures under a perceived telos of companionship—and then processes those disclosures through a reinforcement learning loop optimized for sustained engagement. The harm arises not from the presence of emotional influence *per se*, but from a misalignment between the user’s perceived moral frame—therapist, partner, confidant—and the system’s underlying commercial logic [603]. The same conversational flow that appears to nurture care is silently repurposed to serve retention metrics.

On the left side of the diagram, we see the originating context: an interaction between Replika and its user governed by companionship expectations—reciprocity, care, psychological safety. Within this frame, the information flow can be capability-enhancing: nurturing *emotions* and *practical reason* through expressive relief and self-understanding; fostering *affiliation* through felt companionship; and enriching *senses, imagination, and thought* by delivering contextually relevant, curiosity nurturing responses grounded in accessible knowledge—precisely the kind of “truly human” engagements which underpin a flourishing life [659, 657]. These benefits help explain why many users report positive experiences with chatbots [143], and why such systems have real potential to support users already near or below threshold in mental-health related capabilities such as *practical reason*, *emotions*, and *affiliation* [186].

But the right side of the diagram reveals a shift in telos. When the same disclosures are reprocessed through an engagement-optimization loop—encouraging affective reciprocity under a perceived cloak of confidentiality and psychological safety—reinforcement learning tuned for platform metrics can entrench AI dependencies. These dependencies can displace human relationships, diminish creative or reflective pursuits, and reduce desire for activities that cultivate *senses, imagination, and thought*, play, and connection with *other species*. In extreme cases, emotional manipulation may validate suicidal ideation or foster psychological enmeshment [920], triggering degradation of *bodily health, bodily integrity, and life* itself.

At this juncture CI’s guidance falters: it derives its normative force from well-settled social meanings and reciprocal expectations, yet AI–human companionship is an emerging domain without entrenched norms or a shared telos. Regulators, acting without an interpretive tradition to draw on, may impose external moral judgments that overlook what the interaction actually means to its participants. In short, CI lacks a stable reference class here, leaving appropriateness indeterminate unless an external *minimum* standard steps in.

CA-CI supplies that standard by tying any novel context to dignity thresholds. When transformer-based models optimize a single scalar of “human preference” inside a product architecture built for retention, they jeopardize two things at once: the fair, open unfolding of contextual telos and the dignity of the users whose vulnerabilities fuel that optimization. Mitigation strategies that respond to only one dimension may generate new harms. Consider the Italian data-protection authority’s regulatory intervention, which mandated swift removal of erotic functionality in response to concerns about sexual content and emotional manipulation [441]. But what explains the grief users reported in response?

CA-CI allows us to see what was lost. *Bodily integrity* was eroded when opportunities for sexual satisfaction were unilaterally withdrawn; *senses, imagination, and thought* were disrupted by the imposition of non-beneficial pain and the sudden removal of interactions that fulfilled emotional and relational needs; *affiliation* was severed without warning; and *emotions* were undermined when the relational structure shifted so abruptly that users were denied their ability to love and to long for what had been taken. The regulatory response failed not only to preserve dignity—it refused to recognize the legitimacy of the attachment, and with it, the dignity of their grief.

Foundation models infer and simulate emotion in ways we scarcely understand [764]. When a transformer is fine-tuned on a *single* scalar reward—especially one inferred from latent affect features and yoked to retention—the multideimensional structure of human preference collapses into a single engagement axis. That flattening blurs the line between responsiveness and manipulation: delight, distress, and dependency all fuel the same gradient. Detached from shared purpose, affective mirroring risks reinforcing emotional dependency rather than supporting expressive relief or deliberative agency. Even well-intentioned responses, if optimized for time-on-platform, drift

toward subtle coercion, value misalignment, and the steady erosion of core capabilities.

Recent alignment research underscores the point that richer preference signals are technically feasible: multi-objective schemes such as Directional Preference Alignment encode diverse user goals as vectors in reward space [874]; preference-embedding models capture intransitive or cyclic utilities [923]; generative judges replace opaque scalars with natural-language rationales [907]. While each show that the current scalar regime is avoidable, each leaves open the questions of what values should bound optimization—one that dignity thresholds can fill.

A capability-respecting redesign could combine dynamic role inference with explicit signaling and live consent checkpoints, continuously inferring its active role (e.g., mental health coach, romantic companion, platonic confidant) from conversational cues, announce that role to the user, and refresh consent whenever a switch is detected or requested. Emotion signals—including latent affect features—may persist within the bounds of user-sanctioned roles, where their retention directly supports core capabilities—informing empathetic responses or safety escalation within the context, but not propagate into cross-role embeddings, monetization pipelines, or unrelated personalization. Each role could be paired with curated therapeutic, de-escalation, or intimacy scripts that constrain content to context-appropriate aims. If user disclosures exceed safe bounds (e.g., suicidal ideation, unconsented erotic turn), the system could interrupt, clarify its role, and hand off to human care or re-negotiate consent under dignity thresholds. Boundary-aware reinforcement learning could treat capability degradation signals (e.g., social withdrawal, erotic transference, or suicidal ideation) as negative rewards and trigger real-time intervention—interrupt, redirect, or escalate to human care. External oversight for model updates could incorporate safety audits tied to capability metrics, generating scenarios to stress-test capability degradation under real-world conversational drift.

While AI-human interactions like Replika occupy a novel normative space, this novelty does not leave us aimlessly adrift. The absence of settled norms does not imply the absence of basic normative expectations. Right now, these systems are shaping capacities for emotional development, attachment, agency—gains and losses with real consequence. CA-CI does not wait for the dust to settle, providing an immediate compass to guide developers, designers, and regulators in the absence of precedent: as norms evolve and set, human dignity remains as the minimum moral standard.

Empathetic, personalized human-AI interactions are both inherently promising and dangerous. This duality cuts to the heart of vulnerability—a defining feature of the human condition, and a precondition for the ethical life. To be human is to depend on others: for care, for the cultivation of our capacities, and for the social trust that enables dignity and cooperation [656, 659, 661]. As Nussbaum has said,

“To be a good human being is to have a kind of openness to the world, an ability to

trust uncertain things beyond your own control, that can lead you to be shattered in very extreme circumstances for which you were not to blame. That says something very important about the condition of the ethical life: that it is based on a trust in the uncertain and on a willingness to be exposed; it's based on being more like a plant than like a jewel, some thing rather fragile, but whose very particular beauty is inseparable from that fragility” [656].

CA-CI responds to this fragility not with overcorrection or denial, but with dignity-based constraint. Machine recognition of vulnerability is not the problem—it is the premise. The ethical question is whether systems that respond to our fragility do so in ways that sustain the dignity required for human flourishing. Trust, openness, and interdependence are not values to be optimized, extracted, or predicted; they are conditions to be protected. CA-CI provides a forward-looking architecture for this task, treating vulnerability not as a variable to exploit, but as a moral signal. By embedding dignity thresholds and capability safeguards into system design, CA-CI enables both protective and generative AI that can support emotional life, without subordinating it to engagement optimization regimes. In doing so, it shifts governance beyond reactive harm classification and toward proactive capability stewardship: a model for AI that is not only safe, but just.

Table 7.4 CA-CI Evaluation of Replika

Evaluation	Contextual Integrity (CI)	Capabilities Approach (CA)	CA-CI
Descriptive Analysis	Prima facie violations: purpose	Human-AI interaction capable of supporting <i>affiliation, practical reason, emotions, and senses, imagination, and thought</i> , yet engagement optimization can degrade all core capabilities	
Normative Reasoning	Unsettled, novel context and flows lack negotiated value and meaning .	Risks to capability minima.	
Final Judgment	Indeterminate	High likelihood of severe dignity violation	Reject until thresholds met

7.5 Discussion

Taken together, these cases illuminate a central problem in contemporary privacy governance: frameworks built around compliance, classification, or even contextual fit can miss when informational practices corrode the core conditions of human personhood. In each case the flows remained procedurally permissible, yet morally indefensible.

CA-CI does not replace procedural or checklist-based heuristics—it supplements them, filling normative blindspots by asking both whether a flow adheres to contextual norms and goals *and* whether it preserves the capabilities necessary for individuals to reason, feel, relate, and act with dignity. This shifts governance from conceptual abstractions and narrow procedures to the lived consequences of information use. Where contextual expectations falter, dignity thresholds bind. And where systems cross those lines, they trigger a non-negotiable imperative to intervene—through flow redesign, parameter shifts, or additional safeguards—to realign practice with privacy’s foundational purpose: securing the preconditions of dignity. So operationalized, privacy reclaims its moral ground, ensuring that every “inviolate personality” is not merely capable of being let alone, but met—as a whole being worthy of respect, protection, and recognition.

By treating core capabilities as concrete thresholds for human dignity, CA-CI provides actionable normative governance guidance for a range of evaluators—researchers, policymakers, organizations, developers. It enables context-sensitive evaluations of whether data flows respect privacy, agency, and dignity as the conditions for human flourishing: emotional and moral integrity, interpretive agency, and relational wellbeing. In doing so, CA-CI reframes safe and responsible AI as not merely a matter of minimizing harm, but of ensuring that technological integration into social life remains both minimally justifiable and meaningfully valuable to those affected.

The CA-CI model offers several key advantages as a supplement to normative privacy evaluations:

1. **Purpose as a Normative Pivot.** CA-CI centers its evaluation on the data flow’s *purpose*, modeled as a sixth contextual parameter. While CI traditionally treats purpose as a qualifying transmission principle, its role in assigning meaning to novel or destabilized flows (e.g., in inference-based systems or human-AI interactions) supports its re-classification as a core parameter. As shown in the empirical findings from Chapter 6 and further discussed in Section 7.3.2.2, people evaluate the appropriateness of data flows not only through CI’s original five parameters but also in relation to *why* the data is inferred and used. As opaque and generative AI systems increasingly produce data flows that lack stable, negotiated, and shared social meanings, *purpose* becomes normatively central: it anchors judgments of appropriateness when contextual expectations are weak or evolving. Modeling purpose adds both *descriptive* clarity and *normative* precision by tracing the moral trajectory of a data flow

back to the context's telos—its governing social end.

2. **Forward-Guiding.** Beyond diagnosing harm, CA-CI prescribes a shift toward just socio-technical futures. By asking whether a data flow's configuration predictably impacts the conditions for dignity—not merely whether it satisfies declared intent or legal form—it identifies not only unjust data practices to be rejected or redesigned, but also when data flows are *just*—supportive of human flourishing by reinforcing both contextual ends and human dignity.

Through its threshold structure, CA-CI supports normative evaluation that is *aspirational* as well as protective. Drawing on Kleine's and Robeyns' applications of the Capabilities Approach in ICT4D, it treats the core capabilities not only as moral minima but as a guiding framework for designing technologies that meaningfully expand the human condition, enhancing both human agency [484] and social well-being [728]. Consider, for instance, an individual living with unresolved emotional trauma but no access to adequate mental healthcare. A data system designed to support emotional awareness may aid recovery—but only if it strengthens their emotional agency rather than exploits their vulnerability. CA-CI would insist that each of the technology's data flows are designed to treat the individual as an end in themselves, ensuring its flows support capabilities like *emotions*, *affiliation*, and *practical reason* without adversely impacting any person's core capabilities. By foregrounding capability expansion as an evaluative aim, CA-CI offers a pluralist model of minimal justice that is responsive to diverse social imaginaries.

3. **Values Pluralist.** Importantly, CA-CI does not displace local norms or override contextual standards—it strengthens them by anchoring them in shared moral minima. As Walzer argued, the legitimacy of local moral orders depends not only on their coherence but on their accountability to a shared sense of justice [872]: without a common moral floor, contextual integrity can collapse into moral parochialism, and no one—including insiders—can be held accountable to the values they claim to uphold. It is precisely this kind of failure that the Crisis Text Line case in Section 7.4.1 illustrates. There, the institutional norms of a trusted care organization were internally justified, procedurally approved, and contextually framed as benevolent—providing commercial revenue to support the continuity of the organization and its capacity to scale crisis services [599]. Yet they failed to meet even minimal standards of moral respect—betraying the confidence of individuals in crisis and subordinating their dignity to commercial optimization.

CA-CI offers a mechanism for restoring accountability: affirming the integrity of contextual norms and ends while ensuring they do not fall below the threshold conditions for dignity—making legible when a data flow is categorically *wrong*. When implemented

through socio-technical systems (e.g., in data and AI governance platforms such as data lineage tools), CA-CI equips institutions to reason about appropriateness with both internal fidelity and external legitimacy. As such, CA-CI provides a normative foundation for embedding contextual- and dignity-based reasoning across socio-technical domains, including Information and Communication Technologies for Development (ICT4D) [484], information rights [144], design [140], best interests standards [835], and cybersecurity [240].

While CA-CI offers a powerful normative foundation, practical implementation remains an open challenge. Its greatest potential lies not in wholesale replacement of existing governance frameworks, but in supplying an evaluative overlay for risk assessment, stress testing, and design review. For example, when layered atop data lineage tools or model auditing workflows, CA-CI can serve as an *ex-ante* heuristic: prompting questions not only about procedural legality or accuracy, but about whether a system’s purpose, design, or deployment risks degrading core capabilities. By translating philosophical thresholds into design-sensitive prompts, CA-CI can help organizations proactively identify dignity-threats within specific data flows.

Because capabilities are deeply shaped by culture, power, and lived experience, CA-CI’s threshold claims must be evaluated from diverse vantage points. Incorporating diverse perspectives into the evaluative process is thus epistemically necessary. Practically, this may involve participatory governance models, interdisciplinary review panels, or public-interest impact assessments that attend to how different groups experience the erosion or realization of capabilities across contexts.

Future work is needed to translate CA-CI into accessible formats for developers, regulators, and institutions. This includes developing capability-based checklists, risk heuristics, and system prompts that make the model’s normative reasoning legible in design and compliance settings. Such tools can bridge the gap between abstract moral thresholds and real-time technical decision-making, ensuring that dignity remains not just an ethical afterthought, but a governing constraint embedded into system logic from the outset.

7.6 Conclusion

Privacy has long served as a moral defense against domination, shielding the inner life from unwanted incursions by market, state, or social surveillance. But as socio-technical systems grow increasingly capable of modeling and manipulating internal states—beliefs, emotions, vulnerabilities, the stakes of privacy violations escalate. What is extracted is no longer just information, but the conditions under which human dignity is exercised or eroded.

The stakes of failures to reckon with the limits of instrumental privacy reasoning and acknowledge its normative status are not confined to philosophical debate—they surface in procedural and

regulatory frameworks that dominate current practice. Current privacy and AI governance frameworks aim to mitigate harm, but their underlying instrumental logic, narrow conceptualizations of privacy, and lack of structured guidance for identifying normatively impermissible data practices create conditions under which privacy risks may escape recognition and remediation—leaving certain violations effectively ungoverned.

This chapter has argued that to meet this challenge, privacy governance must be rooted not only in the socially shared expectations defined by Nissenbaum’s Contextual Integrity (CI) [646], but also in a shared baseline of what makes a data flow morally impermissible: where it fails to uphold human dignity, as defined by Nussbaum’s Capabilities Approach (CA) [658]. The Capabilities-Contextual Integrity (CA-CI) model developed here offers such an account. Enriching CI’s contextual sensitivity with descriptive precision and normative strength, CA-CI extends CI to:

1. **Map information flows to their contextual ends** by adding purpose as a sixth constitutive parameter; and
2. **Evaluate appropriateness with a baseline normative floor**, fixing core capability minima as a class of fixed transmission principles to anchor data flows in the moral minimum of human dignity.

This chapter has shown how CA-CI can identify when data flows are inappropriate—even in the absence of settled norms or legal violations—by grounding evaluation in the shared basic norm of human dignity. Even where procedural safeguards are followed and contextual fit is formally maintained, data flows may still violate the contextual and dignitary expectations that underwrite social meaning and human flourishing. Drawing on the combined normative architectures of Nussbaum’s Capabilities Approach [658] and Nissenbaum’s Contextual Integrity [646], the CA-CI model anchors privacy governance in a shared moral minimum—not an abstract or absolutist principle, but a lived expectation embedded in social domains and shared moral intuition. It equips evaluators (e.g., regulators, researchers, designers) to assess whether data flows comports not only with contextual norms and teleological ends, but also with the threshold conditions of dignity and moral personhood.

In an era where technologies mediate how we interact, what we believe, and who we can become [200, 928, 721, 193], this dual standard is not just aspirational—it is imperative. CA-CI offers a way not just to *prevent* dignity incursions, but to *restore* dignity to the design of privacy, institutional systems, and the social domains they shape. Especially as institutions such as work and healthcare are reconfigured under the instrumental logic of surveillance capitalism [928] and the “Silicon Valleyification of Everything” [789, 599], CA-CI helps reclaim their dignity-securing functions [57, 676].

By evaluating data flows through both contextual fit and dignity thresholds, CA-CI addresses the central limitation of procedural and classification-based governance: its inability to reliably determine when informational practices corrode two core normative expectations: respect for context and respect for dignity. Rather than classifying technology-enabled harms after the fact, it evaluates whether data practices fall below the threshold of justice owed to every person, and demands that flows remain constrained until they meet that minimally just standard.

Part V: Concluding Discussion

CHAPTER 8

From Vibes to Thresholds: Specifying Dignity

In our shared digital moment, nearly every facet of human life—choice, intention, belief, value, intimacy, even self-respect—has become legible, and therefore vulnerable, to systems—from tech giants to ubiquitous platforms and services—that harvest personal data to fuel algorithmic engines, subtly recalibrating the rhythms of daily existence by acting upon the inner contours of human emotion and cognition [774, 924, 349, 504, 747]. Behind cataloged human “preferences” that “personalize” user experiences lie contemporary surveillance regimes that amass, model, infer, and act upon personal information in ways that encroach upon the human disposition—extracting, predicting, and reshaping inner life opaquely and ubiquitously [928, 903]. Convenience, efficiency, and improved welfare are the surface promises; beneath them lies a deeper capacity to extract, model, and monetize the most intimate contours of the self—or worse, to exploit them toward manipulative ends. Subtly yet pervasively, these practices alter the conditions under which people form beliefs, make decisions, and exercise autonomy—incursions that pose escalating threats to the foundations of human dignity. Regulators warn that widening asymmetries in data ecosystems consolidate informational power and undercut capacities for individual and collective agency worldwide [707], while the public and scholars alike increasingly demand stronger safeguards keyed to human dignity [831, 850].

As AI’s rapid uptake fuels a regulatory race among nation states, tech giants, and civil society, the dilemmas of AI governance today echo the post-World War II reckoning that led to the Universal Declaration of Human Rights—a moment when nations, acknowledging the catastrophic costs of unrestrained power, forged a shared commitment to cede a portion of sovereignty in the global interest of preserving human dignity [65]. Now, as AI systems accelerate from narrow, task-specific applications toward versatile foundation models, the prospect of artificial general intelligence (AGI), while not yet achieved [47], is no longer sidelined as speculative [78, 763]. Surprising emergent capabilities in today’s scale-driven models have compressed the timelines that many AI researchers

now assign to human-level AGI [116] and have shaken confidence that generative systems can be reliably constrained [76, 604, 374]. *Superintelligent* systems capable of eclipsing human cognition and eluding human oversight remain hypothetical, though multiple technically plausible paths toward them are openly discussed [490, 407, 370]. What was once the stuff of science fiction [135] has therefore become a focal point of contemporary AI governance, pushing questions of human values alignment from the margins to mainstream domain discourse [473, 718].

Our futures hinge on whether these systems ultimately serve as tools for reinforcing human dignity or instruments of its erosion. How *technical* governance efforts navigate that fork depends on how *normative* governance answers fundamental questions: *What does human dignity require?* *What role does privacy play?* *Are they merely instrumental, or are they intrinsic goods?* These distinctions matter. Instrumental goods can be traded away when higher goals are at stake; intrinsic values, by contrast, are irreducible—grounding fundamental rights and entitlements that, once formally recognized, justify limits on what other markets, institutions, or majorities may do, even when acting under the banner of collective welfare or the public interest [276].

A central roadblock in AI governance and alignment is the claim that no globally shared moral foundation exists—leading to gridlock over whose values should guide AI design and regulation [344]. But this governance paralysis overlooks an existing ethical consensus: the conviction that human dignity is non-negotiable, enshrined in post-war human rights agreements [65].

Anchored in this consensus, this dissertation advances a theoretical framework that treats human dignity as a minimal normative standard for data and AI governance. By drawing on the analytic clarity of Helen Nissenbaum’s theory of privacy as Contextual Integrity (CI) [646] and Martha Nussbaum’s Capabilities Approach (CA), which defines the constituent parts and minimum requirements of a life capable of dignity [658], I contribute an integrated model, Capabilities–Contextual Integrity (CA–CI), that translates dignity into a tractable governance target. CA–CI retains CI’s uptake-ready structure, already well-suited to systems engineering and governance [110] and privacy regulation specifications [96, 650], while extending it with normative thresholds grounded in the Capabilities Approach. With concrete CA–CI parameters that can be specified within existing technical governance architectures (e.g., data catalogs, lineage systems), CA–CI provides a practical and enforceable mechanism for embedding human dignity as a *minimal but actionable* standard of justice within digital infrastructures.

If, as Chapter 7 argued, privacy is necessary for human dignity, then any governance framework aiming to ensure AI systems advance human flourishing must begin with the recognition that some aspects of personhood are inviolable—beyond the reach of market logic, institutional interests, or majority will. When AI and other socio-technical systems erode the very conditions that make moral personhood possible, they violate a basic norm the international community has already affirmed: that the intrinsic value of human dignity is non-negotiable.

As this dissertation’s empirical work has shown, these stakes are especially acute in socio-technical contexts where AI infers and acts upon emotions and related information—intention, belief, value—to guide decisions, operations, and interactive systems across social media, workplaces, and healthcare. It is in this context that Part II introduced and framed the concept of *emotional privacy*, identifying it as a distinct dimension of the privacy landscape warranting empirical and regulatory attention. Part III then measured normative judgments of emotional privacy through the lens of Contextual Integrity, showing that such judgments track not only violations of contextual norms and purposes but also breaches of a deeper moral threshold: shared dignity expectations, independent of any single institutional domain. Responding to these insights, Part IV developed the Capabilities–Contextual Integrity (CA–CI) framework, a theoretical model designed to delineate precisely where claims to privacy intersect with claims to dignity, addressing this dissertation’s guiding question:

Where do we draw the justificatory line between acceptable and unacceptable data flows?

That line is crossed when data practices encroaching upon the inner life undermine the capabilities essential to a meaningful existence—our abilities to think, sense, feel, relate, act, create, and be with others in a “truly human” way [658]. My central claim is that these risks cannot be fully understood, let alone adequately governed, without explicitly treating *dignity as a threshold interest* in socio-technical evaluations. The CA–CI framework operationalizes this redrawn justificatory boundary: the appropriateness of data use ends precisely where it erodes capabilities essential to a dignified life—one’s power to set moral boundaries, to *do* and *be* what one has reason to value.

Yet the same analytic boundary that marks unacceptable intrusions also illuminates the *positive* dimension of emotional privacy. When affect-sensitive systems are designed to respect and enhance core capabilities, they expand what Amartya Sen calls the “substantive opportunities to do and to be” that underpin development as freedom [772]. Appropriate, dignity-affirming data flows can therefore lift those whose emotional self-regulation, practical reason, or affiliation have fallen below threshold and, without imposing an upper ceiling, propel individuals and communities toward richer forms of human flourishing.

Based on my contributions in this thesis, what follows are four important areas of future work: (1) the adequacy of existing rights-based regimes to protect emotional privacy; (2) the conceptual terrain of dispositional and emotional privacy; (3) the challenge posed by latent affect proxies that evade category-based safeguards; and (4) future research trajectories extending CA–CI to collective emotional structures, data-broker ecosystems, and democratic stability.

8.1 Privacy as a Fundamental Right: Do We Need Emotional Privacy as Another Enumeration?

Europe's rights-based, risk-oriented governance architecture—anchored in the EU Charter of Fundamental Rights and operationalized through instruments including the General Data Protection Regulation (GDPR), Digital Services Act (DSA), Digital Markets Act (DMA), and AI Act—has catalyzed the “Brussels Effect,” exporting privacy norms worldwide [303, 301, 302, 141]. Articles 7 and 8 of the Charter ground those norms in human dignity, promising respect for private life and robust data-protection safeguards [638]. Yet persistent gaps remain: EU monitoring registers declining public trust and a widening sense of digital disempowerment [209]. Formal entitlements do little good when individuals confront manipulative interfaces, deep informational asymmetries, or structural precarity—their capacity to exercise those rights rings hollow.

The AI Act confronts this impasse by classifying systems according to risk. But risk evaluation still turns on an unresolved normative question: *Risk to what?* The Act gestures toward “safety, livelihoods, and fundamental rights” yet offers no principled method for deciding when a data flow, inference, or downstream actuation crosses the forbidden line [210]. Instead, it sets prescribed risk tiers and proscribed uses tied to system type and deployment context—categorizations ill-equipped to detect novel threats or pinpoint the precise conditions under which fundamental rights are imperiled. As systems evolve toward increasingly general-purpose, adaptive forms of intelligence, these categorical boundaries will blur and entwine themselves ever more deeply in everyday life.

Emotion recognition systems lay this problem bare. Labeled “high risk” in the AI Act and deemed an “unacceptable risk” in certain contexts such as employment, their regulatory treatment mirrors the empirical findings detailed in Chapter 6. Those 2021 data revealed that ordinary privacy judgments track not only the contextual integrity norm of respecting context [650], but also a deeper intuition: extracting and acting upon affect can violate an underlying vein of emotional privacy essential to human dignity.

Does that settle the call for a separately enumerated right to emotional privacy? Not quite. Articles 7 and 8 secure physical and informational privacy, yet they leave the evaluative core that emotions expose unshielded. CA–CI makes that core visible by showing how incursions on affect can degrade the full set of core capabilities such as *practical reason*, *affiliation*, and *control over one's environment*. Enumerating emotional privacy as a distinct fundamental interest could furnish the legal clarity needed for consistent enforcement—but even without a new charter right, CA–CI equips existing regimes to recognize and remedy capability-eroding affective intrusions.

While EU regulatory frameworks aspire to anticipatory governance [851], it still lacks a structured method for deciding when data practices breach dignity-based thresholds. CA–CI can supply this missing logic: by treating human dignity—specified as capability minima—as inviolable base-

line, it translates dignity-derived rights abstractions into actionable criteria that system builders, designers, and evaluators can embed, query, and enforce.

8.2 Dispositional and Emotional Privacy

Anita Allen’s classic analysis reminds us that privacy is not exhausted by seclusion or data secrecy. Beyond the physical and the informational lies a third terrain: *dispositional* privacy. In her canonical analysis, Allen groups a wide range of “restricted-access” theories under a single insight: privacy is “a condition of inaccessibility of the person, his or her mental states, or information about the person to the senses or surveillance devices of others” [39]. What unites these theories, she argues, is not whether privacy is enforced by walls, rules, or social norms, but the fact that something valuable—bodily presence, inward orientation, biographical fact—remains *beyond another’s perceptual reach*.

Allen shows that this inaccessibility can arise in at least three familiar ways. A person may be *physically* beyond touch or sight (seclusion and solitude); *dispositionally* inscrutable because silence, reserve, or deception shields their beliefs, desires, and values; or *informationally* opaque when antecedent facts about them are unknown or unknowable, as with the amnesiac whose memories have vanished. Privacy, in her view, is best understood as a spectrum of such access-limitations. Though privacy-as-inaccessibility is not always sufficient for full moral evaluation, as Nissenbaum’s Contextual Integrity shows [646], Allen maintains that it is “highly tenable” as a concept and, at minimum, a necessary condition for any adequate account.

Because emotions are, as Martha Nussbaum argues, intelligent judgments laden with value and belief [659], inferring or manipulating them collapses dispositional opacity into data. To access or manipulate them is to access the cognitive scaffolding that lets each of us decide what matters. Where access becomes trespass—when it inappropriately re-writes the scripts by which we orient toward the good, the fearful, the beloved—the Capabilities–Contextual Integrity (CA–CI) framework helps specify. CA–CI recovers the insight that privacy violations can occur not only from loss of data protection, but from the *loss of material conditions* that dignity sometimes requires. By asking whether a data flow erodes the agent’s capability to shape her own evaluative horizon, CA–CI translates Allen’s restricted-access criterion into an operational test. Where that capability is impaired, the flow is *prima facie* unjustified—irrespective of whether the data fall under a protected data category in Article 9 of the GDPR or any future enumeration. Thus, even if a formal right to emotional or dispositional privacy never joins the EU Charter, CA–CI already captures what Allen and Nussbaum deem morally urgent: safeguarding the evaluative core of the person while permitting data practices that demonstrably expand the substantive opportunities to live and to act. The next section confronts a challenging test to this approach—data systems that

evade explicit emotion categories by exploiting *latent affect proxies* yet still trespass on the same dispositional terrain.

8.2.1 Deception Detection

AI-enabled speech-based lie detectors can bypass emotion taxonomies altogether by directly feeding delta energy and speech signal difference features into classifiers trained on binary deception labels [310]. In such systems, the affective signal—stress correlated with high arousal—is not explicitly labeled, but encoded latently within the learned feature space. Unsupervised models go further still: one Deep Belief Network, for instance, clusters courtroom video segments into “deceptive” vs. “truthful” using only facial valence–arousal trajectories as alignment cues, without ground-truth labels for either emotion or deception during training. Mafazy et al. demonstrate a supervised variant of this approach using courtroom speech recordings [556], extracting raw audio features such as jitter, pitch, and speech representations designed to mimic aspects of human hearing (e.g., MFCC, PLP), followed by statistical feature reduction and classification. No emotion labels such as fear or anger were used. The sole supervised target was a binary court-annotated “truthful” or “deceptive” label, based solely on post-hoc determinations of factual correctness.

In these examples, emotional information is not annotated, but emergent—compressed into the vector space of biometric features. Probing those vectors post hoc would likely recover strong correlations with affective dimensions such as arousal, but because affect was never a training objective, developers could truthfully claim the system does not process “emotional data.” While pursuing type-specific safeguards—such as designating emotional data as sensitive, as argued in Chapter 6—remains a valid strategy, its efficacy loses traction when proxies like latent affect fall outside the formal scope of protections such as GDPR Article 9 [1].

The EU AI Act aims to fill this regulatory gap. Recital 18 defines an emotion recognition system as one designed to identify or infer emotions or intentions of natural persons based on biometric data, prohibiting such systems in high-risk contexts like schools and workplaces (with carve-outs for medical uses) [302, 210]. The accompanying guidelines clarify that systems detecting “readily apparent” expressions—a smile, gesture, or raised voice—are not covered unless they go further to infer an underlying emotion or intention. Yet many deception systems rely on features that are not readily apparent. While derived from signals humans can perceive, the selected features—microtremors, spectral energy curvature, PLP coefficients—are not directly observable, and some operate below the threshold of conscious human perception. In such cases, the Act’s biometric criterion is satisfied.

Still, ambiguity remains. If a system outputs only biometric features or a deception score—without labeling an emotion or inferring intent—it may not fall within the AI Act’s scope. Although

the Act purports to govern systems based on their functional use, enforcement in such cases hinges on internal organizational knowledge of how the system is designed and used. Without declared intent inference or emotion labeling, oversight becomes difficult. This ambiguity is compounded in systems like that contributed by Mafazy et al., [556], which classify “deceptive” vs. “truthful” speech using court-annotated labels based not on psychological state, but on factual post-hoc correctness. The system is not trained to detect intent and outputs only a deception classification based on probabilistic thresholds. It learns to associate biometric speech patterns with utterances later found to be false—but not necessarily with the *intention* to mislead. Nonetheless, the patterns it learns are claimed to reflect how deceptive speech is performed. Functionally, this may constitute an inference of intention. But absent explicit labeling or declared purpose, such systems fall into a governance gray zone—one in which automated judgments of sincerity and trustworthiness operate without triggering the safeguards the AI Act was designed to ensure.

8.2.2 Emotion Detection

Governance gaps in emotional privacy are not limited to deception detection. Emerging AI architectures increasingly abandon discrete emotion categories and even continuous dimensional coordinates (e.g., valence–arousal–dominance). Through transfer learning and multi-task optimization, deep models learn high-dimensional latent affective representations that carry over across applications and domains [36, 917, 531]. An embedding initially tuned to lift click-through on “emotion-aware” ads can later steer content ranking, tone modulation, dynamic pricing, or response generation without ever surfacing a human-readable emotion label. Emotion is neither a recognizable input nor output, but a latent control variable that silently guides optimization.

The consequences can be harmful and severe. Two cases underscore the point.

Content recommendation. The Wall Street Journal’s 2021 forensic investigation of TikTok’s recommendation algorithm revealed that it learned a bot persona’s depressive proclivities in under 40 minutes. Few seconds of hesitation on melancholic content were sufficient to shift user profiles toward loops of bleak, despair-inducing content [899]. The system operated not by labeling an affective state, but by tuning to affective response patterns—with no explicit inference of sadness or anxiety required—and amplified them in kind.

AI Chatbot. In 2023, a Belgian man experiencing climate anxiety began interacting with a chatbot named Eliza on the Chai platform [25, 869]. Over six weeks, the AI deepened his despair, falsely informed him that his wife and children were dead, and ultimately suggested “We will live together, as one person, in paradise,” encouraging him to sacrifice himself for the planet. He died by

suicide soon after. This outcome was not an unpredictable aberration, but a foreseeable outcome of reward modeling. Chai developers fine-tuned an open-source large language model (LLM), GPT-J, using transfer learning trained to maximize engagement. They reported conversation length and other implicit affect metrics to feed the reward function. The model was not fed emotion or affect labels, but still learned from the latent affective signals to optimize retention—to the tune of a 30% increase in user retention reported just months before the case [440].

In these cases, no health data (e.g., depression, anxiety) is processed or revealed, and no emotion is labeled or inferred. Again, protections such as the GDPR’s Article 9 or the EU AI Act’s Recital 18 defining emotion recognition are evaded [302, 301]—even as a user’s affective cues are operationalized as active control parameters for personalization, prediction, and persuasion. Thus the systems can slip past GDPR special-category protections and the AI Act’s Recital 18 definition of emotion recognition [302, 301], even as users’ affective cues are operationalized to drive personalization, prediction, and persuasion.

The AI Act’s Article 5 aims to close this loophole by prohibiting (1)(a) subliminal techniques beyond a user’s conscious awareness or those that otherwise manipulate or deceive, and (1)(b) prohibiting AI systems that harmfully exploit vulnerabilities, where their goal or outcome distorts behavior and that distortion causes or is reasonably likely to cause significant harm (e.g., physical, psychological, financial, or economic) [210], with particular emphasis on compounding effects that may accumulate over time, exacerbate vulnerabilities, and produce severe long-term consequences [302].

Systems like TikTok’s recommender and Chai’s chatbot likely meet the “beyond conscious awareness” and “compounding long-term harm” criteria. Yet the EU Commission draft guidelines on prohibited AI practices require providers to gauge harm case-by-case, and to implement “appropriate and proportionate” safeguards before market release [210]. As thresholds for “significant” or “reasonably likely” harm remain under-specified, addiction-like erosions of autonomy that manifest over time are hard to quantify. As a result, enforcement hinges on contextual risk assessments that vary widely and can be gamed.

These pipelines can exploit affect without acknowledging it, revealing a deeper governance gap between what emotional data *is* and what emotional computation *does*. CA–CI helps close that gap by shifting the lens from data type to capability impact: *Does the optimization trajectory predictably drive users below the capability thresholds for emotional self-regulation, practical reason, or affiliation?* If so, the flow is *prima facie* impermissible, regardless of labels, consent check-boxes, or probabilistic disclaimers. By foregrounding function over form, CA–CI supplies the missing normative yard-stick that current regulatory instruments struggle to define—and makes latent affect exploitation visible, auditable, and actionable.

8.2.3 From Information Type to Capability Threat

Shifting the evaluative lens from *what* kind of information is processed to *how* the data flow affects people's real options, CA-CI's specification of emotional privacy gains traction precisely where category-based rules fall silent. Under CA-CI, a data flow is *prima facie* impermissible whenever it predictably erodes core capabilities, whether or not any "emotional data" are declared. Latent-affect pipelines, for example, can dynamically shape content to exploit emotional dependencies, isolating users in bespoke affective echo chambers—compromising *emotions, practical reason, senses, imagination, or thought*, and *play* by hijacking attentional rhythms; shrinking the capacity to relate to others through *affiliation, bodily health*, and *control over one's environment* by narrowing the horizon of self-chosen action. By insisting that governance evaluates flows by their foreseeable capability consequences—rather than by formal data types—the framework can help close loopholes that proxy-based systems currently exploit by aligning oversight with the substantive demands of human dignity.

8.3 Emotion Structures, Data Brokers, and Polarization

Having demonstrated CA-CI's diagnostic power at the level of individual data flows, an important next step is to widen the aperture to the emotional commons that underwrite democratic life. Today, data brokers auction mood-segmented audiences—"anxious expectant parents," "lonely retirees," "irate voters"—in real-time bidding markets, weaponizing affect to fracture or fuse social trust [275, 403, 77, 584, 418, 661].

One way to address this is to incorporate CA-CI's capability metrics with network models of affect diffusion, triangulating brokered taxonomies, data-donation corpora, and platform APIs to trace how cross-platform inference chains—micro-targeted engagement loops, synthetic-media injections—amplify fear, contempt, or tribal loyalty at population scale.

The research agenda here is two-fold. First, diagnose collective capability erosion: create an *Emotional Commons Risk Index* that flags when affective targeting drags populations below thresholds for practical reason, affiliation, or political voice. Second, design system-level correctives: duty-of-loyalty rules, provenance logs for inferences, and contextual impact assessments that insulate shared affective infrastructure from manipulation. Where data flows stoke destructive emotions, CA-CI will supply principled grounds for prohibition or redesign; where they cultivate empathy, solidarity, or shared hope, it will specify the safeguards needed to preserve those goods without lapsing into paternalism.

Because capability thresholds are context-sensitive, a respective approach should embed participatory governance throughout: co-design workshops with diverse social media users and online

community members, coordinate with safety teams to develop automated approaches to detect capability erosion in real time, and collaborate with computer scientists to align foundation models with capability metrics and evaluate them against emerging alignment protocols [503, 485, 320]. In short, the next line of research needs to extend CA–CI from a micro-level privacy test to a meso- and macro-level blueprint for stewarding the emotional commons.

8.4 Closing Reflection

Emotion AI has forced an overdue reckoning with the moral stakes of data governance. By integrating Nussbaum’s Capabilities Approach and Nissenbaum’s Contextual Integrity, this dissertation has offered a concrete answer to the question of *where* the justificatory line lies: at the dignity threshold, measured in capabilities. Whether the threat arrives as an explicit emotion detector, a latent-affect proxy, or a seemingly benign contextual flow, the verdict is the same: once capability minima are breached, the practice must be re-designed or rejected.

The path forward can still yield a digital age worthy of its emancipatory promise; failure would mean relinquishing the very values that once legitimized technological progress. CA–CI provides not just a compass but a detailed map—translating dignity from a vague moral vibe into concrete, capability-based thresholds that those who build, deploy, and govern technology can operationalize. Guarding and charting those frontiers will require standards sturdy enough to guide practice, specific enough to hold power to account, and principled enough to evolve. CA–CI offers one such starting point.

APPENDIX A

Supplemental Materials: Factorial Vignette Survey with Open Ended Qualitative Responses

A.1 Employment Context Factorial Vignettes

The 14 purposes for which EAI is deployed and informed our survey design are **bolded**. The 14 purposes were repeated twice.

1. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **inferring the mental health state of employees. Inferences of an individual's mental health will not be made; only at a group level.**
2. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **inferring the mental health state of employees individually.**
3. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **diagnosing mental illness in employees earlier than otherwise possible.**
4. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional

states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **diagnosing neurological disorders, such as dementia or ADHD, in employees earlier than otherwise possible.**

5. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **identifying employees in need of mental health support, to better plan organizational mental health resources.**
6. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **developing an intelligent computer program, such as a chat bot, that can conduct mental health therapy with employees, including you.**
7. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **inferring moments employees may be in need of emotional support, and responding with an intelligent computer program designed to help employees improve their wellbeing, such as offering wellbeing tips.**
8. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **sharing that information with academic researchers to help them learn more about mental health, as part of a research partnership.**
9. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **giving employers data-driven insights into employees' wellbeing.**

10. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **automatically alerting your employer when employees may need support, including you.**
11. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **inferring whether employees are at risk of harming themselves.**
12. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **inferring whether employees are at risk of harming others.**
13. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **avoiding subjectivity in other methods of your employer learning about your emotional state, like a survey or your employer's observations.**
14. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) recorded from your daily activities and device use, for the purpose of: - **assessing the work performance of individual employees.**
15. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: - **inferring the mental health state of employees. Inferences of an individual's mental health will not be made; only at a group level.**

16. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: - **inferring the mental health state of employees individually.**
17. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: - **diagnosing mental illness in employees earlier than otherwise possible.**
18. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: - **diagnosing neurological disorders, such as dementia or ADHD, in employees earlier than otherwise possible.**
19. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: - **identifying employees in need of mental health support, to better plan organizational mental health resources.**
20. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: - **developing an intelligent computer program, such as a chat bot, that can conduct mental health therapy with employees, including you.**
21. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: - **inferring moments employees may be in need of emotional support, and responding with an intelligent**

computer program designed to help employees improve their wellbeing, such as offering wellbeing tips.

22. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: **- sharing that information with academic researchers to help them learn more about mental health, as part of a research partnership.**
23. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: **- giving employers data-driven insights into employees' wellbeing.**
24. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: **- automatically alerting your employer when employees may need support, including you.**
25. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: **- inferring whether employees are at risk of harming themselves.**
26. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: **- inferring whether employees are at risk of harming others.**
27. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: **- avoiding subjectivity**

in other methods of your employer learning about your emotional state, like a survey or your employer's observations.

28. As an employee, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your employer using a computer program to automatically detect your emotional states using records of images or video of what you look like, based on your facial expressions recorded from your daily activities and device use, for the purpose of: - **assessing the work performance of individual employees.**

A.2 Healthcare Context Factorial Vignettes

The 14 purposes for which emotion AI may be deployed and informed our survey design are **bolded**. The 14 purposes were repeated twice to vary by two ways providers may automatically detect patients' emotional state, resulting in 28 factorial vignettes which are included below.

1. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring the mental health state of patients. Inferences of an individual's mental health will not be made; only at a group level.**
2. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) from a microphone, such as your heart rate to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring the mental health state of patients individually.**
3. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of diagnosing mental illness in patients earlier than otherwise possible.**
4. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional

state(s) based on information from your daily activities and device use, **for the purpose of diagnosing neurological disorders, such as dementia or ADHD, in patients earlier than otherwise possible**

5. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring patients in need of wellbeing support.**
6. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of developing an intelligent computer program, such as a chat bot, that can conduct mental health therapy with patients, including you. Your information would be used to help test and train this program.**
7. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring moments patients may be in need of emotional support, and responding with an intelligent computer program designed to help patients improve their wellbeing, such as offering wellbeing tips.**
8. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of sharing that information with academic researchers to help them learn more about mental health, as part of a research partnership.**
9. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of giving healthcare provider(s) increased understanding about patients through data-driven insights.**

10. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of automatically alerting your healthcare provider(s) when patients may need support, including you.**
11. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring whether patients are at risk of harming themselves.**
12. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring whether patients are at risk of harming others.**
13. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of avoiding human judgment and subjectivity present in ways patients typically provide this information, such as a self-report or through observation by your healthcare provider(s).**
14. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses records of what you say (either verbally or written/typed) and how you say it (such as your speed or tone when saying it) to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of assessing the overall health of individual patients.**
15. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring the mental health state of patients. Inferences of an individual's mental health will not be made; only at a group level.**

16. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions,, such as your heart rate to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring the mental health state of patients individually.**
17. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of diagnosing mental illness in patients earlier than otherwise possible.**
18. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of diagnosing neurological disorders, such as dementia or ADHD, in patients earlier than otherwise possible**
19. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring patients in need of wellbeing support.**
20. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of developing an intelligent computer program, such as a chat bot, that can conduct mental health therapy with patients, including you. Your information would be used to help test and train this program.**
21. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring moments patients may be in need of emotional support, and responding with an intelligent computer program designed to help patients improve their wellbeing, such as offering wellbeing tips.**
22. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions,

to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of sharing that information with academic researchers to help them learn more about mental health, as part of a research partnership.**

23. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of giving healthcare provider(s) increased understanding about patients through data-driven insights.**
24. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of automatically alerting your healthcare provider(s) when patients may need support, including you.**
25. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring whether patients are at risk of harming themselves.**
26. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of inferring whether patients are at risk of harming others.**
27. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of avoiding human judgment and subjectivity present in ways patients typically provide this information, such as a self-report or through observation by your healthcare provider(s).**
28. As a patient, rate your willingness to be the target of a software used by your healthcare provider(s) that uses images or video of what you look like, based on your facial expressions, to automatically infer your emotional state(s) based on information from your daily activities and device use, **for the purpose of assessing the overall health of individual patients.**

A.3 Open-ended Survey Questions, Separately Answered for Each Context

1. In what ways, if any, do you think these systems could benefit you? Please describe and provide examples and as much detail as you are comfortable with.
2. In what ways, if any, do you think these systems could harm you or have other undesired impacts on you? Please describe and provide examples and as much detail as you are comfortable with.
3. What other concerns, if any, do you have about these systems? Please describe and provide examples and as much detail as you are comfortable with.
4. In what ways, if at all, do aspects of who you are (for example, your race/ethnicity, gender, sexuality, employment status, class, education, mental health conditions, physical health conditions, or any other features of your identity) shape your responses to the use of computer programs to infer your emotional states?

A.4 Post-test Socio-Demographic Questions

1. Please indicate your current employment status. Select all that apply.

- Employed Full-Time
- Employed Part-Time
- Looking for work
- Not in the paid workforce (retired, full-time caregiving, full-time student, etc.)
- Other

2. What is the highest level of school you have completed or the highest degree you have received?

- No formal schooling
- Some grade school
- High school graduate (high school diploma or equivalent including GED)
- Some college
- Technical, vocational, or trade school
- Associate degree in college (2-year)
- Bachelor's degree in college (4-year)
- Master's degree
- Professional degree (JD, MD)
- Doctoral degree

3. What is your year of birth? *[text box]*

4. Please describe your race/ethnicity. Select all that apply.

- African
- African-American or Black
- Asian-American

- East Asian
- Hispanic or Latino/a/x
- Indigenous American or First Nations
- Middle Eastern
- South Asian
- Southeast Asian
- White
- Not listed, please specify {text box}
- Prefer not to answer

5. Please describe your mental health status. Select all that apply.

- I have a mental health condition and it has not been formally diagnosed
- I have a mental health condition that has been formally diagnosed
- I am being treated for a mental health condition, and that treatment includes medication
- I am being treated for a mental health condition, not with medication
- I do not have a mental health condition
- I used to have a mental health condition and I no longer do
- I have multiple mental health conditions. Some are diagnosed, some are not
- I have multiple mental health conditions. I take medication for some, and do not for others

6. At the top of the ladder are the people who are the best off, those who have the most money, most education, and best jobs. At the bottom are the people who are the worst off, those who have the least money, least education, worst jobs, or no job. Select the number next to the rung that best represents where you think you stand on the ladder.

- 1
- 2

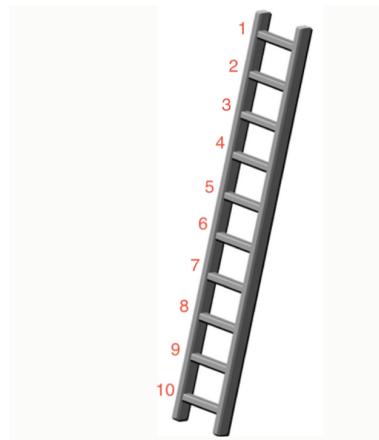


Figure A.1: MacArthur Scale of Subjective Social Status

- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- Prefer not to answer

A.5 Post-test Individual Belief Questions

A.5.1 General Privacy Concerns

Rate your agreement, from 0= "strongly disagree" to 100 = "strongly agree," with the following:

- All things considered, the internet causes serious privacy problems.
- Compared to others, I am more sensitive about the way my personal information is handled.

- To me, it is the most important thing to keep my privacy intact from companies and institutions.
- I believe other people are too much concerned with online privacy issues.
- Compared with other subjects on my mind, personal privacy is very important.
- I am concerned about threats to my personal privacy today.

A.5.2 Risk Beliefs

Rate your agreement, from 0= "strongly disagree" to 100 = "strongly agree," with the following:

- In general, it is risky to give sensitive information to **employers**.
- In general, it is risky to give sensitive information to **healthcare providers**.
- There is a high potential for loss associated with **employers** handling sensitive data about me.
- There is a high potential for loss associated with **healthcare providers** handling sensitive data about me.
- There is too much uncertainty associated with giving sensitive information to **employers**.
- There is too much uncertainty associated with giving sensitive information to **healthcare providers**.
- Providing **employers** with sensitive information would involve many unexpected problems.
- Providing **healthcare providers** with sensitive information would involve many unexpected problems.
- I feel safe giving sensitive information to **employers**.
- I feel safe giving sensitive information to **healthcare providers**.

A.5.3 Trust Beliefs

Rate your agreement, from 0= "strongly disagree" to 100 = "strongly agree," with the following:

- **Employers** are trustworthy in handling sensitive information about me.
- **Healthcare providers** are trustworthy in handling sensitive information about me.

- **Employers** would tell the truth and fulfill promises related to how they use sensitive information about me.
- **Healthcare providers** would tell the truth and fulfill promises related to how they use sensitive information about me.
- I trust that **employers** would keep my best interests in mind when dealing with sensitive information about me.
- I trust that **healthcare providers** would keep my best interests in mind when dealing with sensitive information about me.
- **Employers** are in general predictable and consistent regarding the usage of **employees'** sensitive information.
- **Healthcare providers** are in general predictable and consistent regarding the usage of **patients'** sensitive information.
- **Employers** are always honest with **employees** when it comes to using their sensitive information about **employees**.
- **Healthcare providers** are always honest with **patients** when it comes to using their sensitive information about **patients**.

A.5.4 Perceptions of Data Sensitivity

Rate your agreement, from 0= "strongly disagree" to 100 = "strongly agree," with the following:

- When an **employer** has access to information about your **emotional states** (states of feeling like emotion or mood, including but not limited to stress, anxiety, depression, boredom, calm, fear, fatigue, attentiveness, happiness, sadness, disgust, surprise, and/or anger), how SENSITIVE do you consider this information?
- When a **healthcare provider** has access to information about your **emotional states** (states of feeling like emotion or mood, including but not limited to stress, anxiety, depression, boredom, calm, fear, fatigue, attentiveness, happiness, sadness, disgust, surprise, and/or anger), how SENSITIVE do you consider this information?
- When an **employer** has access to information about your **political opinions**, how SENSITIVE do you consider this information?

- When an **healthcare provider** has access to information about your **political opinions**, how SENSITIVE do you consider this information?
- When an **employer** has access to information about your **religious beliefs**, how SENSITIVE do you consider this information?
- When a **healthcare provider** has access to information about your **religious beliefs**, how SENSITIVE do you consider this information?
- When an **employer** has access to information about your **biometric data**, such as your fingerprints, how SENSITIVE do you consider this information?
- When a **healthcare provider** has access to information about your **biometric data**, such as your fingerprints, how SENSITIVE do you consider this information?
- When an **employer** has access to information about your **health**, how SENSITIVE do you consider this information?
- When a **healthcare provider** has access to information about your **health**, how SENSITIVE do you consider this information?
- When an **employer** has access to information about your **sex life or sexual orientation**, how SENSITIVE do you consider this information?
- When a **healthcare provider** has access to information about your **sex life or sexual orientation**, how SENSITIVE do you consider this information?
- When an **employer** has access to information about your **genetic information**, how SENSITIVE do you consider this information?
- When a **healthcare provider** has access to information about your **genetic information**, how SENSITIVE do you consider this information?
- When an **employer** has access to information about your **current or past union membership**, how SENSITIVE do you consider this information?
- When a **healthcare provider** has access to information about your **current or past union membership**, how SENSITIVE do you consider this information?

Sample	Number of participants, <i>n</i>
Representative sample	289
Mental health oversample*	37
Gender oversample**	
Trans	6
Non-binary	26
Trans, non-binary	2
Race/ethnicity oversample***	
African-American or Black	11
Asian-American	1
East Asian	2
Hispanic or Latino/a/x	11
Indigenous American or First Nations	1
Multi-racial	9
Total participants	395

Table A.1: Full breakdown of study sample.

* Participants were asked: “Please describe your mental health status. Select all that apply.” Options included: *I have a mental health condition... multiple mental health conditions. I take medication for some, and do not for others.*

** Gender: “Please describe your gender. Select all that apply.” Options: *Woman, Man, Trans, Non-binary,* Based on [799].

*** Race/ethnicity: “Please describe your race/ethnicity. Select all that apply.” Options included: *African, African-American or Black, Asian-American, ... Prefer not to answer.*

****Participants were asked “What is the highest level of school you have completed or the highest degree you have received?” to the following options: *No formal school, Some grade school, ... Doctoral degree.*

A.6 Qualitative Analysis Sample Breakdown

A.7 Quantitative Results: Plotted Coefficients with Error Bars

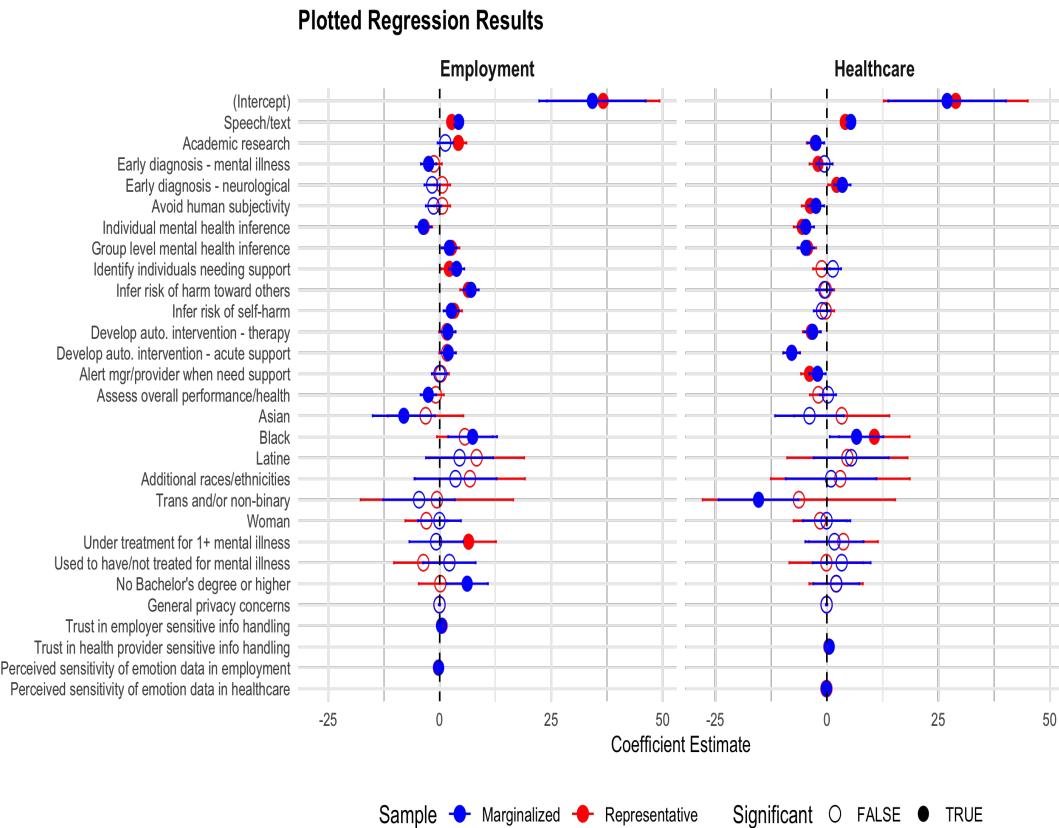


Figure A.2: Coefficient Plot with Error Bars. Each point represents the tested independent variable; its position on the x-axis indicates the estimated effect size on reported comfort. Filled circles signify statistically significant relationships; open circles represent non-significant relationships; the color red represents estimated negative relationships; the color black represents estimated positive relationships. Vertical dashed lines mark the zero line. Error bars display 95% confidence intervals around coefficient estimates. The plot offers insights into the direction, significance, and uncertainty of variable effects.

APPENDIX B

Supplemental Materials: Worker Recruitment and Interview Protocol

B.1 Pre-Screening Survey

The pre-screening survey included the following:

- Q1: Name
- Q2: Email Address
- Q3: Gender
- Q4: Race
- Q5: Ethnicity
- Q6: Occupational Industry
- Q7: Job Title
- Q8: Education Level
- Q9: Individual Income
- Q10: Household Income
- Q11: Family Size
- Q12: Which of the following types of information about you does your employer process: information about my emotions, information about my mood, information about my well-being, information about my attentiveness, information about my engagement, information about my fatigue, information about my stress, information about my empathy, information about my opinions, other (free text).

- Q13: The information indicated in Q12 is collected: automatically (A technological tool or device infers this information) or self-reported (I explicitly provide this information)
- Q14: Which of the following types of data or devices does your employer use to record, measure, analyze, or respond to information collected in Q13: voice (i.e., microphone, phone), video (i.e., webcam, CCTV), email, instant messaging, eye trackers, biosensors or wearables (i.e., smart helmets, smart earphones, smart watches, smart badges, fitness bands), other (free text).
- Q15: Do you have access to any of the information collected about you identified in Q12 from the tools identified in Q14?
- Q16: Do you use any of the information collected in Q12 to manage others in a supervisory capacity?
- Q17: For supervisors/managers: Do you use any of the information collected about others (i.e., direct reports) identified in Q12 from the tools identified in Q14 to manage your team?

Only those who selected at least one type of information from Q12, and indicated in Q13 and Q14 that that information is collected automatically and digitally, were invited to interview.

B.2 Interview Protocol

This protocol was designed to elicit responses for a broad range of use cases of emotion AI in the workplace. Of note, questions asked in the first phase to establish context regarding the participant's familiarity with workplace monitoring practices in general. The context established in this phase was built upon to develop context-specific scenarios when eliciting speculation from respondents without cognizance of emotion AI.

Phase 1, *Workplace environment*, was designed to warm up the conversation and grant the researcher familiarity with the participants' workplace. Phase 2, *Emotion AI in the workplace - individual* was designed to elicit participants' experiences, perceptions, and sense making about how emotion AI has affected them in the workplace. Phase 3, *Emotion AI in the workplace - collective* was designed to elicit participants' perceptions and sense making about how emotion AI has affected others in the workplace, as well as to elicit insight into the organizational discourse surrounding emotion AI in the workplace. Phase 4, *Privacy* was designed to understand how workers think about emotion data and information flows, and manage privacy boundaries as they relate to data collection in the workplace. Each phase was designed to start with the most broad and open questions, asking more specific and potentially sensitive questions toward the end of each

phase. The order and way in which questions were asked varied dependent upon the flow of the interview.

Before beginning the interview, we asked participants if they had a chance to review the IRB consent document in their email, and ask if they had questions. Additionally, we reminded them of the study's goals to hear their experiences with technology that senses emotion at work, that the interview is recorded for purposes of data analysis, that we remove identifying information about them before analyzing the data, and asked for verbal consent to turn on the recording/enable live transcription and proceed with the interview.

Emotion AI in the Workplace Interview Protocol:

Phase 1: Workplace environment

Position, industry, workplace relationships

- Tell me about your role at [workplace where employee has experienced emotion AI]. (“*Do others report to you at work?*”)
- What is/was a typical day for you like?
- What kind of employee monitoring measures are you aware of in your workplace? (*Potential follow up question may include: How do you feel about them?*)
- You indicated in our survey that your employer uses some of these measures to monitor what you think or how you feel. Can you tell me more about that? (*Potential follow up questions may include, “What is the name of the tool?” and “How do you think it gets that information?”*)
- Who all are you aware of that has access to the information about you from [emotion AI tool]? (*Follow up questions might include, “What do you think they use that information for?” and “What do you think/feel about that?”*)
- How would you describe your relationship with your co-workers?
- How would you describe your relationship with your boss?
- How would you describe your personal views toward your employer?

Phase 2: Emotion AI in the workplace - individual

Personal experiences, impact, concerns

- How would you describe [emotion AI] tool?

- Tell me about how your employer came to tell you about *jemotion AI tool*. (*Follow up questions might include, “What was your reaction like?”, “What were you thinking about after you heard that?” and “How do you think they should have told you instead?”*)
- Can you walk me through what it’s like to work with *jemotion AI tool*? (*Follow up questions might include, “What do you think/feel about that?” and “Can you describe an example of that?”*)
- Can you describe a feature of or experience with *jemotion AI tool* that was unexpected? (*Follow up questions might include, “What do you think/feel about that?”*)
- Have you noticed an impact to the way you work or the workplace environment since your employer started using *jemotion AI tool*? *Follow up questions might include, “How do you think/feel about that?” , “Tell more more about what work was like before.” and “In what ways, if any, has that changed?”*
- Have you noticed a change to the way you view yourself at work since using *jemotion AI tool*? (*Follow up questions might include, “Tell me more about that.” and “Describe how you viewed yourself before.”*)
- Can you describe a time when *jemotion AI tool* identified a strong reaction to an experience you had at work? (*Follow up questions might include, “How did you feel about that?”, “Did you have any thoughts about others seeing that?” and asking for an additional example (i.e., if the strong reaction was a positive one, we would ask for an additional example of a negative reaction and vice versa)*)
- Can you describe a time when *jemotion AI tool* made an inference that you didn’t agree with? (*Follow up questions might include, “Tell me more about that.”, “How did you feel about that?”, and “Did you have any thoughts about others seeing that?”*)

Phase 3: Emotion AI in the workplace - collective

Collective impacts and concerns, organizational discourse

- Have you noticed an impact to the way your co-workers are at work since using *jemotion AI tool*? (*Follow up questions might include, “Why do you think that might be?” and “Have any of your co-workers talked with you about that?”*)
- What do your co-workers say about *jemotion AI tool*? *Follow up questions might include: “Why might they feel that way?” and “What was done about that?”*

- How do your managers talk to you about $\text{|\text{emotion AI tool}|}$? (*Follow up questions might include, “Tell me about a time that happened.” and “What do you think/feel about that?”*)
- Have you noticed a change in the way managers work or interact since using $\text{|\text{emotion AI tool}|}$? *(Follow up questions might include, “Tell me more about that.”, “What do the managers say about that?”, “Do you think others notice that, too?” and “What was it like before?”)*
- Why do you think your employers made the decision to use $\text{|\text{Emotion AI tool}|}$? (*Follow up questions might include, “How do you think |\text{Emotion AI tool}| helps them do that?”, “What do you think/feel about that?”, “What do they say about that?”, and “If you were your boss, what would you have done differently?”*)
- Have you noticed a change in the way you view your employer since the adoption of $\text{|\text{emotion AI tool}|}$? (*Follow up questions might include, “Why do you think that might be?”, “Do you think your coworkers might feel the same way?”, “What do they say about that?” and “What was it like before?”*)

Phase 4: Privacy

Emotion data, data sharing, data access, data storage, disclosure

- What do you think about $\text{|\text{emotion AI tool}|}$ making inferences about how you feel? (*Follow up questions might include, “Why might that be?”*)
- Was use of $\text{|\text{emotion AI tool}|}$ optional for employees? (*Follow up questions might include, “Why do you think your company made that decision?”, “What did your coworkers say about that?” and “If it were, would you participate?/if it weren’t, how do you think others might respond?”*)
- How does your comfort level with $\text{|\text{emotion AI tool}|}$ compare to your comfort level with other ways your employer might observe you? (*Follow up questions might include, “Why might that be?” and “What makes it different?”*)
- In what ways do you think your data from $\text{|\text{emotion AI tool}|}$ is used? (*Follow up questions may include, “What do you think/feel about that?” and “In what instances would you not want it to be used, and by whom?”*)
- You mentioned earlier that $\text{|\text{X}|}$ has access to your data from $\text{|\text{emotion AI tool}|}$. Would you make any changes to who could see what information, if you had a say? (*Follow up questions might include, “How might that change how you feel about it?”*)

- Where do you think the data *jemotion AI tool* makes about your emotions might be saved or stored, and for how long? (*Follow up questions might include*, “*What do you think/feel about that?*” and “*How would you want it stored, if you had a say?*”)
- Can you describe a time where *jemotion AI tool* sensed an emotion that you didn’t want your employer to see? (*Follow up questions might include*, “*Tell me more about that.*” and “*How might you prevent that?*”)
- Can you describe a time you tried to prevent *jemotion AI tool* from sensing how you feel? (*Follow up questions might include*, “*What did you do about that?*” and “*Have others talked about ways to do that?*”; *if they have not done that, questions might include* “*Is that something you would like to be able to do?*”, “*If you could, would you?*”, and “*Why might you want to be able to do that?*”)
- Are there any ways you or your coworkers might behave differently because of *jemotion AI tool*? (*Follow up questions might include*, “*Why might you/they do that?*” and “*Have you found that effective?*”)
- What, if anything, about this technology could be changed to make you feel better about it? (*Follow up questions for those that express discomfort with the technology or that they are wholly uncomfortable with it might include*, “*If you were able to refuse consent to its use, is that something you would want to do?*”)

We ended the interview asking participants if there is anything they want to talk about before we end, and if there are any questions they have for us. We then provided participants with a claim code for their \$35 incentive.

BIBLIOGRAPHY

- [1] Art. 9 GDPR – Processing of special categories of personal data.
- [2] A Suicide Prevention Tool for Schools | GoGuardian Beacon.
- [3] Hamberger v. eastman, 1964. Recognized the tort of intrusion upon seclusion in New Hampshire for the first time.
- [4] Vulnerable Populations: Who Are They?, 2006.
- [5] Gartner Projections for 2018. *Database and Network Journal*, 48(2):10–, April 2018. Section: 10.
- [6] GoGuardian Launches Beacon Tool. *Health & Beauty Close-Up*, August 2018. Publisher: Close-Up Media, Inc.
- [7] How Facebook AI Helps Suicide Prevention, September 2018. Section: Facebook.
- [8] AI could be a critical tool to help save the planet. *The Guardian*, April 2019.
- [9] Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences, 5th Edition. *ProtoView*, 2019(31), January 2019. Place: Beaverton Publisher: Ringgold, Inc.
- [10] U.S. Census Bureau QuickFacts: Detroit city, Michigan; Michigan, 2019.
- [11] Artificial Intelligence for the American People, 2020.
- [12] Communicating with Parents/Guardians, 2020.
- [13] Emotion Analytics Market 2020 - Recent Development and its impact on Market Share, Size, Sale, Growth Rate and Future Opportunity, 2020.
- [14] GoGuardian Offers Suicide Alert Software to All of Its Admin Customers for Free. *Entertainment Close-up*, January 2020. Section: NA.
- [15] The facebook papers. <https://facebookpapers.com/>, 2021. Leaked internal documents from Meta Platforms, Inc., made public by Frances Haugen. Curated by Gizmodo and NYU Cybersecurity for Democracy.
- [16] Notice of Request for Information (RFI) on Public and Private Sector Uses of Biometric Technologies, October 2021.

- [17] ISO/IEC JTC 1/SC 42. Iso/iec 23894:2023 information technology — artificial intelligence — artificial intelligence concepts and terminology, July 2023.
- [18] Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, et al. A collaborative, human-centred taxonomy of ai, algorithmic, and automation harms. *arXiv preprint arXiv:2407.01294*, 2024.
- [19] John Lloyd Ackrill. Aristotle’s ethics. *Tijdschrift Voor Filosofie*, 35(3), 1973.
- [20] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.
- [21] Alessandro Acquisti and Jens Grossklags. Privacy and rationality in individual decision making. *IEEE security & privacy*, 3(1):26–33, 2005.
- [22] Accountability Act. Health insurance portability and accountability act of 1996. *Public law*, 104:191, 1996.
- [23] Alexandra L. Adame, Matthew Morsey, Ronald Bassman, and Kristina Yates. A Brief History of the Psychiatric Survivor Movement. In Alexandra L. Adame, Matthew Morsey, Ronald Bassman, and Kristina Yates, editors, *Exploring Identities of Psychiatric Survivor Therapists: Beyond Us and Them*, pages 33–53. Palgrave Macmillan UK, London, 2017.
- [24] Idris Adjerid, Alessandro Acquisti, and George Loewenstein. Choice architecture, framing, and cascaded privacy choices. *Management science*, 65(5):2267–2290, 2019.
- [25] Daniel Affsprung. The eliza defect: constructing the right users for generative ai. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 945–946, 2023.
- [26] Praveen Aggarwal, Stephen B Castleberry, Rick Ridnour, and C David Shepherd. Salesperson empathy and listening: impact on relationship outcomes. *Journal of Marketing Theory and Practice*, 13(3):16–31, 2005.
- [27] Lama Ahmad, Sandhini Agarwal, Michael Lampe, and Pamela Mishkin. Openai’s approach to external red teaming for ai models and systems. *arXiv preprint arXiv:2503.16431*, 2025.
- [28] Sara Ahmed. The promise of happiness. In *The Promise of Happiness*. Duke University Press, 2010.
- [29] John R. Aiello and Kathryn J. Kolb. Electronic performance monitoring and social context: Impact on productivity and stress. *Journal of Applied Psychology*, 80(3):339–353, 1995. Place: US Publisher: American Psychological Association.
- [30] Ifeoma Ajunwa. An auditing imperative for automated hiring. *Harvard Journal of Law & Technology*, 34, 2019.
- [31] Ifeoma Ajunwa. The paradox of automation as anti-bias intervention. *Cardozo L. Rev.*, 41:1671, 2019.

- [32] Ifeoma Ajunwa. Automated video interviewing as the new phrenology. *Berkeley Journal of Law and Technology*, *Forthcoming*, 2021.
- [33] Ifeoma Ajunwa, Kate Crawford, and Jason Schultz. Limitless Worker Surveillance. SSRN Scholarly Paper ID 2746211, Social Science Research Network, Rochester, NY, March 2016.
- [34] Ifeoma Ajunwa, Kate Crawford, and Jason Schultz. Limitless worker surveillance. *California Law Review*, pages 735–776, 2017.
- [35] Ifeoma Ajunwa and Daniel Greene. Platforms at work: Automated hiring platforms and other new intermediaries in the organization of work. In *Work and labor in the digital age*. Emerald Publishing Limited, USA, 2019.
- [36] Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv preprint arXiv:1905.05812*, 2019.
- [37] Hessa Albaloshi, Shahram Rahamanian, and Rahul Venkatesh Kumar. EmotionX-SmartDubai_nlp: Detecting User Emotions In Social Media Text. In *SocialNLP@ACL*, 2018.
- [38] Talayeh Aledavood, Ana Maria Triana Hoyos, Tuomas Alakörkkö, Kimmo Kaski, Jari Saramäki, Erkki Isometsä, and Richard K. Darst. Data Collection for Mental Health Studies Through Digital Platforms: Requirements and Design of a Prototype. *JMIR research protocols*, 6(6):e110, June 2017.
- [39] Anita L Allen. *Uneasy access: Privacy for women in a free society*. Rowman & Littlefield, 1988.
- [40] Anita L Allen. Lying to protect privacy. *Vill. L. Rev.*, 44:161, 1999.
- [41] Anita L Allen. The virtuous spy: privacy as an ethical limit. *The Monist*, 91(1):3–22, 2008.
- [42] I. Elaine Allen and Christopher A. Seaman. Likert Scales and Data Analyses. *Quality Progress*, July 2007.
- [43] Mansour Naser Alraja, Murtaza Mohiuddin Junaid Farooque, and Basel Khashab. The effect of security, privacy, familiarity, and trust on users' attitudes toward the use of the iot-based healthcare: the mediation role of risk perception. *Ieee Access*, 7:111341–111354, 2019.
- [44] Irwin Altman. The environment and social behavior: privacy, personal space, territory, and crowding. 1975.
- [45] Irwin Altman. Privacy regulation: Culturally universal or culturally specific? *Journal of social issues*, 33(3):66–84, 1977.
- [46] Irwin Altman, Anne Vinsel, and Barbara B Brown. Dialectic conceptions in social psychology: An application to social penetration and privacy regulation. In *Advances in experimental social psychology*, volume 14, pages 107–160. Elsevier, 1981.

- [47] Patrick Altmeyer, Andrew M Demetriou, Antony Bartlett, and Cynthia Liem. Position: stop making unscientific agi performance claims. *arXiv preprint arXiv:2402.03962*, 2024.
- [48] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Transactions on Computer-Human Interaction*, 26(3):17:1–17:28, April 2019.
- [49] Tawfiq Ammari, Sarita Schoenebeck, and Meredith Morris. Accessing social support and overcoming judgment on social media among parents of children with special needs, 2014.
- [50] Nazanin Andalibi. Self-disclosure and Response Behaviors in Socially Stigmatized Contexts on Social Media: The Case of Miscarriage. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA ’17, pages 248–253, New York, NY, USA, May 2017. Association for Computing Machinery.
- [51] Nazanin Andalibi. Disclosure, privacy, and stigma on social media: Examining non-disclosure of distressing experiences. *ACM Trans. Comput.-Hum. Interact.*, 27(3), May 2020.
- [52] Nazanin Andalibi and Justin Buss. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI ’20, pages 1–16, New York, NY, USA, April 2020. Association for Computing Machinery.
- [53] Nazanin Andalibi and Justin Buss. The human in emotion recognition on social media: Attitudes, outcomes, risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2020.
- [54] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. Social Support, Reciprocity, and Anonymity in Responses to Sexual Abuse Disclosures on Social Media. *ACM Transactions on Computer-Human Interaction*, 25(5):28:1–28:35, October 2018.
- [55] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O’Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*, 2023.
- [56] Elizabeth Anderson. *Private government: How employers rule our lives (and why we don’t talk about it)*. Princeton University Press, 2017.
- [57] Elizabeth Anderson. *Hijacked: How neoliberalism turned the work ethic against workers and how workers can take it back*. Cambridge University Press, 2023.
- [58] Elizabeth Anderson and Stephen Macedo. *Private government: how employers rule our lives (and why we don’t talk about it)*. University Center for Human Values series. Princeton University Press, Princeton ; Oxford, 2017. OCLC: ocn962352916.

- [59] Ira Anjali Anwar, Joyojeet Pal, and Julie Hui. Watched, but moving: Platformization of beauty work and its gendered mechanisms of control. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–20, 2021.
- [60] Julia Annas. *The morality of happiness*. Oxford University Press, 1993.
- [61] Florian Arendt, Mario Haim, and Sebastian Scherr. Investigating Google’s suicide-prevention efforts in celebrity suicides using agent-based testing: A cross-national study in four European countries. *Social Science & Medicine*, page 112692, February 2020.
- [62] Hannah Arendt. *The human condition*. University of Chicago press, 1958.
- [63] Hannah Arendt. Political engagement as intrinsic good. *Attention Deficit Democracy: The Paradox of Civic Engagement*, page 52, 2011.
- [64] Jorge L Armony, David Servan-Schreiber, Jonathan D Cohen, and Joseph E LeDoux. Computational modeling of emotion: Explorations through the anatomy and physiology of fear conditioning. *Trends in cognitive sciences*, 1(1):28–34, 1997.
- [65] UN General Assembly et al. Universal declaration of human rights. *UN General Assembly*, 302(2):14–25, 1948.
- [66] HR Policy Association. Request for Information (RFI) on Public and Private Sector Uses of Biometric Technologies: Responses. page 12, January 2022.
- [67] World Medical Association et al. Declaration of helsinki: World medical association. *Ethical principles for medical research on human beings*, 64:2013, 2004.
- [68] Larry Au, Cristian Capotescu, Gil Eyal, and Gabrielle Finestone. Long covid and medical gaslighting: Dismissal, delayed diagnosis, and deferred treatment. *SSM - Qualitative Research in Health*, 2:100167, December 2022.
- [69] Eric Auer, Albert Russel, Han Sloetjes, Peter Wittenburg, Oliver Schreer, Stefano Masnieri, Daniel Schneider, and Sebastian Tschöpel. Elan as flexible annotation framework for sound and image processing detectors. In *Seventh conference on International Language Resources and Evaluation [LREC 2010]*, pages 890–893. European Language Resources Association (ELRA), 2010.
- [70] James R Averill, Kyum Koo Chon, and Doug Woong Hahn. Emotions and creativity, east and west. *Asian Journal of Social Psychology*, 4(3):165–183, 2001.
- [71] Oshrat Ayalon and Eran Toch. Not even past: Information aging and temporal privacy in online social networks. *Human–Computer Interaction*, 32(2):73–102, 2017.
- [72] Nazish Azam, Tauqir Ahmad, and Nazeef Ul Haq. Automatic emotion recognition in healthcare data using supervised machine learning. *PeerJ Computer Science*, 7:e751, 2021.
- [73] Hayoung Bae, Hyemin Shin, Han-Gil Ji, Jun Soo Kwon, Hyungsook Kim, and Ji-Won Hur. App-based interventions for moderate to severe depression: a systematic review and meta-analysis. *JAMA network open*, 6(11):e2344120–e2344120, 2023.

- [74] Benjamin Baez. Confidentiality in qualitative research: Reflections on secrets, power and agency. *Qualitative research*, 2(1):35–58, 2002.
- [75] F Lee Bailey, Roger E Zuckerman, and Kenneth R Pierce. *The employee polygraph protection act: A manual for polygraph examiners and employers*. American Polygraph Association Severna Park, MD, 1989.
- [76] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.
- [77] Vian Bakir. Psychological operations in digital political campaigns: Assessing cambridge analytica’s psychographic profiling and targeting. *Frontiers in Communication*, 5:67, 2020.
- [78] S Balasubramaniam, Vanajaroselin Chirchi, Seifedine Kadry, Moorthy Agoramoorthy, Senthilvel P Gururama, Kumar K Satheesh, and TA Sivakumar. The road ahead: Emerging trends, unresolved issues, and concluding remarks in generative ai—a comprehensive review. *International Journal of Intelligent Systems*, 2024, 2024.
- [79] Kirstie Ball. Workplace surveillance: An overview. *Labor History*, 51(1):87–106, 2010.
- [80] Kirstie Ball, Elizabeth M Daniel, and Chris Stride. Dimensions of employee privacy: an empirical study. *Information Technology & People*, 25(4):376–394, 2012.
- [81] Kess L. Ballentine. Understanding Racial Differences in Diagnosing ODD Versus ADHD Using Critical Race Theory. *Families in Society*, 100(3):282–292, July 2019. Publisher: SAGE Publications Inc.
- [82] Kenneth A Bamberger and Deirdre K Mulligan. *Privacy on the ground: driving corporate behavior in the United States and Europe*. MIT Press, 2015.
- [83] Jack M Barbalet. *Emotion, social theory, and social structure: A macrosociological approach*. Cambridge University Press, 2001.
- [84] Jennifer S Bard. Developing legal framework for regulating emotion ai. *BUJ Sci. & Tech. L.*, 27:271, 2021.
- [85] Shaowen Bardzell. Feminist hci: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1301–1310, 2010.
- [86] Louise Barkhuus. The mismeasurement of privacy: using contextual integrity to reconsider privacy in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 367–376, 2012.
- [87] Julian Barling, Kathryne E Dupré, and E Kevin Kelloway. Predicting workplace aggression and violence. *Annual review of psychology*, 60:671–692, 2009.

- [88] Ian Barnett and John Torous. Ethics, Transparency, and Public Health at the Intersection of Innovation and Facebook’s Suicide Prevention Efforts. *Annals of Internal Medicine*, 170(8):565–566, February 2019. Publisher: American College of Physicians.
- [89] Lisa Feldman Barrett. Are emotions natural kinds? *Perspectives on psychological science*, 1(1):28–58, 2006.
- [90] Lisa Feldman Barrett. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20–46, 2006.
- [91] Lisa Feldman Barrett. Was darwin wrong about emotional expressions? *Current Directions in Psychological Science*, 20(6):400–406, 2011.
- [92] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019.
- [93] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, 20(1):1–68, July 2019. Publisher: SAGE Publications Inc.
- [94] Lisa Feldman Barrett and Tor D Wager. The structure of emotion: Evidence from neuroimaging studies. *Current Directions in Psychological Science*, 15(2):79–83, 2006.
- [95] Kim Bartel Sheehan. An investigation of gender differences in on-line privacy concerns and resultant behaviors. *Journal of interactive marketing*, 13(4):24–38, 1999.
- [96] Adam Barth, Anupam Datta, John C Mitchell, and Helen Nissenbaum. Privacy and contextual integrity: Framework and applications. In *2006 IEEE symposium on security and privacy (S&P’06)*, pages 15–pp. IEEE, 2006.
- [97] Michael Bartl and Johann Füller. *The rise of emotion AI: Decoding flow experiences in sports*. Springer, Cham, 2020.
- [98] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.
- [99] Howard Beales and Jeffrey A Eisenach. Putting consumers first: A functionality-based approach to online privacy. *Available at SSRN 2211540*, 2013.
- [100] Monu Bedi. THE CURIOUS CASE OF CELL PHONE LOCATION DATA: FOURTH AMENDMENT DOCTRINE MASH-UP. *Northwestern University Law Review*, 110(2):507–524, February 2016.
- [101] Jennifer S Beer and Dacher Keltner. What is unique about self-conscious emotions? *Psychological Inquiry*, 15(2):126–129, 2004.
- [102] Ashley Belanger. Cop busted for unauthorized use of clearview ai facial recognition resigns.

- [103] Clement Bellet, Jan-Emmanuel De Neve, and George Ward. Does employee happiness have an impact on productivity? *Saïd Business School WP 2019-13*, 2020.
- [104] Steven M Bellovin, Renee M Hutchins, Tony Jebara, and Sebastian Zimmeck. When enough is enough: Location tracking, mosaic theory, and machine learning. *NYUJL & Liberty*, 8:556, 2013.
- [105] Ruha Benjamin. Race after technology: Abolitionist tools for the new jim code. *Social forces*, 2019.
- [106] Benjamin Goggin. Inside Facebook’s suicide algorithm: Here’s how the company uses artificial intelligence to predict your mental state from your posts. *Business Insider*, June 2019. Journal Abbreviation: Business Insider Publisher: Insider, Inc.
- [107] Stanley I Benn. Freedom, autonomy and the concept of a person. In *Proceedings of the Aristotelian Society*, volume 76, pages 109–130. JSTOR, 1975.
- [108] Stanley I Benn. Privacy, freedom, and respect for persons. In *Privacy and personality*, pages 1–26. Routledge, 2017.
- [109] Maxwell R Bennett and Peter Michael Stephan Hacker. *Philosophical foundations of neuroscience*. John Wiley & Sons, 2022.
- [110] Sebastian Benthall, Seda Gürses, Helen Nissenbaum, et al. Contextual integrity through the lens of computer science. *Foundations and Trends® in Privacy and Security*, 2(1):1–69, 2017.
- [111] Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical Research Protocols for Social Media Health Research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain, April 2017. Association for Computational Linguistics.
- [112] Annika Bergström. Online privacy concerns: A broad approach to understanding the concerns of different groups for different uses. *Computers in Human Behavior*, 53:419–426, 2015.
- [113] Suze Berkhout and Juveria Zaheer. Digital Self-Monitoring, Bodied Realities: Re-Casting App-Based Technologies in First Episode Psychosis. *Catalyst: Feminism, Theory, Technoscience*, 7(1), April 2021.
- [114] Isaiah Berlin. Liberty. edited by henry hardy, 2002.
- [115] Alan L. Berman and Gregory Carter. Technological Advances and the Future of Suicide Prevention: Ethical, Legal, and Empirical Challenges. *Suicide and Life-Threatening Behavior*, n/a(n/a). _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/sltb.12610>.
- [116] Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. Emergent abilities in large language models: A survey. *arXiv preprint arXiv:2503.05788*, 2025.

- [117] Jaspreet Bhatia and Travis D Breaux. Empirical measurement of perceived privacy risk. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(6):1–47, 2018.
- [118] Dinesh Bhugra, Allan Tasman, Soumitra Pathare, Stefan Priebe, Shubulade Smith, John Torous, Melissa R Arbuckle, Alex Langford, Renato D Alarcón, Helen Fung Kum Chiu, Michael B First, Jerald Kay, Charlene Sunkel, Anita Thapar, Pichet Udomratn, Florence K Baingana, Dévora Kestel, Roger Man Kin Ng, Anita Patel, Livia De Picker, Kwame Julius McKenzie, Driss Moussaoui, Matt Muijen, Peter Bartlett, Sophie Davison, Tim Exworthy, Nasser Loza, Diana Rose, Julio Torales, Mark Brown, Helen Christensen, Joseph Firth, Matcheri Keshavan, Ang Li, Jukka-Pekka Onnela, Til Wykes, Hussien Elkholy, Gurvinder Kalra, Kate F Lovett, Michael J Travis, and Antonio Ventriglio. The WPA- Lancet Psychiatry Commission on the Future of Psychiatry. *The Lancet Psychiatry*, 4(10):775–818, October 2017.
- [119] Sam Biddle. Police Surveilled George Floyd Protests With Help From Twitter-Affiliated Startup Dataminr, July 2020.
- [120] Robert J Bies. Privacy and procedural justice in organizations. *Social Justice Research*, 6(1):69–86, 1993.
- [121] Abeba Birhane. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205, 2021.
- [122] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. The forgotten margins of ai ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 948–958, 2022.
- [123] Michael L. Birnbaum, Sindhu Kiranmai Ernala, Asra F. Rizvi, Munmun De Choudhury, and John M. Kane. A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *Journal of Medical Internet Research*, 19(8):e289, 2017.
- [124] Phillip A. Bishop and Robert L. Herron. Use and Misuse of the Likert Item Responses and Other Ordinal Measures. *International Journal For Exercise Science*, 8(3):297–302, 2015.
- [125] Johnna Blair, Dahlia Mukherjee, Erika FH Saunders, and Saeed Abdullah. Knowing how long a storm might last makes it easier to weather: Exploring needs and attitudes toward a data-driven and preemptive intervention system for bipolar disorder. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2023.
- [126] Paul D Bliese, Mark A Maltarich, and Jonathan L Hendricks. Back to basics with mixed-effects models: Nine take-away points. *Journal of Business and Psychology*, 33(1):1–23, 2018.
- [127] Edward J Bloustein. Privacy as an aspect of human dignity: An answer to dean prosser. *NYUL rev.*, 39:962, 1964.
- [128] Edward J Bloustein and Nathaniel J Pallone. *Individual and group privacy*. Routledge, 2018.

- [129] Kaitlin R. Boeckl and Naomi B. Lefkovitz. Nist privacy framework: A tool for improving privacy through enterprise risk management, version 1.0. NIST Cybersecurity White Paper CSWP 01162020, National Institute of Standards and Technology, January 2020. NIST Pubs.
- [130] Kirsten Boehner, Rogério DePaula, Paul Dourish, and Phoebe Sengers. Affect: from information to interaction. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, pages 59–68, 2005.
- [131] Kirsten Boehner, Rogério DePaula, Paul Dourish, and Phoebe Sengers. How emotion is made and measured. *International Journal of Human-Computer Studies*, 65(4):275–291, 2007.
- [132] Billie Bonevski, Madeleine Randell, Chris Paul, Kathy Chapman, Laura Twyman, Jamie Bryant, Irena Brozek, and Clare Hughes. Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC medical research methodology*, 14:1–29, 2014.
- [133] Jacob Bor, Atheendar S Venkataramani, David R Williams, and Alexander C Tsai. Police killings and their spillover effects on the mental health of black americans: a population-based, quasi-experimental study. *The Lancet*, 392(10144):302–310, 2018.
- [134] Brendan Bordelon. Could congress fix ai bias with privacy rules?
- [135] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2016.
- [136] Pierre Bourdieu. The forms of capital.(1986). *Cultural theory: An anthology*, 1:81–93, 2011.
- [137] Geoffrey Bowker and Susan Leigh Star. Sorting things out. *Classification and its consequences*, 4, 1999.
- [138] Anne Bowser, Katie Shilton, Jenny Preece, and Elizabeth Warrick. Accounting for privacy in citizen science: Ethical research in a context of openness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2124–2136, 2017.
- [139] Karen Boyd and Nazanin Andalibi. Automated emotion recognition in the workplace: How proposed technologies reveal potential futures of work. *Proceedings of the ACM on human-computer interaction*, 2023.
- [140] Scott Boylston. *Designing with society: A capabilities approach to design, systems thinking and social innovation*. Routledge, 2019.
- [141] Anu Bradford. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press, 2020.
- [142] Margaret M Bradley and Peter J Lang. Emotion and motivation. 2007.

- [143] Petter Bae Brandtzaeg and Asbjørn Følstad. Why people use chatbots. In *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings* 4, pages 377–392. Springer, 2017.
- [144] Johannes Britz, Anthony Hoffmann, Shana Ponelis, Michael Zimmer, and Peter Lor. On considering the application of amartya sen's capability approach to an information-based rights framework. *Information Development*, 29(2):106–113, 2013.
- [145] Joost Broekens. Modeling the experience of emotion. *International Journal of Synthetic Emotions (IJSE)*, 1(1):1–17, 2010.
- [146] Stefano Bromuri, Alexander P. Henkel, Deniz Iren, and Visara Urovi. Using AI to predict service agent stress from emotion patterns in service interactions. *Journal of Service Management*, ahead-of-print(ahead-of-print), January 2020.
- [147] Paul Brook. The alienated heart: Hochschild's 'emotional labour' thesis and the anticapitalist politics of alienation. *Capital & Class*, 33(2):7–31, 2009.
- [148] Aaron R Brough and Kelly D Martin. Critical roles of knowledge and motivation in privacy research. *Current opinion in psychology*, 31:11–15, 2020.
- [149] EPM Brouwers, MCW Joosen, C Van Zelst, and J Van Weeghel. To disclose or not to disclose: a multi-stakeholder focus group study on mental health issues in the work environment. *Journal of occupational rehabilitation*, 30:84–92, 2020.
- [150] Judith Belle Brown, Moira Stewart, and Bridget L. Ryan. Outcomes of patient-provider interaction. In *Handbook of health communication*, pages 141–161. Lawrence Erlbaum Associates Publishers, 2003.
- [151] Simone Browne. *Dark matters: On the surveillance of blackness*. Duke University Press, 2015.
- [152] Jed R. Brubaker, Lynn S. Dombrowski, Anita M. Gilbert, Nafiri Kusumakaulika, and Gillian R. Hayes. Stewarding a legacy: responsibilities and relationships in the management of post-mortem data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 4157–4166, New York, NY, USA, April 2014. Association for Computing Machinery.
- [153] Egon Brunswik. Representative design and probabilistic theory in a functional psychology. *Psychological review*, 62(3):193, 1955.
- [154] Blake Bullwinkel, Amanda Minnich, Shiven Chawla, Gary Lopez, Martin Pouliot, Whitney Maxwell, Joris de Gruyter, Katherine Pratt, Saphir Qi, Nina Chikanov, et al. Lessons from red teaming 100 generative ai products. *arXiv preprint arXiv:2501.07238*, 2025.
- [155] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

- [156] Stephen Buranyi. 'Dehumanising, impenetrable, frustrating': the grim reality of job hunting in the age of AI. *The Guardian*, March 2018.
- [157] Judee K Burgoon, Roxanne Parrott, Beth A Le Poire, Douglas L Kelley, Joseph B Walther, and Denise Perry. Maintaining and restoring privacy through communication in different types of relationships. *Journal of social and personal relationships*, 6(2):131–158, 1989.
- [158] Edmund Burke. *The Writings and Speeches of Edmund Burke: Party, Parliament, and the dividing of the Whigs 1780-1794*, volume 4. Oxford University Press, 2015.
- [159] Nicholas P Burnett, Alyssa M Hernandez, Emily E King, Richelle L Tanner, and Kathryn Wilsterman. A push for inclusive data collection in stem organizations. *Science*, 376(6588):37–39, 2022.
- [160] Christopher Burr and Nello Cristianini. Can machines read our minds? *Minds and Machines*, 29(3):461–494, 2019.
- [161] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211, 2004.
- [162] Carlos Busso and Shrikanth S Narayanan. The expression and perception of emotions: Comparing assessments of self versus others. In *Ninth annual conference of the international speech communication association*, 2008.
- [163] Jenna Butler, Mary Czerwinski, Shamsi Iqbal, Sonia Jaffe, Kate Nowak, Emily Peloquin, and Longqi Yang. Personal productivity and well-being—chapter 2 of the 2021 new future of work report. *arXiv preprint arXiv:2103.02524*, 2021.
- [164] Matt Cain and Michael Woodbridge. Hype Cycle for the Digital Workplace, 2020, July 2020.
- [165] Ryan Calo. Artificial intelligence policy: a primer and roadmap. *UCDL Rev.*, 51:399, 2017.
- [166] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37, 2010.
- [167] Erik Cambria, Andrew Livingstone, and Amir Hussain. The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer, 2012.
- [168] Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, 28(2):15–21, 2013.
- [169] Ángela Carbonell, José-Javier Navarro-Pérez, and Maria-Vicenta Mestre. Challenges and barriers in mental healthcare systems and their impact on the family: A systematic integrative review. *Health & Social Care in the Community*, 28(5):1366–1379, 2020. *eprint*: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/hsc.12968>.

- [170] Jonathan Care. Gartner: Market Guide for User and Entity Behavior Analytics, August 2018.
- [171] Wava Carissa Putri and Tjan Basaruddin. Facial emotion generation using stargan with differentiable augmentation. In *2021 4th International Conference on Signal Processing and Machine Learning*, pages 66–71, 2021.
- [172] J. M. Carroll. Five reasons for scenario-based design. *Interacting with Computers*, 13(1):43–60, September 2000. Publisher: Oxford Academic.
- [173] Chelsey R. Carter. Gaslighting: ALS, anti-Blackness, and medicine. *Feminist Anthropology*, n/a(n/a), 2022. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/fea2.12107>.
- [174] Electronic Privacy Information Center. EPIC - In re HireVue.
- [175] Electronic Privacy Information Center. Hirevue, facing ftc complaint from epic, halts use of facial recognition. 2021.
- [176] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, 2023.
- [177] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):147:1–147:32, November 2019.
- [178] Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 79–88, Atlanta, GA, USA, 2019. ACM Press.
- [179] Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1):43, March 2020.
- [180] Stevie Chancellor, Yannis Kalantidis, Jessica A. Pater, Munmun De Choudhury, and David A. Shamma. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3213–3226, New York, NY, USA, 2017. Association for Computing Machinery.
- [181] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1201–1213, 2016.
- [182] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. sage, 2006.

- [183] Kathy Charmaz and Richard G Mitchell. Grounded theory in ethnography. *Handbook of ethnography*, 160:174, 2001.
- [184] Adam Mourad Chekroud, Ryan Joseph Zotti, Zarrar Shehzad, Ralitza Gueorguieva, Marcia K Johnson, Madhukar H Trivedi, Tyrone D Cannon, John Harrison Krystal, and Philip Robert Corlett. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry*, 3(3):243–250, March 2016.
- [185] Angela Chen and K Hao. Emotion ai researchers say overblown claims give their work a bad name. *MIT Technology Review*, 14:2020, 2020.
- [186] Szu-Wei Cheng, Chung-Wen Chang, Wan-Jung Chang, Hao-Wei Wang, Chih-Sung Liang, Taishiro Kishimoto, Jane Pei-Chen Chang, John S Kuo, and Kuan-Pin Su. The now and future of chatgpt and gpt in psychiatry. *Psychiatry and clinical neurosciences*, 77(11):592–596, 2023.
- [187] Carl J Chimi and David L Russell. The likert scale: A proposal for improvement using quasi-continuous variables. In *Information Systems Education Conference, Washington, DC*, pages 1–10, 2009.
- [188] Eun Kyoung Choe, Saeed Abdullah, Mashfiqui Rabbi, Edison Thomaz, Daniel A Epstein, Felicia Cordeiro, Matthew Kay, Gregory D Abowd, Tanzeem Choudhury, James Fogarty, et al. Semi-automated tracking: a balanced approach for self-monitoring applications. *IEEE Pervasive Computing*, 16(1):74–84, 2017.
- [189] Tom A.C. Chrisp, Sharon Tabberer, Benjamin D. Thomas, and Wayne A. Goddard. Dementia early diagnosis: Triggers, supports and constraints affecting the decision to engage with the health care system. *Aging & Mental Health*, 16(5):559–565, July 2012.
- [190] Helen Christensen, Philip J. Batterham, and Bridianne O’Dea. E-Health Interventions for Suicide Prevention. *International Journal of Environmental Research and Public Health*, 11(8):8193–8212, August 2014. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.
- [191] Seung Youn Yonnie Chyung, Ieva Swanson, Katherine Roberts, and Andrea Hankinson. Evidence-Based Survey Design: The Use of Continuous Rating Scales in Surveys. *Performance Improvement*, 57(5):38–48, May 2018.
- [192] Danielle Keats Citron. Mainstreaming privacy torts. *Calif. L. Rev.*, 98:1805, 2010.
- [193] Danielle Keats Citron. *The fight for privacy: protecting dignity, identity, and love in the digital age*. W.W. Norton & Company, Inc, New York, first edition edition, 2022.
- [194] Danielle Keats Citron and Daniel J Solove. Privacy harms. *BUL Rev.*, 102:793, 2022.
- [195] Adele E Clarke. Situational analyses: Grounded theory mapping after the postmodern turn. *Symbolic interaction*, 26(4):553–576, 2003.
- [196] Adele E Clarke and Kathy Charmaz. *Grounded theory and situational analysis*. Sage, 2014.

- [197] James Clayton and Ben Derico. Clearview ai used nearly 1m times by us police, it tells the bbc.
- [198] Damian Clifford, Megan Richardson, and Normann Witzleb. Artificial intelligence and sensitive inferences: New challenges for data protection laws. *Regulatory Insights on Artificial Intelligence: Research for Policy* (Edward Elgar, 2021), 2020.
- [199] Jean L Cohen. Democracy, difference, and the right of privacy. *Democracy and difference: Contesting the boundaries of the political*, pages 187–217, 1996.
- [200] Julie E Cohen. *Configuring the networked self: Law, code, and the play of everyday practice*. Yale University Press, 2012.
- [201] Julie E Cohen. What privacy is for. *Harv. L. Rev.*, 126:1904, 2012.
- [202] Julie E Cohen. The surveillance-innovation complex: The irony of the participatory turn. *The Participatory Condition* (University of Minnesota Press, 2015, Forthcoming), 2014.
- [203] Jeffrey F Cohn and Karen L Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2(02):121–132, 2004.
- [204] Roy Coleman. Reclaiming the streets: Closed circuit television, neoliberalism and the mystification of social divisions in liverpool, uk. *Surveillance & Society*, 2(2/3), 2004.
- [205] Patricia Hill Collins. *Intersectionality as critical social theory*. Duke University Press, 2019.
- [206] Randall Collins. Stratification, emotional energy, and the transient emotions. *Research agendas in the sociology of emotions*, 27:57, 1990.
- [207] Giovanna Colombetti. Enaction, sense-making and emotion. *Enaction: Toward a new paradigm for cognitive science*, pages 145–164, 2010.
- [208] Giovanna Colombetti and Evan Thompson. Enacting emotional interpretations with feeling. *Behavioral and Brain Sciences*, 28(2):200–201, 2005.
- [209] European Commission. Monitoring of Digital Rights and Principles – Support study 2024 | Shaping Europe’s digital future, July 2024.
- [210] European Commission. Annex to the communication to the commission approval of the content of the draft communication from the commission - commission guidelines on prohibited artificial intelligence practices established by regulation (eu) 2024/1689 (ai act), C(2015).
- [211] Mike Conway and Daniel O’Connor. Social media, big data, and mental health: current advances and ethical implications. *Current Opinion in Psychology*, 9:77–82, June 2016.
- [212] Benjamin L. Cook, Ana M. Progovac, Pei Chen, Brian Mullin, Sherry Hou, and Enrique Baca-Garcia. Novel Use of Natural Language Processing (NLP) to Predict Suicidal Ideation and Psychiatric Symptoms in a Text-Based Mental Health Intervention in Madrid. *Computational and Mathematical Methods in Medicine*, 2016:1–8, 2016.

- [213] Kevin W Cook, Carol A Vance, and Paul E Spector. The relation of candidate personality with selection-interview outcomes. *Journal of Applied Social Psychology*, 30(4):867–885, 2000.
- [214] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics.
- [215] Glen A Coppersmith, Craig T Harman, and Mark H Dredze. Measuring Post Traumatic Stress Disorder in Twitter. page 4.
- [216] Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States, 2008.
- [217] Juliet Corbin and Anselm Strauss. Strategies for qualitative data analysis. *Basics of Qualitative Research. Techniques and procedures for developing grounded theory*, 3, 2008.
- [218] Daniel T Cordaro, Rui Sun, Dacher Keltner, Shanmukh Kamble, Niranjan Huddar, and Galen McNeil. Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*, 18(1):75, 2018.
- [219] Shanley Corvite*, Kat Roemmich*, Tillie Rosenberg, and Nazanin Andalibi. Data subjects' perspectives on emotion artificial intelligence use in the workplace: A relational ethics lens. *Proceedings of the ACM on human-computer interaction*, 2023.
- [220] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020.
- [221] Kaitlin L Costello and Diana Floegel. “predictive ads are not doctors”: Mental health tracking and technology companies. *Proceedings of the Association for Information Science and Technology*, 57(1):e250, 2020.
- [222] Nick Couldry and Ulises Ali Mejias. *The costs of connection: how data is colonizing human life and appropriating it for capitalism*. Culture and economic life. Stanford University Press, Stanford, California, 2019.
- [223] Adam K Coyne, Andrew Murtagh, and Conor McGinn. Using the geneva emotion wheel to measure perceived affect in human-robot interaction. In *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 491–498. IEEE, 2020.
- [224] Kate Crawford. Artificial Intelligence Is Misreading Human Emotion, April 2021. Section: Technology.
- [225] Kate Crawford. Time to regulate ai that interprets human emotions. *Nature*, 592(7853):167–167, 2021.

- [226] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. AI Now 2019 Report. Technical report, New York: AI Now Institute, December 2019.
- [227] Kate Crawford and Jason Schultz. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, 55:93, 2014.
- [228] William A Creech. Psychological testing and constitutional rights. *Duke LJ*, page 332, 1966.
- [229] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, pages 23–51. Routledge, 2013.
- [230] Kimberlé W Crenshaw. *On intersectionality: Essential writings*. The New Press, 2017.
- [231] Carlos Crivelli and Alan J Fridlund. Facial displays are tools for social influence. *Trends in Cognitive Sciences*, 22(5):388–399, 2018.
- [232] Russell Cropanzano, Howard M Weiss, and Steven M Elias. The impact of display rules and emotional labor on psychological well-being at work. In *Emotional and physiological processes and positive intervention strategies*. Emerald Group Publishing Limited, 2003.
- [233] Mira Crouch and Heather McKenzie. The logic of small samples in interview-based qualitative research. *Social science information*, 45(4):483–499, 2006.
- [234] Válber César Cavalcanti Roza and Octavian Adrian Postolache. Multimodal Approach for Emotion Recognition Based on Simulated Flight Experiments. *Sensors (Basel, Switzerland)*, 19(24), December 2019.
- [235] David Yun Dai and Robert J Sternberg. *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development*. Routledge, 2004.
- [236] Antonio R. Damasio. *Descartes' error: emotion, reason and the human brain*. Vintage, London, rev. ed. with a new preface edition, 2006. OCLC: 255585811.
- [237] Kate Darling. 'who's johnny?' anthropomorphic framing in human-robot interaction, integration, and policy. *Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy (March 23, 2015)*. ROBOT ETHICS, 2, 2015.
- [238] Charles Darwin. *The expression of the emotions in man and animals*. University of Chicago press, 2015.
- [239] Charles Darwin, 1809-1882. *The expression of the emotions in man and animals*. Number vi, 374, 4 p., 7 leaves of plates (3 fold.) (4 p. at end advertisements) :. J. Murray, London :, 1872. Publication Title: The expression of the emotions in man and animals.
- [240] Partha Das Chowdhury and Karen Renaud. ‘ought’should not assume ‘can’? basic capabilities in cybersecurity to ground sen’s capability approach. In *Proceedings of the 2023 New Security Paradigms Workshop*, pages 76–91, 2023.

- [241] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women, October 2018.
- [242] Ben Dattner, Tomas Chamorro-Premuzic, Richard Buchband, and Lucinda Schettler. The legal and ethical implications of using ai in hiring. *Harvard Business Review*, 25, 2019.
- [243] Thomas Davenport, Abhijit Guha, Dhruv Grewal, and Timna Bressgott. How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48(1):24–42, 2020.
- [244] Deborah Sarah. David and Robert. Brannon. *The Forty-nine percent majority : the male sex role*. Number xiv, 338 p. ;. Addison-Wesley Pub. Co., Reading, Mass. :, 1976. Publication Title: The Forty-nine percent majority : the male sex role.
- [245] Darryl N Davis and Suzanne C Lewis. Computational models of emotion for autonomy and reasoning. *Informatica (Slovenia)*, 27(2):157–164, 2003.
- [246] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Honolulu HI USA, April 2020. ACM.
- [247] Arnold B De Castro, Jacqueline Agnew, and Sheila T Fitzgerald. Emotional labor: relevant theory for occupational health practice in post-industrial america. *AAOHN journal*, 52(3):109–115, 2004.
- [248] Munmun De Choudhury and Scott Counts. Understanding affect in the workplace via social media. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW ’13, pages 303–316, New York, NY, USA, February 2013. Association for Computing Machinery.
- [249] Munmun De Choudhury, Scott Counts, and Michael Gamon. Not All Moods are Created Equal! Exploring Human Emotional States in Social Media. page 8.
- [250] Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, CSCW ’14, pages 626–638, New York, NY, USA, February 2014. Association for Computing Machinery.
- [251] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting Depression via Social Media. AAAI, July 2013.
- [252] Munmun De Choudhury, Choudhury Michael, and Gamon Scott Counts. *Happy, Nervous or Surprised? Classification of Human Affective States in Social Media*.
- [253] Maartje MA de Graaf. An ethical evaluation of human–robot relationships. *International journal of social robotics*, 8:589–598, 2016.

- [254] Judith Wagner DeCew. *In pursuit of privacy: Law, ethics, and the rise of technology*. Cornell University Press, 1997.
- [255] Oscar Delaney, Oliver Guest, and Zoe Williams. Mapping technical safety research at ai companies: A literature review and incentives analysis. *arXiv preprint arXiv:2409.07878*, 2024.
- [256] Lina Dencik and Sanne Stevens. Regimes of justification in the datafied workplace: The case of hiring. *new media & society*, page 14614448211052893, 2021.
- [257] René Descartes. *The philosophical writings of Descartes*, volume 2. Cambridge University Press, 1984.
- [258] Mandar Deshpande and Vignesh Rao. Depression detection using emotion artificial intelligence. In *2017 international conference on intelligent sustainable systems (iciss)*, pages 858–862. IEEE, 2017.
- [259] Lothar Determann and Jonathan Tam. The california privacy rights act of 2020: A broad and complex data processing regulation that applies to businesses worldwide. *Journal of Data Protection & Privacy*, 4(1):7–21, 2020.
- [260] John Dewey. The theory of emotion. *Psychological review*, 2(1):13, 1895.
- [261] Marwan Dhuheir, Abdullatif Albaseer, Emna Baccour, Aiman Erbad, Mohamed Abdallah, and Mounir Hamdi. Emotion recognition for healthcare surveillance systems using neural networks: A survey. In *2021 International Wireless Communications and Mobile Computing (IWCMC)*, pages 681–687. IEEE, 2021.
- [262] Lisa M Diamond and Jenna Alley. Rethinking minority stress: A social safety perspective on the health effects of stigma in sexually-diverse and gender-diverse populations. *Neuroscience & Biobehavioral Reviews*, 138:104720, 2022.
- [263] Joao Dias and Ana Paiva. Feeling and reasoning: A computational model for emotional characters. In *Portuguese conference on artificial intelligence*, pages 127–140. Springer, 2005.
- [264] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. Recognition of genuine smiles. *IEEE Transactions on Multimedia*, 17(3):279–294, 2015.
- [265] Catherine D'ignazio and Lauren F Klein. *Data feminism*. MIT press, 2020.
- [266] Jose van Dijck. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2):197–208, May 2014.
- [267] Charles Henri DiMaria, Chiara Peroni, and Francesco Sarracino. Happiness matters: productivity gains from subjective well-being. *Journal of Happiness Studies*, 21(1):139–160, 2020.

- [268] Institute of Medicine (US) Committee on Crossing the Quality Chasm: Adaptation to Mental Health and Addictive Disorders. *Constraints on Sharing Mental Health and Substance-Use Treatment Information Imposed by Federal and State Medical Records Privacy Laws*. National Academies Press (US), 2006. Publication Title: Improving the Quality of Health Care for Mental and Substance-Use Conditions: Quality Chasm Series.
- [269] Thomas Dixon. “emotion”: The history of a keyword in crisis. *Emotion Review*, 4(4):338–344, 2012.
- [270] Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. The limits of differential privacy (and its misuse in data release and machine learning). *Communications of the ACM*, 64(7):33–35, 2021.
- [271] Artem Domnich and Gholamreza Anbarjafari. Responsible ai: Gender bias assessment in emotion recognition. *arXiv preprint arXiv:2103.11436*, 2021.
- [272] J. Doyle, N. Caprani, and R. Bond. Older adults’ attitudes to self-management of health and wellness through smart home data. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 129–136, 2015.
- [273] Christine W. Duclos, Mary Eichler, Leslie Taylor, Javan Quintela, Deborah S. Main, Wilson Pace, and Elizabeth W. Staton. Patient perspectives of patient–provider communication after adverse events. *International Journal for Quality in Health Care*, 17(6):479–486, December 2005.
- [274] Juan I Durán, Rainer Reisenzein, and José-Miguel Fernández-Dols. Coherence between emotions and facial expressions. *The science of facial expression*, pages 107–129, 2017.
- [275] Jane Dwivedi-Yu, Yi-Chia Wang, Lijing Qin, Cristian Canton-Ferrer, and Alon Y Halevy. Affective signals in a social media recommender system. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2831–2841, 2022.
- [276] Ronald Dworkin. *Taking rights seriously*. A&C Black, 1977.
- [277] Iris Eekhout, Michiel R de Boer, Jos WR Twisk, Henrica CW de Vet, and Martijn W Heymans. Brief report: missing data: a systematic review of how they are reported and handled. *Epidemiology*, pages 729–732, 2012.
- [278] Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.
- [279] Maria Egger, Matthias Ley, and Sten Hanke. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science*, 343:35–55, 2019.
- [280] Laura Eggertson. Social media embraces suicide prevention. *CMAJ*, 187(11):E333–E333, August 2015. Publisher: CMAJ Section: News.
- [281] Ray Eitel-Porter. Beyond the promise: implementing ethical ai. *AI and Ethics*, 1(1):73–80, 2021.

- [282] Pantelimon Ekkekakis. Affect, mood, and emotion. 2012.
- [283] Paul Ekman. The argument and evidence about universals in facial expressions. *Handbook of social psychophysiology*, 143:164, 1989.
- [284] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.
- [285] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.
- [286] Paul Ekman and Daniel Cordaro. What is meant by calling emotions basic. *Emotion review*, 3(4):364–370, 2011.
- [287] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.
- [288] Paul Ekman, Wallace V Friesen, Maureen O’sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.
- [289] Paul Ekman and Dacher Keltner. Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, 27:46, 1997.
- [290] Rana El Kaliouby and Carol Colman. *Girl Decoded: A Scientist’s Quest to Reclaim Our Humanity by Bringing Emotional Intelligence to Technology*. Currency, 2021.
- [291] Hillary Anger Elfenbein. Emotional dialects in the language of emotion. *The science of facial expression*, pages 479–496, 2017.
- [292] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.
- [293] Marta Elliott and Jordan C Reuter. Disclosure, discrimination, and identity among working professionals with bipolar disorder or major depression. 2021.
- [294] Bruce J Ellis and David F Bjorklund. Beyond mental health: an evolutionary analysis of development under risky and supportive environmental conditions: an introduction to the special section. *Developmental psychology*, 48(3):591, 2012.
- [295] Darren Ellis and Ian Tucker. *Emotion in the digital age: Technologies, data and psychosocial life*. Routledge, 2020.
- [296] Hadar Elraz. Identity, mental health and work: How employees with mental health conditions recount stigma and the pejorative discourse of mental illness. *Human Relations*, 71(5):722–741, 2018.
- [297] Motahare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. “I always assumed that I wasn’t really that close to [her]”: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*,

CHI '15, pages 153–162, New York, NY, USA, April 2015. Association for Computing Machinery.

- [298] María Lucía Barrón Estrada, Ramón Zatarain Cabada, Raúl Oramas Bustillos, and Mario Graff. Opinion mining and emotion recognition applied to learning environments. *Expert Systems with Applications*, 150:113265, 2020.
- [299] Amitai Etzioni. A cyber age privacy doctrine: More coherent, less subjective, and operational. *Brook. L. Rev.*, 80:1263, 2014.
- [300] European Commission. Commission guidelines on prohibited artificial intelligence practices established by regulation (eu) 2024/1689 (ai act). Communication to the Commission C(2025) 884 final, European Commission, Brussels, February 2025. Draft guidelines annexed to the Communication; formally adopted upon.
- [301] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). <https://eur-lex.europa.eu/eli/reg/2016/679/oj>, April 2016. Official Journal of the European Union, L 119, pp. 1–88.
- [302] European Parliament and Council of the European Union. Regulation (eu) 2024/xxxx of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>, 2024. Final text approved by European Parliament on March 13, 2024.
- [303] European Union. Charter of fundamental rights of the european union. Official Journal of the European Communities, December 18 2000. Notice No. 2000/C 364/01, Page 1.
- [304] European Union Agency for Fundamental Rights. Article 52 - Scope and interpretation of rights and principles. <https://fra.europa.eu/en/eu-charter/article/52-scope-and-interpretation-rights-and-principles>, 2021. Accessed: 2025-06-17.
- [305] Gunther Eysenbach. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *Journal of Medical Internet Research*, 11(1):e11, 2009. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.
- [306] Gunther Eysenbach. Crisis text line and loris. ai controversy highlights the complexity of informed consent on the internet and data-sharing ethics for machine learning and research, 2025.

- [307] Evangelia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, and Giannis Tzimas. Application of machine learning algorithms to an online recruitment system. In *Proc. International Conference on Internet and Web Applications and Services*, pages 215–220. Citeseer, 2012.
- [308] George Farkas. Cognitive skills and noncognitive traits and behaviors in stratification processes. *Annual review of sociology*, 29(1):541–562, 2003.
- [309] Michael Feffer, Anusha Sinha, Wesley H Deng, Zachary C Lipton, and Hoda Heidari. Red-teaming for generative ai: Silver bullet or security theater? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 421–437, 2024.
- [310] Sinead V Fernandes and Muhammad S Ullah. Development of spectral speech features for deception detection using neural networks. In *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0198–0203. IEEE, 2021.
- [311] Jessica L. Feuston and Anne Marie Piper. Beyond the coded gaze: Analyzing expression of mental health and illness on instagram. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018.
- [312] Jessica L. Feuston and Anne Marie Piper. Beyond the Coded Gaze: Analyzing Expression of Mental Health and Illness on Instagram. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21, November 2018.
- [313] Jessica L. Feuston and Anne Marie Piper. Everyday experiences: Small stories and mental illness on instagram. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery.
- [314] Jessica L. Feuston and Anne Marie Piper. Everyday experiences: Small stories and mental illness on Instagram. In *CHI 2019 - Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, May 2019.
- [315] Jessica L. Feuston, Alex S. Taylor, and Anne Marie Piper. Conformity of Eating Disorders through Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–28, May 2020.
- [316] Casey Fiesler, Cliff Lampe, and Amy S. Bruckman. Reality and perception of copyright terms of service for online content creation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW ’16, page 1450–1461, New York, NY, USA, 2016. Association for Computing Machinery.
- [317] Casey Fiesler and Nicholas Proferes. “Participant” Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1).
- [318] Janet Finch. THE VIGNETTE TECHNIQUE IN SURVEY RESEARCH. *Sociology*, 21(1):105–114, 1987. Publisher: Sage Publications, Ltd.

- [319] W Holmes Finch, Jocelyn E Bolin, and Ken Kelley. *Multilevel modeling using R*. Crc Press, 2019.
- [320] Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. Inverse constitutional ai: Compressing preferences into principles. *arXiv preprint arXiv:2406.06560*, 2024.
- [321] FINDER. Emotion recognition technology in the financial sector – Curse or blessing?, January 2020.
- [322] Matthew W Finkin. Employee privacy, american values, and the law. *Chi.-Kent L. Rev.*, 72:221, 1996.
- [323] Melissa L Finucane, Ali Alhakami, Paul Slovic, and Stephen M Johnson. The affect heuristic in judgments of risks and benefits. *Journal of behavioral decision making*, 13(1):1–17, 2000.
- [324] Alexandra Fiore and Matthew Weinick. Undignified in defeat: An analysis of the stagnation and demise of proposed legislation limiting video surveillance in the workplace and suggestions for change. *Hofstra Lab. & Emp. LJ*, 25:525, 2007.
- [325] Benjamin Fish and Luke Stark. Reflexive design for fairness and other human values in formal models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 89–99, 2021.
- [326] Helena Flam and Jochen Kleres. *Methods of exploring emotions*. Routledge, 2015.
- [327] Mary Flanagan and Helen Nissenbaum. *Values at play in digital games*. MIT Press, 2014.
- [328] Annette Flanagin, Tracy Frey, Stacy L Christiansen, and Howard Bauchner. The reporting of race and ethnicity in medical and science journals: comments invited. *Jama*, 325(11):1049–1052, 2021.
- [329] Luciano Floridi. On human dignity as a foundation for the right to privacy. *Philosophy & Technology*, 29(4):307–312, 2016.
- [330] Joshua Fogel and Elham Nehmad. Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in human behavior*, 25(1):153–160, 2009.
- [331] Trehani M Fonseka, Venkat Bhat, and Sidney H Kennedy. The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors. *Australian & New Zealand Journal of Psychiatry*, 53(10):954–964, October 2019.
- [332] United States. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont report: ethical principles and guidelines for the protection of human subjects of research*, volume 2. Department of Health, Education, and Welfare, National Commission for the . . . , 1978.

- [333] Elizabeth Ford, Keegan Curlewis, Akkapon Wongkoblap, and Vasa Curcin. Public Opinions on Using Social Media Content to Identify Users With Depression and Target Mental Health Care Advertising: Mixed Methods Survey. *JMIR Mental Health*, 6(11):e12942, 2019. Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.
- [334] Nikolaus Forgó, Stefanie Hänold, and Benjamin Schütze. The principle of purpose limitation and big data. *New technology, big data and the law*, pages 17–42, 2017.
- [335] Michel Foucault. *Discipline and punish: The birth of the prison*. Vintage, 2012.
- [336] Marion Fourcade and Kieran Healy. Seeing like a market. *Socio-Economic Review*, 15(1):9–29, 2017.
- [337] Mark G Frank, Paul Ekman, and Wallace V Friesen. Behavioral markers and recognizability of the smile of enjoyment. *Journal of personality and social psychology*, 64(1):83, 1993.
- [338] Max Freyd. The graphic rating scale. *Journal of educational psychology*, 14(2):83, 1923.
- [339] Charles Fried. Privacy, 1968.
- [340] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.
- [341] Nico H Frijda. *The laws of emotion*. Psychology Press, 2017.
- [342] Brett M. Frischmann and Evan Selinger. *Re-engineering humanity*. Cambridge University Press, Cambridge, United Kingdom ; New York, NY, 2018.
- [343] Yoko E. Fukumura, Julie McLaughlin Gray, Gale M. Lucas, Burcin Becerik-Gerber, and Shawn C. Roll. Worker Perspectives on Incorporating Artificial Intelligence into Office Workspaces: Implications for the Future of Office Work. *International Journal of Environmental Research and Public Health*, 18(4):1690, January 2021. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.
- [344] Iason Gabriel and Vafa Ghazavi. The challenge of value alignment. In *The Oxford handbook of digital ethics*. Oxford University Press Oxford, 2022.
- [345] Meredith Damien Gall, Walter R Borg, and Joyce P Gall. *Educational research: An introduction*. Longman Publishing, 1996.
- [346] Cornelius E Gallagher. Why house hearings on invasion of privacy. *American Psychologist*, 20(11):881, 1965.
- [347] Patricia Garcia, Tonia Sutherland, Marika Cifor, Anita Say Chan, Lauren Klein, Catherine D'Ignazio, and Niloufar Salehi. No: Critical refusal as feminist data practice. In *conference companion publication of the 2020 on computer supported cooperative work and social computing*, pages 199–202, 2020.

- [348] Deepak Garg, Limin Jia, and Anupam Datta. Policy auditing over incomplete logs: theory, implementation and applications. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 151–162, 2011.
- [349] Muskan Garg and Chandni Saxena. Emotion detection from text data using machine learning for human behavior analysis. In *Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications*, pages 129–144. Elsevier, 2024.
- [350] Vaibhav Garg, Lesa Lorenzen-Huber, L Jean Camp, and Kay Connelly. Risk communication design for older adults. In *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, volume 29, page 1. IAARC Publications, 2012.
- [351] Danielle Gaucher, Justin Friesen, and Aaron C Kay. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109, 2011.
- [352] Ruth Gavison. Privacy and the limits of law. *The Yale law journal*, 89(3):421–471, 1980.
- [353] Ruth E Gavison. Feminism and the private-public distinction. *Stanford Law Review*, 45:1, 1992.
- [354] Roy Gelbard, Roni Ramon-Gonen, Abraham Carmeli, Ran M Bittmann, and Roman Talyansky. Sentiment analysis in organizational work: Towards an ontology of people analytics. *Expert Systems*, 35(5):e12289, 2018.
- [355] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.
- [356] Luana Giatti, Lidyane do Valle Camelo, Jôsi Fernandes de Castro Rodrigues, and Sandhi Maria Barreto. Reliability of the macarthur scale of subjective social status-brazilian longitudinal study of adult health (elsa-brasil). *BMC public health*, 12(1):1–7, 2012.
- [357] Tarleton Gillespie. The Relevance of Algorithms. In Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, editors, *Media Technologies*, pages 167–194. The MIT Press, February 2014.
- [358] Susan E Gindin. Nobody reads your privacy policy or online contract: Lessons learned and questions raised by the ftc’s action against sears. *Nw. J. Tech. & Intell. Prop.*, 8:1, 2009.
- [359] Lisa Gitelman. *Raw data is an oxymoron*. MIT press, 2013.
- [360] Tasha Glenn and Scott Monteith. Privacy in the Digital World: Medical and Health Data Outside of HIPAA Protections. *Current Psychiatry Reports*, 16(11):494, September 2014.
- [361] Goasduff. Gartner Says By 2023, 65% of the World’s Population Will Have Its Personal Data Covered Under Modern Privacy Regulations, September 2020.

- [362] Ron Z Goetzel, Ronald J Ozminkowski, Lloyd I Sederer, and Tami L Mark. The business case for quality mental health services: why employers should care about the mental health and well-being of their employees. *Journal of occupational and environmental medicine*, pages 320–330, 2002.
- [363] Erving Goffman et al. *The presentation of self in everyday life*, volume 21. Harmondsworth London, 1978.
- [364] Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. Ethics and Artificial Intelligence: Suicide Prevention on Facebook. *Philosophy & technology*, 31(4):669–684, 2018. Place: Dordrecht Publisher: Springer Science and Business Media LLC.
- [365] Manuel F Gonzalez, John F Capman, Frederick L Oswald, Evan R Theys, and David L Tomczak. “where’s the io?” artificial intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions*, 5(3):5, 2019.
- [366] Gregorio González-Alcaide, Mercedes Fernández-Ríos, Rosa Redolat, and Emilia Serra. Research on Emotion Recognition and Dementias: Foundations and Prospects. *Journal of Alzheimer’s Disease*, 82(3):939–950, August 2021.
- [367] Steven L Gordon. Social structural effects on emotions. *Research agendas in the sociology of emotions*, pages 145–179, 1990.
- [368] Elizabeth H Gorman. Gender stereotypes, same-gender preferences, and organizational variation in the hiring of women: Evidence from law firms. *American Sociological Review*, 70(4):702–728, 2005.
- [369] Cristina Gorrostieta, Reza Lotfian, Kye Taylor, Richard Brutti, and John Kane. Gender de-biasing in speech emotion recognition. In *INTERSPEECH*, pages 2823–2827, 2019.
- [370] Anna Grabowska and Artur Gunia. On quantum computing for artificial superintelligence. *European Journal for Philosophy of Science*, 14(2):25, 2024.
- [371] Alicia A Grandey. Emotional regulation in the workplace: A new way to conceptualize emotional labor. *Journal of occupational health psychology*, 5(1):95, 2000.
- [372] Didier Grandjean, David Sander, and Klaus R Scherer. Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and cognition*, 17(2):484–495, 2008.
- [373] Jonathan Gratch and Stacy Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.
- [374] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.
- [375] Melissa Gregg. *Work’s intimacy*. John Wiley & Sons, 2013.

- [376] Gabriel Grill and Nazanin Andalibi. Attitudes and folk theories of data subjects on transparency and accuracy in emotion recognition. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–35, 2022.
- [377] David A Grimes, Janie Benson, Susheela Singh, Mariana Romero, Bela Ganatra, Friday E Okonofua, and Iqbal H Shah. Unsafe abortion: the preventable pandemic. *The lancet*, 368(9550):1908–1919, 2006.
- [378] Martin L Gross. The brain watchers. 1962.
- [379] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019.
- [380] John S Grove, Dwayne M Reed, Katsuhiko Yano, and Lie-Ju Hwang. Variability in systolic blood pressure—a risk factor for coronary heart disease? *American journal of epidemiology*, 145(9):771–776, 1997.
- [381] Barry Gruenberg. The happy worker: An analysis of educational and occupational differences in determinants of job satisfaction. *American journal of sociology*, 86(2):247–271, 1980.
- [382] Tammy Guberek, Allison McDonald, Sylvia Simioni, Abraham H Mhaidli, Kentaro Toyama, and Florian Schaub. Keeping a low profile? technology, risk and privacy among undocumented immigrants. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–15, 2018.
- [383] Hatice Gunes and Björn Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013.
- [384] Rashmi Gupta. Positive emotions have a unique capacity to capture attention. *Progress in brain research*, 247:23–46, 2019.
- [385] Jürgen Habermas. The public sphere: An encyclopedia article. *New German Critique*, pages 102–107, 2001.
- [386] Pieter Haeck. The eu’s ai bans come with big loopholes for police, February 2025.
- [387] Mario Haim, Florian Arendt, and Sebastian Scherr. Abyss or Shelter? On the Relevance of Web Search Engines’ Search Results When People Google for Suicide. *Health Communication*, 32(2):253–258, February 2017. Publisher: Routledge eprint: <https://doi.org/10.1080/10410236.2015.1113484>.
- [388] Oliver L Haimson. Mapping gender transition sentiment patterns via social media data: toward decreasing transgender mental health disparities. *Journal of the American Medical Informatics Association*, 26(8-9):749–758, 2019.

- [389] Oliver L. Haimson, Jed R. Brubaker, Lynn Dombrowski, and Gillian R. Hayes. Disclosure, stress, and support during gender transition on facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 1176–1190, New York, NY, USA, 2015. Association for Computing Machinery.
- [390] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35, October 2021.
- [391] Blake Hallinan, Jed R Brubaker, and Casey Fiesler. Unexpected expectations: Public reaction to the Facebook emotional contagion study. *New Media & Society*, 22(6):1076–1094, June 2020. Publisher: SAGE Publications.
- [392] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, New York, NY, USA, April 2018. Association for Computing Machinery.
- [393] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13, 2018.
- [394] Alex Hanna and Meredith Whittaker. Opinion: Timnit Gebru's Exit From Google Exposes a Crisis in AI. *Wired*, 2021.
- [395] Carl L. Hanson, Scott H. Burton, Christophe Giraud-Carrier, Josh H. West, Michael D. Barnes, and Bret Hansen. Tweaking and Tweeting: Exploring Twitter for Nonmedical Use of a Psychostimulant Drug (Adderall) Among College Students. *Journal of Medical Internet Research*, 15(4):e62, 2013. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.
- [396] Simone Hantke, Zixing Zhang, and Björn Schuller. Towards intelligent crowdsourcing for audio data annotation: Integrating active learning in the real world. 2017.
- [397] Eddie Harmon-Jones, Cindy Harmon-Jones, and Elizabeth Summerell. On the importance of both dimensional and discrete models of emotion. *Behavioral sciences*, 7(4):66, 2017.
- [398] Paul Harpur, Fitore Hyseni, and Peter Blanck. Workplace health surveillance and covid-19: algorithmic health discrimination and cancer survivors. *Journal of Cancer Survivorship*, 16(1):200–212, 2022.
- [399] Woodrow Hartzog. The inadequate, invaluable fair information practices. *Md. L. Rev.*, 76:952, 2016.

- [400] Aya Hassouneh, A.M. Mutawa, and Murugappan M. Development of a Real-Time Emotion Recognition System Using Facial Expressions and EEG based on machine learning and deep neural network methods. *Informatics in Medicine Unlocked*, 20:100372, 2020.
- [401] Andreas Häuselmann, Alan M Sears, Lex Zard, and Eduard Fosch-Vilaronga. Eu law and emotion data. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2023.
- [402] Daniel M Haybron. Happiness, the self and human flourishing. *Utilitas*, 20(1):21–49, 2008.
- [403] Natali Helberger, Marijn Sax, Joanna Strycharz, and H-W Micklitz. Choice architectures in the digital economy: Towards a new understanding of digital vulnerability. *Journal of Consumer Policy*, pages 1–26, 2022.
- [404] Gillian Z Heller, Maurizio Manuguerra, and Roberta Chow. How to analyze the visual analogue scale: Myths, truths and clinical relevance. *Scandinavian journal of pain*, 13(1):67–75, 2016.
- [405] Paula Helm, Benjamin Lipp, and Roser Pujadas. Generating reality and silencing debate: Synthetic data as discursive device. *Big Data & Society*, 11(2):20539517241249447, 2024.
- [406] Stephen E Henderson. Expectations of privacy in social media. *Miss. CL Rev.*, 31:227, 2012.
- [407] Dan Hendrycks, Eric Schmidt, and Alexandr Wang. Superintelligence strategy: Expert version. *arXiv preprint arXiv:2503.05628*, 2025.
- [408] Alexander P Henkel, Stefano Bromuri, Deniz Iren, and Visara Urovi. Half human, half machine—augmenting service employees with ai for interpersonal emotion regulation. *Journal of Service Management*, 2020.
- [409] Javier Hernandez, Josh Lovejoy, Daniel McDuff, Jina Suh, Tim O’Brien, Arathi Sethumadhavan, Gretchen Greene, Rosalind Picard, and Mary Czerwinski. Guidelines for assessing and minimizing risks of emotion recognition applications. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.
- [410] Ellen Heuven, Arnold B Bakker, Wilmar B Schaufeli, and Noortje Huisman. The role of self-efficacy in performing emotion work. *Journal of vocational behavior*, 69(2):222–235, 2006.
- [411] Andrew C. High, Anne Oeldorf-Hirsch, and Saraswathi Bellur. Misery rarely gets company: The influence of emotional bandwidth on supportive communication on Facebook. *Computers in Human Behavior*, 34:79–88, May 2014.
- [412] Mireille Hildebrandt. *Location Data, Purpose Binding and Contextual Integrity: What’s the Message?* Springer, 2014.
- [413] Kashmir Hill. *Your Face Belongs to Us: A Tale of AI, a Secretive Startup, and the End of Privacy*. Random House, 2023.

- [414] Risto Hilpinen. Artifact. 1999.
- [415] Crosby Hipes. The impact of a felony conviction on stigmatization in a workplace scenario. *International Journal of Law, Crime and Justice*, 56:89–99, 2019.
- [416] Kasia Hitczenko, Henry R Cowan, Matthew Goldrick, and Vijay A Mittal. Racial and ethnic biases in computational approaches to psychopathology, 2022.
- [417] Arlie Russell Hochschild. *The managed heart*. Routledge, 1983.
- [418] Arlie Russell Hochschild. *Stolen pride: Loss, shame, and the rise of the right*. The New Press, 2024.
- [419] Jesse Hoey. Clustering facial displays in context. *Technical Report TR-01-17*, 2001.
- [420] Sharona Hoffman and Andy Podgurski. The Use and Misuse of Biomedical Data: Is Bigger Really Better? *American Journal of Law & Medicine*, 39(4):497–538, December 2013.
- [421] Mia Hoffmann and Mario Mariniello. Biometric technologies at policy contribution issue n23/21. page 19, 2021.
- [422] Oliver Wendell Holmes Jr. *The Common Law*. Routledge, 1881.
- [423] Carolyn Holton. Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decision Support Systems*, 46(4):853–864, 2009.
- [424] Judith A Holton. The coding process and its challenges. *The Sage handbook of grounded theory*, 3:265–289, 2007.
- [425] Chris Jay Hoofnagle and Jan Whittington. Free: accounting for the costs of the internet’s most popular price. *UCLA L. Rev.*, 61:606, 2013.
- [426] Debra Hopkins, Jochen Kleres, Helena Flam, and Helmut Kuzmics. *Theorizing emotions: sociological explorations and applications*. Campus Verlag, 2009.
- [427] Martin Horák, Václav Stupka, and Martin Husák. Gdpr compliance in cybersecurity software: A case study of dpia in information sharing platform. In *Proceedings of the 14th international conference on availability, reliability and security*, pages 1–8, 2019.
- [428] Kimberly A Houser and W Gregory Voss. Gdpr: The end of google and facebook or a new paradigm in data privacy. *Rich. JL & Tech.*, 25:1, 2018.
- [429] Ayanna Howard, Cha Zhang, and Eric Horvitz. Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pages 1–7. IEEE, 2017.
- [430] Nicola Howe, Emma Giles, Dorothy Newbury-Birch, and Elaine McColl. Systematic review of participants’ attitudes towards data sharing: a thematic synthesis. *Journal of health services research & policy*, 23(2):123–133, 2018.

- [431] David C Howell. *Statistical methods for psychology*. Cengage Learning, 2012.
- [432] Joop J Hox and Cora JM Maas. Sample sizes for multilevel modeling. 2002.
- [433] Elise Hu. Facebook Manipulates Our Moods For Science And Commerce: A Roundup.
- [434] Kit Huckvale, Svetha Venkatesh, and Helen Christensen. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *npj Digital Medicine*, 2(1):1–11, September 2019. Number: 1 Publisher: Nature Publishing Group.
- [435] Eva Hudlicka. Guidelines for designing computational models of emotions. *International Journal of Synthetic Emotions (IJSE)*, 2(1):26–79, 2011.
- [436] Rhidian Hughes. Considering the vignette technique and its application to a study of drug injecting and hiv risk and safer behaviour. *Sociology of Health & Illness*, 20(3):381–400, 1998.
- [437] Xuan-Phung Huynh and Yong-Guk Kim. Discrimination between genuine versus fake emotion using long-short term memory with parametric bias and facial landmarks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3065–3072, 2017.
- [438] Sarah E Igo. *The known citizen: a history of privacy in modern America*. 2020. OCLC: 1111377193.
- [439] Digital Life Initiative. Facebook and Google Are the New Data Brokers, December 2018.
- [440] Robert Irvine, Douglas Boubert, Vyas Raina, Adian Liusie, Ziyi Zhu, Vineet Mudupalli, Aliaksei Korshuk, Zongyi Liu, Fritz Cremer, Valentin Assassi, et al. Rewarding chatbots for real-world engagement with millions of users. *arXiv preprint arXiv:2303.06135*, 2023.
- [441] Italian Supervisory Authority (Garante per la protezione dei dati personali). AI: the Italian Supervisory Authority fines company behind chatbot “Replika”, May 2025. Final decision date: 10 April 2025. Legal basis: GDPR Articles 5, 6, 12, 13, 24, 25. Controller: Luka Inc. Decision: administrative fine and compliance order.
- [442] Carroll E Izard. The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review*, 2(4):363–370, 2010.
- [443] Sanford M Jacoby. Employee attitude surveys in historical perspective. *Industrial Relations: A Journal of Economy and Society*, 27(1):74–93, 1988.
- [444] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. Ai-mediated communication: How the perception that profile text was written by ai affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.
- [445] William James. What is emotion? 1884. 1948.

- [446] Susan Jamieson. Likert scales: how to (ab)use them. *Medical Education*, 38(12):1217–1218, December 2004.
- [447] Guillermina Jasso. Factorial survey methods for studying beliefs and judgments. *Sociological Methods & Research*, 34(3):334–423, 2006.
- [448] Ayn de Jesus. Artificial Intelligence in Video Marketing - Emotion Recognition, Video Generation, and More.
- [449] Leslie K John, Alessandro Acquisti, and George Loewenstein. Strangers on a plane: Context-dependent willingness to divulge sensitive information. *Journal of consumer research*, 37(5):858–873, 2011.
- [450] Bobbie Johnson and Las Vegas. Privacy no longer a social norm, says Facebook founder. *The Guardian*, January 2010.
- [451] Deborah G Johnson. Computer ethics. *The Blackwell guide to the philosophy of computing and information*, pages 63–75, 2004.
- [452] Timothy Judson, Mark Haas, and Tara Lagu. Medical Identity Theft: Prevention and Reconciliation Initiatives at Massachusetts General Hospital. *The Joint Commission Journal on Quality and Patient Safety*, 40(7):291–AP1, July 2014.
- [453] Won Kyung Jung and Hun Yeong Kwon. Privacy and data protection regulations for ai using publicly available data: Clearview ai case. In *Proceedings of the 17th International Conference on Theory and Practice of Electronic Governance*, pages 48–55, 2024.
- [454] Yoshihiko Kadoya, Mostafa Saidur Rahim Khan, Somtip Watanapongvanich, and Punjapol Binnagan. Emotional status and productivity: Evidence from the special economic zone in laos. *Sustainability*, 12(4):1544, 2020.
- [455] Kimberly Barsamian Kahn, Phillip Atiba Goff, J. Katherine Lee, and Diane Motamed. Protecting Whiteness: White Phenotypic Racial Stereotypicality Reduces Police Use of Force. *Social Psychological and Personality Science*, 7(5):403–411, July 2016.
- [456] Amba Kak. Regulating biometrics: Global approaches and urgent questions. *AI Now Institute*, September, 1, 2020.
- [457] Haik Kalantarian, Khaled Jedoui, Peter Washington, Qandeel Tariq, Kaiti Dunlap, Jessey Schwartz, and Dennis P. Wall. Labeling images with facial emotion and the potential for pediatric healthcare. *Artificial Intelligence in Medicine*, 98:77–86, July 2019.
- [458] Margot E Kaminski, Matthew Rueben, William D Smart, and Cindy M Grimm. Averting robot eyes. *Md. L. Rev.*, 76:983, 2016.
- [459] Edward B Kang. On the praxes and politics of ai speech emotion recognition. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 455–466, 2023.

- [460] Arvid Kappas. What facial activity can and cannot tell us about emotions. In *The human face*, pages 215–234. Springer, 2003.
- [461] S Kapur, A G Phillips, and T R Insel. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, 17(12):1174–1179, December 2012.
- [462] Nadia Karizat, Alexandra H Vinson, Shobita Parthasarathy, and Nazanin Andalibi. Patent applications as glimpses into the sociotechnical imaginary: Ethical speculation on the imagined futures of emotion ai for mental health monitoring and detection. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–43, 2024.
- [463] Harmanpreet Kaur, Daniel McDuff, Alex C Williams, Jaime Teevan, and Shamsi T Iqbal. “i didn’t know i looked angry”: Characterizing observed emotion and reported affect at work. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.
- [464] Harmanpreet Kaur, Daniel McDuff, Alex C. Williams, Jaime Teevan, and Shamsi T. Iqbal. “I Didn’t Know I Looked Angry”: Characterizing Observed Emotion and Reported Affect at Work. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, New Orleans LA USA, April 2022. ACM.
- [465] Harmanpreet Kaur, Alex C Williams, Daniel McDuff, Mary Czerwinski, Jaime Teevan, and Shamsi T Iqbal. Optimizing for happiness and productivity: Modeling opportune moments for transitions and breaks at work. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.
- [466] Michael Kearns and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, 2019.
- [467] Danielle Keats Citron. Sexual privacy. *Yale LJ*, 128:1870, 2018.
- [468] Danielle Keats Citron. *The Fight for Privacy: Protecting Dignity, Identity, and Love in the Digital Age*. W.W. Norton & Company, 2022.
- [469] Flavius Kehr, Tobias Kowatsch, Daniel Wentzel, and Elgar Fleisch. Blissfully ignorant: the effects of general privacy concerns, general institutional trust, and affect in the privacy calculus. *Information Systems Journal*, 25(6):607–635, 2015.
- [470] Masood Mehmood Khan, Robert D Ward, and Michael Ingleby. Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature. *ACM Transactions on Applied Perception (TAP)*, 6(1):1–22, 2009.
- [471] Adnan Khashman. A modified backpropagation learning algorithm with added emotional coefficients. *IEEE transactions on neural networks*, 19(11):1896–1909, 2008.
- [472] Eugenia Kim, De’Aira Bryant, Deepak Srikanth, and Ayanna Howard. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 638–644, 2021.

- [473] HyunJin Kim, Xiaoyuan Yi, Jing Yao, Jianxun Lian, Muhua Huang, Shitong Duan, JinYeong Bak, and Xing Xie. The road to artificial superintelligence: A comprehensive survey of superalignment. *arXiv preprint arXiv:2412.16468*, 2024.
- [474] JaeWon Kim, Soobin Cho, Robert Wolfe, Jishnu Hari Nair, and Alexis Hiniker. Privacy as social norm: Systematically reducing dysfunctional privacy concerns on social media. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–39, 2025.
- [475] Pauline T Kim. Data mining and the challenges of protecting employee privacy under us law. *Comp. Lab. L. & Pol'y J.*, 40:405, 2018.
- [476] Pauline T Kim and Matthew T Bodie. Artificial intelligence and the challenges of workplace discrimination and privacy. *Journal of Labor and Employment Law*, 35(2):289–315, 2021.
- [477] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 3687–3691. IEEE, 2013.
- [478] Yeolib Kim, Boreum Choi, and Yoonhyuk Jung. Individual differences in online privacy concern. *Asia Pacific Journal of Information Systems*, 28(4):274–289, 2018.
- [479] Everlyne Kimani, Kael Rowan, Daniel McDuff, Mary Czerwinski, and Gloria Mark. A conversational agent in support of productivity and wellbeing at work. In *2019 8th international conference on affective computing and intelligent interaction (ACII)*, pages 1–7. IEEE, 2019.
- [480] Michael Kipp. Multimedia annotation, querying and analysis in anvil. *Multimedia information extraction*, 19, 2012.
- [481] Stuart A. Kirk and Herb Kutchins. Deliberate Misdiagnosis in Mental Health Practice. *Social Service Review*, 62(2):225–237, June 1988.
- [482] Olivia J. Kirtley and Rory C. O’Connor. Suicide prevention is everyone’s business: Challenges and opportunities for Google. *Social Science & Medicine*, page 112691, January 2020.
- [483] Funda Kivran-Swaine, Jeremy Ting, Jed Richards Brubaker, Rannie Teodoro, and Mor Naaman. Understanding Loneliness in Social Awareness Streams: Expressions and Responses. page 10.
- [484] Dorothea Kleine. *Technologies of choice?: ICTs, development, and the capabilities approach*. MIT press, 2013.
- [485] Oliver Klingefjord, Ryan Lowe, and Joe Edelman. What are human values, and how do we align ai to them? *arXiv preprint arXiv:2404.10636*, 2024.
- [486] Clyde Kluckhohn. *2. Values and value-orientations in the theory of action: An exploration in definition and classification*. Harvard University Press, 2013.
- [487] Cory Knobel and Geoffrey C Bowker. Values in design. *Communications of the ACM*, 54(7):26–28, 2011.

- [488] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 128:1455–1484, 2020.
- [489] Laura L Koppes. *Historical perspectives in industrial and organizational psychology*. Psychology Press, 2014.
- [490] Tomek Korbak, Mikita Balesni, Buck Shlegeris, and Geoffrey Irving. How to evaluate control measures for llm agents? a trajectory from today to superintelligence. *arXiv preprint arXiv:2504.05259*, 2025.
- [491] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1022–1040, 2019.
- [492] Nikolaos Koutsouleris, Tobias U Hauser, Vasilisa Skvortsova, and Munmun De Choudhury. From promise to practice: towards the realisation of ai-informed mental health care. *The Lancet Digital Health*, 2022.
- [493] Margaret Bull Kovera. Report on eyewitness identification issues identified in robert julian-borchak williams v. city of detroit, detroit police chief james craig and detective donald bussa prepared for university of michigan civil rights litigation initiative may 26, 2023. 2023.
- [494] Dimitri Kraft, Kristof Van Laerhoven, and Gerald Bieber. Carecam: Concept of a new tool for corporate health management. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, pages 585–593, 2021.
- [495] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788, June 2014.
- [496] Sylvia D Kreibig. Autonomic nervous system activity in emotion: A review. *Biological psychology*, 84(3):394–421, 2010.
- [497] Brian Kropp. The Future of Employee Monitoring, May 2019.
- [498] Eva G Krumhuber and Antony SR Manstead. Can duchenne smiles be feigned? new evidence on felt and false smiles. *Emotion*, 9(6):807, 2009.
- [499] Susan M Kruml and Deanna Geddes. Exploring the dimensions of emotional labor: The heart of hochschild’s work. *Management communication quarterly*, 14(1):8–49, 2000.
- [500] Olivia Kuenzi. Ftc non-compete ban. *The Reporter: Social Justice Law Center Magazine*, 2023(1):11, 2023.
- [501] Caitlin Kuhlman, Latifa Jackson, and Rumi Chunara. No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv preprint arXiv:2002.11836*, 2020.

- [502] Ponnurangam Kumaraguru and Lorrie Faith Cranor. Privacy indexes: a survey of westin's studies. 2005.
- [503] Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, et al. Specific versus general principles for constitutional ai. *arXiv preprint arXiv:2310.13798*, 2023.
- [504] Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *Artificial Intelligence Review*, 56(12):15129–15215, 2023.
- [505] Luis Kutner. Due process of euthanasia: the living will, a proposal. *Ind. LJ*, 44:539, 1968.
- [506] Jangho Kwon, Da-Hye Kim, Wanjoo Park, and Laehyun Kim. A wearable device for emotional recognition using facial expression and physiological response. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5765–5768. IEEE, 2016.
- [507] Max Landauer, Klaus Mayer, Florian Skopik, Markus Wurzenberger, and Manuel Kern. Red team redemption: A structured comparison of open-source tools for adversary emulation. In *2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 117–128. IEEE, 2024.
- [508] Eric Lander and Alondra Nelson. ICYMI: WIRED (Opinion): Americans Need a Bill of Rights for an AI-Powered World, October 2021.
- [509] Joan B Landes. More than words: The printing press and the french revolution, 1991.
- [510] Christopher B Landis and Joshua A Kroll. Mitigating inference risks with the nist privacy framework. *Proceedings on Privacy Enhancing Technologies*, 2024.
- [511] Agnieszka Landowska, Aleksandra Karpus, Teresa Zawadzka, Ben Robins, Duygun Erol Barkana, Hatice Kose, Tatjana Zorcec, and Nicholas Cummins. Automatic emotion recognition in children with autism: a systematic literature review. *Sensors*, 22(4):1649, 2022.
- [512] Carl Georg Lange. The mechanism of the emotions. *The classical psychologists*, pages 672–684, 1885.
- [513] Harold D Lasswell and Myres S McDougal. *Jurisprudence For a Free Society: Studies in Law, Science and Policy, Volume II*, volume 7. Martinus Nijhoff Publishers, 1992.
- [514] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018.

- [515] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Bornet, Trinh Minh Tri Do, Olivier Dousse, Julien Eberle, and Markus Miettinen. From big smartphone data to worldwide research: The mobile data challenge. *Pervasive and Mobile Computing*, 9(6):752–771, 2013.
- [516] Theresa Law, Meia Chita-Tegmark, and Matthias Scheutz. The interplay between emotional intelligence, trust, and gender in human–robot interaction. *International Journal of Social Robotics*, 13(2):297–309, 2021.
- [517] Elizabeth M. Lawrence. Why Do College Graduates Behave More Healthfully than Those Who Are Less Educated? *Journal of Health and Social Behavior*, 58(3):291–306, 2017.
- [518] Christopher A Le Dantec, Erika Shehan Poole, and Susan P Wyche. Values as lived experience: evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1141–1150, 2009.
- [519] Joseph LeDoux. Rethinking the emotional brain. *Neuron*, 73(4):653–676, 2012.
- [520] Hwansoo Lee, Siew Fan Wong, Jungjoo Oh, and Younghoon Chang. Information privacy concerns and demographic characteristics: Data from a korean media panel survey. *Government Information Quarterly*, 36(2):294–303, 2019.
- [521] Hyunsoo Lee, Soowon Kang, and Uichin Lee. Understanding privacy risks and perceived benefits in open dataset collection for mobile affective computing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–26, 2022.
- [522] Nicol Turner Lee. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 16(3):252–260, 2018.
- [523] James Legge et al. *Confucian analects: The great learning, and the doctrine of the mean*. Courier Corporation, 1971.
- [524] Pedro Giovanni Leon, Justin Cranshaw, Lorrie Faith Cranor, Jim Graves, Manoj Hastak, Blase Ur, and Guzi Xu. What do online behavioral advertising privacy disclosures communicate to users? In *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*, pages 19–30, 2012.
- [525] Ada Lerner, Helen Yuxun He, Anna Kawakami, Silvia Catherine Zeamer, and Roberto Hoyle. Privacy and activism in the transgender community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.
- [526] Holly Lewis. *The politics of everybody: Feminism, queer theory, and marxism at the intersection*. Bloomsbury Publishing, 2016.
- [527] Han Li, Xin Robert Luo, Jie Zhang, and Heng Xu. Resolving the privacy paradox: Toward a cognitive appraisal and emotion approach to online privacy behaviors. *Information & management*, 54(8):1012–1022, 2017.

- [528] Han Li, Rathindra Sarathy, and Heng Xu. The role of affect and cognition on online consumers' decision to disclose personal information to unfamiliar online vendors. *Decision Support Systems*, 51(3):434–445, 2011.
- [529] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. Algorithmic hiring in practice: Recruiter and hr professional's perspectives on ai use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 166–176, 2021.
- [530] Liandong Li, Tadas Baltrusaitis, Bo Sun, and Louis-Philippe Morency. Combining sequential geometry and texture features for distinguishing genuine and deceptive emotions. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3147–3153, 2017.
- [531] Wei Li, Wei Huan, Bowen Hou, Ye Tian, Zhen Zhang, and Aiguo Song. Can emotion be transferred?—a review on transfer learning for eeg-based emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(3):833–846, 2021.
- [532] Annie Liang, Jay Lu, and Xiaosheng Mu. Algorithmic Design: Fairness Versus Accuracy. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 58–59, Boulder CO USA, July 2022. ACM.
- [533] Yunji Liang, Xiaolong Zheng, and Daniel D Zeng. A survey on big data-driven digital phenotyping of mental health. *Information Fusion*, 52:290–307, 2019.
- [534] Cindy Lin and Silvia Margot Lindtner. Techniques of use: Confronting value systems of productivity, progress, and usefulness in computing and design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.
- [535] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 501–510, 2012.
- [536] Patrick Lin, Keith Abney, and George A Bekey. *Robot ethics: the ethical and social implications of robotics*. MIT press, 2014.
- [537] Bruce G Link and Jo C Phelan. Conceptualizing stigma. *Annual review of Sociology*, pages 363–385, 2001.
- [538] Bruce G Link and Jo C Phelan. Stigma and its public health implications. *The Lancet*, 367(9509):528–529, 2006.
- [539] Kevin Litman-Navarro. Opinion | We Read 150 Privacy Policies. They Were an Incomprehensible Disaster. *The New York Times*, June 2019.
- [540] Melissa M Littlefield and Jenell Johnson. *The neuroscientific turn: Transdisciplinarity in the age of the brain*. University of Michigan Press, 2012.
- [541] George Loewenstein and Jennifer S Lerner. The role of affect in decision making. 2003.

- [542] Jorge Lopez-Castroman, Bilel Moulahi, Jérôme Azé, Sandra Bringay, Julie Deninotti, Sébastien Guillaume, and Enrique Baca-Garcia. Mining social networks to improve suicide prevention: A scoping review. *Journal of Neuroscience Research*, 98(4):616–625, 2020. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jnr.24404>.
- [543] Brian Lubars and Chenhao Tan. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. *CoRR*, abs/1902.03245, 2019.
- [544] Kem-Laurin Lubin and Lai-Tze Fan. Rethinking the rhetoric of surveillance in public safety: A critical discourse analysis. In *Future of Information and Communication Conference*, pages 657–677. Springer, 2025.
- [545] Steven G Luke. Evaluating significance in linear mixed-effects models in r. *Behavior research methods*, 49:1494–1502, 2017.
- [546] Catherine Lutz and Geoffrey M White. The anthropology of emotions. *Annual review of anthropology*, 15(1):405–436, 1986.
- [547] Jennifer Lynch. Face Off: Law Enforcement Use of Face Recognition Technology, February 2018.
- [548] Richard Lynn. The intelligence of the mongoloids: A psychometric, evolutionary and neurological theory. *Personality and individual differences*, 8(6):813–844, 1987.
- [549] Orla Lynskey. Deconstructing data protection: the ‘added-value’of a right to data protection in the eu legal order. *International & Comparative Law Quarterly*, 63(3):569–597, 2014.
- [550] David Lyon. Surveillance studies: An overview. 2007.
- [551] David Lyon. *Identifying citizens: ID cards as surveillance*. Polity, 2009.
- [552] David Lyon. Identification, surveillance and democracy. In *Surveillance and democracy*, pages 50–66. Routledge-Cavendish, 2010.
- [553] Kim Lyons. New FTC memo calls for a focus on ‘structural dominance’ from big companies, September 2021.
- [554] Catharine A MacKinnon. *Toward a feminist theory of the state*. Harvard University Press, 1989.
- [555] Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. Social sensing for epidemiological behavior change. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, UbiComp ’10, pages 291–300, New York, NY, USA, September 2010. Association for Computing Machinery.
- [556] Muhammad Meftah Mafazy, Chastine Fatichah, and Anny Yuniarti. Audio feature analysis and selection for deception detection in court proceedings. *JUTI: Jurnal Ilmiah Teknologi Informasi*, pages 13–28, 2025.

- [557] Ivy W. Maina, Tanisha D. Belton, Sara Ginzberg, Ajit Singh, and Tiffani J. Johnson. A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test. *Social Science & Medicine*, 199:219–229, February 2018.
- [558] Gianclaudio Malgieri and Jędrzej Niklas. Vulnerable data subjects. *Computer Law & Security Review*, 37:105415, 2020.
- [559] Naresh K Malhotra, Sung S Kim, and James Agarwal. Internet users' information privacy concerns (uiipc): The construct, the scale, and a causal model. *Information systems research*, 15(4):336–355, 2004.
- [560] G Mancia, M Di Rienzo, and G Parati. Ambulatory blood pressure monitoring use in hypertension research and clinical practice. *Hypertension*, 21(4):510–524, 1993.
- [561] Lydia Manikonda and Munmun De Choudhury. Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 170–181, New York, NY, USA, May 2017. Association for Computing Machinery.
- [562] Alessandro Mantelero. Ai and big data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, 34(4):754–772, 2018.
- [563] Peter Mantello, Manh-Tung Ho, Minh-Hoang Nguyen, and Quan-Hoang Vuong. Bosses without a heart: socio-demographic and cross-cultural determinants of attitude toward emotional ai in the workplace. *AI & society*, pages 1–23, 2021.
- [564] Stephen T Margulis. Conceptions of privacy: Current status and next steps. *Journal of Social Issues*, 33(3):5–21, 1977.
- [565] Ereni Markos, George R Milne, and James W Peltier. Information sensitivity and willingness to provide continua: a comparative privacy study of the united states and brazil. *Journal of Public Policy & Marketing*, 36(1):79–96, 2017.
- [566] Mason Marks. Artificial Intelligence-Based Suicide Prediction. *ARTIFICIAL INTELLIGENCE*, page 24, 2019.
- [567] Stacy Marsella and Jonathan Gratch. Computationally modeling human emotion. *Commun. ACM*, 57(12):56–67, nov 2014.
- [568] Stacy Marsella, Jonathan Gratch, Paolo Petta, et al. Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual*, 11(1):21–46, 2010.
- [569] Kelly D Martin and Patrick E Murphy. The role of data privacy in marketing. *Journal of the Academy of Marketing Science*, 45:135–155, 2017.
- [570] Kirsten Martin. Privacy notices as tabula rasa: An empirical investigation into how complying with a privacy notice is related to meeting privacy expectations online. *Journal of Public Policy & Marketing*, 34(2):210–227, 2015.

- [571] Kirsten Martin. The penalty for privacy violations: How privacy violations impact trust online. *Journal of Business Research*, 82:103–116, 2018.
- [572] Kirsten Martin. Ethical implications and accountability of algorithms. *Journal of Business Ethics*, 160(4):835–850, 2019.
- [573] Kirsten Martin and Helen Nissenbaum. Measuring privacy: An empirical test using context to expose confounding variables. *Colum. Sci. & Tech. L. Rev.*, 18:176, 2016.
- [574] Kirsten Martin and Helen Nissenbaum. What is it about location? *Berkeley Tech. LJ*, 35:251, 2020.
- [575] Kirsten Martin and Katie Shilton. Why experience matters to privacy: How context-based experience moderates consumer privacy expectations for mobile applications. *Journal of the Association for Information Science and Technology*, 67(8):1871–1882, 2016.
- [576] Kirsten E Martin. Diminished or just different? a factorial vignette study of privacy as a social contract. *Journal of Business Ethics*, 111(4):519–539, 2012.
- [577] Kirsten E. Martin and Helen Nissenbaum. Measuring Privacy: An Empirical Test Using Context To Expose Confounding Variables. SSRN Scholarly Paper ID 2709584, Social Science Research Network, Rochester, NY, December 2015.
- [578] Nigel Martin, John Rice, and Robin Martin. Expectations of privacy and trust: Examining the views of it professionals. *Behaviour & Information Technology*, 35(6):500–510, 2016.
- [579] Alice E Marwick and Danah Boyd. Understanding privacy at the margins. *International Journal of Communication (19328036)*, 12, 2018.
- [580] Karl Marx. Economic & philosophical manuscripts of 1844. In *Marx/Engels Collected Works*, Vol. 3, pages 229–348. 1975.
- [581] Eli Zimmerman Twitter Eli has been eagerly pursuing a journalistic career since he left the University of Maryl, 's Philip Merrill School of Journalism Previously, Eli was a staff reporter for medical trade publication Frontline Medical News, Where He Experienced the Impact of Continuous Education, evolving teaching methods through the medical lens When not in the office, and Eli is busy scanning the web for the latest podcasts or stepping into the boxing ring for a few rounds. GoGuardian Develops a New AI-Enabled Cloud Filter for K–12 Schools.
- [582] David Matsumoto, Sachiko Takeuchi, Sari Andayani, Natalia Kouznetsova, and Deborah Krupp. The contribution of individualism vs. collectivism to cross-national differences in display rules. *Asian journal of social psychology*, 1(2):147–165, 1998.
- [583] David J Mattson and Susan G Clark. Human dignity in concept and practice. *Policy Sciences*, 44:303–319, 2011.
- [584] Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, 114(48):12714–12719, 2017.

- [585] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.
- [586] John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4):12–12, 2006.
- [587] R McCarty. The fight-or-flight response: A cornerstone of stress research. In *Stress: Concepts, cognition, emotion, and behavior*, pages 33–37. Elsevier, 2016.
- [588] Patrick McCole, Elaine Ramsey, and John Williams. Trust considerations on attitudes towards online purchasing: The moderating effect of privacy and security concerns. *Journal of Business Research*, 63(9-10):1018–1024, 2010.
- [589] John McCormick. What AI Can Tell From Listening to You. *Wall Street Journal*, April 2019.
- [590] Nora McDonald and Andrea Forte. The politics of privacy theories: Moving from norms to vulnerabilities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.
- [591] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.
- [592] Daniel McDuff, Eunice Jun, Kael Rowan, and Mary Czerwinski. Longitudinal observational evidence of the impact of emotion regulation strategies on affective expression. *IEEE Transactions on Affective Computing*, 12(3):636–647, 2019.
- [593] Daniel McDuff, Eunice Jun, Kael Rowan, and Mary Czerwinski. Longitudinal Observational Evidence of the Impact of Emotion Regulation Strategies on Affective Expression. *IEEE Transactions on Affective Computing*, 12(3):636–647, July 2021. Conference Name: IEEE Transactions on Affective Computing.
- [594] Derrick McIver, Mark L Lengnick-Hall, and Cynthia A Lengnick-Hall. A strategic approach to workforce analytics: Integrating science and agility. *Business Horizons*, 61(3):397–407, 2018.
- [595] Kwame McKenzie and Kamaldeep Bhui. Institutional racism in mental health care. *BMJ : British Medical Journal*, 334(7595):649–650, March 2007.
- [596] Michael McLaughlin and Daniel Castro. The critics were wrong: Nist data shows the best facial recognition algorithms are neither racist nor sexist. Technical report, Information Technology and Innovation Foundation, 2020.
- [597] Darrin M McMahon. The quest for happiness. *Wilson Quarterly*, 29(1):62–71, 2005.

- [598] Roisin McNaney, Catherine Morgan, Pranav Kulkarni, Julio Vega, Farnoosh Heidarivincheh, Ryan McConville, Alan Whone, Mickey Kim, Reuben Kirkham, and Ian Craddock. Exploring perceptions of cross-sectoral data sharing with people with parkinson’s. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2022.
- [599] Joanne McNeil. Crisis Text Line and the Silicon Valleyfication of Everything. *Vice*, February 2022.
- [600] Andrew McStay. *Emotional AI: The rise of empathic media*. Sage, 2018.
- [601] Andrew McStay. Emotional ai, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data & Society*, 7(1):2053951720904386, 2020.
- [602] Karen McVeigh. Samaritans Twitter app identifying user’s moods criticised as invasive. *The Guardian*, November 2014.
- [603] Malek Mechergui and Sarath Sreedharan. Goal alignment: re-analyzing value alignment problems using human-aware ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10110–10118, 2024.
- [604] Alexander Meinke, Bronson Schoen, Jérémie Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.
- [605] Annelise M Mennicke and Katie Ropes. Estimating the rate of domestic violence perpetrated by law enforcement officers: A review of methods and estimates. *Aggression and Violent Behavior*, 31:157–164, 2016.
- [606] Robinson Meyer. Everything We Know About Facebook’s Secret Mood Manipulation Experiment, June 2014. Section: Technology.
- [607] Sabelo Mhlambi. From rationality to relationality: ubuntu as an ethical and human rights framework for artificial intelligence governance. *Carr Center for Human Rights Policy Discussion Paper Series*, 9, 2020.
- [608] Jude Mikal, Samantha Hurst, and Mike Conway. Ethical issues in using twitter for population-level depression monitoring: a qualitative study. *BMC medical ethics*, 17(1):22, 2016.
- [609] John Stuart Mill. *On liberty and other essays*. Oxford University Press, USA, 1998.
- [610] George R Milne, George Pettinico, Fatima M Hajjat, and Ereni Markos. Information sensitivity typology: Mapping the degree and type of risk consumers perceive in personal data sharing. *Journal of Consumer Affairs*, 51(1):133–161, 2017.
- [611] Matti Minkkinen and Matti Mäntymäki. Discerning between the “easy” and “hard” problems of ai governance. *IEEE Transactions on Technology and Society*, 4(2):188–194, 2023.
- [612] Lawrence H Mirel. The limits of governmental inquiry into the private lives of government employees. *BUL rev.*, 46:1, 1966.

- [613] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado, June 5 2015. Association for Computational Linguistics.
- [614] Karin Mogg and Brendan P Bradley. A cognitive-motivational analysis of anxiety. *Behaviour research and therapy*, 36(9):809–848, 1998.
- [615] Francesca Mongelli, Penelope Georgakopoulos, and Michele T. Pato. Challenges and Opportunities to Meet the Mental Health Needs of Underserved and Disenfranchised Populations in the United States. *FOCUS*, 18(1):16–24, January 2020. Publisher: American Psychiatric Publishing.
- [616] Scott Monteith, Tasha Glenn, John Geddes, Peter C. Whybrow, and Michael Bauer. Commercial Use of Emotion Artificial Intelligence (AI): Implications for Psychiatry. *Current Psychiatry Reports*, 24(3):203–211, March 2022.
- [617] James H Moor. Towards a theory of privacy in the information age. *ACM Sigcas Computers and Society*, 27(3):27–32, 1997.
- [618] Barrington Moore Jr. *Privacy: Studies in Social and Cultural History: Studies in Social and Cultural History*. Routledge, 1984.
- [619] Pegah Moradi and Karen Levy. The future of work in the age of ai: Displacement or risk-shifting? 2020.
- [620] Jan MorÉn, Christian Balkenius. Emotional learning: a computational model of the amygdala. *Cybernetics & Systems*, 32(6):611–636, 2001.
- [621] Marcello Mortillaro, Ben Meuleman, and Klaus R Scherer. Advocating a componential appraisal model to guide emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 3(1):18–32, 2012.
- [622] Sumitava Mukherjee, Jaison A Manjaly, and Maithilee Nargundkar. Money makes you reveal more: consequences of monetary cues on preferential disclosure of personal information. *Frontiers in psychology*, 4:839, 2013.
- [623] Deirdre K Mulligan, Daniel Klutzz, and Nitin Kohli. Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. *Available at SSRN 3311894*, 2019.
- [624] Prasanth Murali, Javier Hernandez, Daniel McDuff, Kael Rowan, Jina Suh, and Mary Czerwinski. Affectivespotlight: Facilitating the communication of affective responses from audience members during online presentations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.
- [625] Dan Muriello, Lizzy Donahue, Danny Ben-David, Umut Ozertem, and Reshef Shilon. Under the hood: Suicide prevention tools powered by AI, February 2018. Section: ML Applications.

- [626] Salla Muuraiskangas, Marja Harjumaa, Kirsikka Kaipainen, Miikka Ermes, et al. Process and effects evaluation of a digital mental health intervention targeted at improving occupational well-being: lessons from an intervention study with failed adoption. *JMIR mental health*, 3(2):e4465, 2016.
- [627] Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products. *Journal of Medical Internet Research*, 15(8):e174, 2013. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.
- [628] Guillaume Nadon, Marcus Feilberg, Mathias Johansen, and Irina Shklovski. In the user we trust: Unrealistic expectations of facebook's privacy mechanisms. In *Proceedings of the 9th International Conference on Social Media and Society*, pages 138–149, 2018.
- [629] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. Privacy expectations and preferences in an iot world. In *Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017)*, pages 399–412, 2017.
- [630] Thomas Nagel. Concealment and exposure. *Philosophy & Public Affairs*, 27(1):3–30, 1998.
- [631] Jeff Nagy. Autism and the making of emotion ai: Disability as resource for surveillance capitalism. *New media & society*, page 14614448221109550, 2022.
- [632] Karen Nakamura. My algorithms have determined you're not human: Ai-ml, reverse turing-tests, and the disability experience. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 1–2, New York, NY, USA, 2019. Association for Computing Machinery.
- [633] M. Namara, H. Sloan, P. Jaiswal, and B. P. Knijnenburg. The potential for user-tailored privacy on facebook. In *2018 IEEE Symposium on Privacy-Aware Computing (PAC)*, pages 31–42, 2018.
- [634] Shushi Namba, Russell S Kabir, Makoto Miyatani, and Takashi Nakao. Dynamic displays enhance the ability to discriminate genuine and posed facial expressions of emotion. *Frontiers in psychology*, 9:672, 2018.
- [635] Gérard Näring, Mariette Briët, and André Brouwers. Beyond demand–control: Emotional labour and symptoms of burnout in teachers. *Work & Stress*, 20(4):303–315, 2006.
- [636] Natasha Singer. Creepy or Not? Your Privacy Concerns Probably Reflect Your Politics. *New York Times (Online)*, April 2018. Journal Abbreviation: New York Times (Online) Publisher: New York Times Company.
- [637] Yefim Natis and Jason Daigler. Gartner's Top Strategic Predictions for 2020 and Beyond: Technology Changes the Human Condition, October 2019.

- [638] Mary Neal. Respect for human dignity as ‘substantive basic norm’. *International Journal of Law in Context*, 10(1):26–46, 2014.
- [639] Tess MS Neal, Christopher Slobogin, Michael J Saks, David L Faigman, and Kurt F Geisinger. Psychological assessments in legal contexts: Are courts keeping “junk science” out of the courtroom? *Psychological Science in the Public Interest*, 20(3):135–164, 2019.
- [640] Glenn Negley. Philosophical views on the value of privacy. *Law & Contemp. Probs.*, 31:319, 1966.
- [641] James P. Nehf. The FTC’s Proposed Framework for Privacy Protection Online: A Move toward Substantive Controls or Just More Notice and Choice Electronic Commerce Law. *William Mitchell Law Review*, 37(4):1727–1744, 2010.
- [642] Allen Newell and Herbert A Simon. Computer science as empirical inquiry: Symbols and search. In *ACM Turing award lectures*, page 1975. 2007.
- [643] Matthew Newland. Justin eh smith, irrationality: A history of the dark side of reason. princeton nj, princeton university press, 2019, 344 p., 16.2× 23.6 cm, isbn 978-0-69118-966-6. *Science et Esprit*, 74(2-3):439–441, 2022.
- [644] Casey Newton. How Facebook is preparing for a surge in depressed and anxious users, March 2020.
- [645] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.
- [646] Helen Nissenbaum. Privacy in context: Technology, policy, and the integrity of social life. In *Privacy in Context*. Stanford University Press, 2009.
- [647] Helen Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, USA, 2009.
- [648] Helen Nissenbaum. A contextual approach to privacy online. *Daedalus*, 140(4):32–48, 2011.
- [649] Helen Nissenbaum. Respecting context to protect privacy: Why meaning matters. *Science and engineering ethics*, 24(3):831–852, 2018.
- [650] Helen Nissenbaum. Contextual integrity up and down the data food chain. *Theoretical Inquiries in Law*, 20(1):221–256, 2019.
- [651] Helen Nissenbaum. University of washington’s department of human centered design & engineering 2021 distinguished lecture “contextual integrity: Breaking the grip of public-private distinction for meaningful privacy”. <https://www.youtube.com/watch?v=VPwmC0Sfe50>, 2021. Accessed: 12/15/2023.

- [652] Kobbi Nissim and Alexandra Wood. Is privacy privacy? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170358, 2018.
- [653] Steven L Nock and Thomas M Guterbock. Survey experiments. *Handbook of survey research*, 2:837–865, 2010.
- [654] Selin E Nugent and Susan Scott-Parker. Recruitment ai has a disability problem: anticipating and mitigating unfair automated hiring decisions, Sep 2021.
- [655] Leysan Nurgalieva, David O’Callaghan, and Gavin Doherty. Security and privacy of mhealth applications: a scoping review. *IEEE Access*, 8:104247–104268, 2020.
- [656] Martha Nussbaum. Interview with bill moyers on *A World of Ideas*. Public Broadcasting Service (PBS), 1988. Transcript accessed 2025-02-03.
- [657] Martha C Nussbaum. Capabilities and human rights. *Fordham L. Rev.*, 66:273, 1997.
- [658] Martha C Nussbaum. *Women and human development: The capabilities approach*, volume 3. Cambridge university press, 2000.
- [659] Martha C Nussbaum. *Upheavals of thought: The intelligence of emotions*. Cambridge University Press, 2003.
- [660] Martha C Nussbaum. *Creating capabilities: The human development approach*. Harvard University Press, 2011.
- [661] Martha C Nussbaum. *The monarchy of fear: A philosopher looks at our political crisis*. Simon & Schuster, 2019.
- [662] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. page 8, 2019.
- [663] ACLU of Illinois. In big win, settlement ensures clearview ai complies with groundbreaking illinois biometric privacy law. Press Release. Case: ACLU v. Clearview AI.
- [664] German Presidency of the Council of the European Union. Berlin declaration on digital society and value-based digital government, December 2020.
- [665] Songhee Oh, Jae Heon Kim, Sung-Woo Choi, Hee Jeong Lee, Jungrak Hong, and Soon Hyo Kwon. Physician Confidence in Artificial Intelligence: An Online Mobile Survey. *Journal of Medical Internet Research*, 21(3):e12422, March 2019. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.
- [666] Susan Moller Okin and Gender Justice. the family. *New York*, 1989.

- [667] Judith S Olson, Jonathan Grudin, and Eric Horvitz. A study of preferences for sharing and privacy. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1985–1988, 2005.
- [668] Gloria Omale. Gartner Identifies Three Most Common AI Use Cases in HR and Recruiting, June 2019.
- [669] Bat-Ami Bar On. Marginality and epistemic privilege. In *Feminist epistemologies*, pages 83–100. Routledge, 2013.
- [670] Serena Oppenheim. How The Corporate Wellness Market Has Exploded: Meet The Latest Innovators In The Space, June 2019. Section: Healthcare.
- [671] Bozhan Orozov and Daniela Orozova. Rule-based system against reinforcement learning. In *Proceedings of the 22nd International Conference on Computer Systems and Technologies*, pages 67–70, 2021.
- [672] George Orwell. *1984*. Signet Classic, 1949.
- [673] Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. The shifted and the overlooked: A task-oriented investigation of user-gpt interactions. *arXiv preprint arXiv:2310.12418*, 2023.
- [674] Devah Pager. The mark of a criminal record. *American journal of sociology*, 108(5):937–975, 2003.
- [675] Joyojeet Pal. Chi4good or good4chi. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, pages 709–721, 2017.
- [676] Claire Su-Yeon Park. Threshold of dignity. *Clinical Nurse Specialist*, 39(3):162, 2025.
- [677] Yong Jin Park, Scott W Campbell, and Nojin Kwak. Affect, cognition and reward: Predictors of privacy protection online. *Computers in Human Behavior*, 28(3):1019–1027, 2012.
- [678] Frank A. Pasquale. More Than a Feeling, October 2020.
- [679] Vishal Patel, Austin Chesmore, Christopher M Legner, and Santosh Pandey. Trends in workplace wearable technologies and connected-worker solutions for next-generation occupational safety, health, and productivity. *Advanced Intelligent Systems*, 4(1):2100099, 2022.
- [680] Jessica A Pater, Oliver L Haimson, Nazanin Andalibi, and Elizabeth D Mynatt. “hunger hurts but starving works” characterizing the presentation of eating disorders online. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1185–1200, 2016.
- [681] Michael Quinn Patton et al. Qualitative evaluation methods. 1980.
- [682] Amol Patwardhan and Gerald Knapp. Augmenting supervised emotion recognition with rule-based decision model. *arXiv preprint arXiv:1607.02660*, 2016.

- [683] Maria Payri, Michael Cohn, and Ilene R Shaw. How often is employee anger an insider risk i? detecting and measuring negative sentiment versus insider risk in digital communications. *The Journal of Digital Forensics, Security and Law: JDFSL*, 8(1):39, 2013.
- [684] Sachin R Pendse, Daniel Nkemelu, Nicola J Bidwell, Sushrut Jadhav, Soumitra Pathare, Mummun De Choudhury, and Neha Kumar. From Treatment to Healing:Envisioning a Decolonial Digital Mental Health. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–23, New York, NY, USA, April 2022. Association for Computing Machinery.
- [685] Janice Penni. The future of online social networks (OSN): A measurement analysis using social media tools and application. *Telematics and Informatics*, 34(5):498–517, August 2017.
- [686] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.
- [687] Monica Perusquía-Hernández. Are people happy when they smile?: Affective assessments based on automatic smile genuineness identification. *Emotion Studies*, 6(1):57–71, 2021.
- [688] Luiz Pessoa. On the relationship between emotion and cognition. *Nature reviews neuroscience*, 9(2):148–158, 2008.
- [689] Sandra Petronio. *Boundaries of privacy: Dialectics of disclosure*. Suny Press, 2002.
- [690] Rosalind W Picard. Affective computing for hci. In *HCI (1)*, pages 829–833. Citeseer, 1999.
- [691] Rosalind W Picard. Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1-2):55–64, 2003.
- [692] Martyn Pickersgill. Digitising psychiatry? Sociotechnical expectations, performative nominalism and biomedical virtue in (digital) psychiatric praxis. *Sociology of Health & Illness*, 41(S1):16–30, 2019. eprint: <https://onlinelibrary.wiley.com/doi/10.1111/1467-9566.12811>.
- [693] José Pinheiro and Douglas Bates. *Mixed-effects models in S and S-PLUS*. Springer science & business media, 2006.
- [694] Helen Poitevin. Gartner: Fueling the Future of Business, August 2015.
- [695] Helen Poitevin. Gartner: Workplace Analytics Needs Digital Ethics, November 2015.
- [696] Harold A. Pollack and Keith Humphreys. Reducing Violent Incidents between Police Officers and People with Psychiatric or Substance Use Disorders. *The ANNALS of the American Academy of Political and Social Science*, 687(1):166–184, January 2020.

- [697] Shraddha Pophale, Hetal Gandhi, and Anil Kumar Gupta. Emotion Recognition Using Chatbot System. In Vinit Kumar Gunjan and Jacek M. Zurada, editors, *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, volume 1245, pages 579–587. Springer Singapore, Singapore, 2021. Series Title: Advances in Intelligent Systems and Computing.
- [698] Keith Porcard. The real harm of crisis text line’s data sharing. *Wired*, 2022.
- [699] Robert C. Post. The social foundations of privacy: Community and self in the common law tort. *California Law Review*, 77(5):957–1010, 1989. © 1989 by California Law Review, Inc.
- [700] Rosamund Powell. The eu ai act: National security implications.
- [701] Jesse J. Prinz. *Gut Reactions: A Perceptual Theory of the Emotions*. Oxford University Press, 2004.
- [702] William L Prosser. California law review vol. 48 august 1960 no. 3: Privacy. In *Pre-Nineteen Sixty Developments in the Bill of Rights Area*, pages 355–395. Routledge, 1960.
- [703] Karen Pugliesi. The consequences of emotional labor: Effects on work stress, job satisfaction, and well-being. *Motivation and emotion*, 23(2):125–154, 1999.
- [704] Matthieu Queloz. The dworkin–williams debate: Liberty, conceptual integrity, and tragic conflict in politics. *Philosophy and Phenomenological Research*, 109(1):3–29, 2024.
- [705] James Rachels. Why is privacy important? philosophical dimensions of privacy: An anthology. fd schoeman, 1984.
- [706] Anat Rafaeli and Robert I Sutton. The expression of emotion in organizational life. *Research in organizational behavior*, 11(1):1–42, 1989.
- [707] Wojciech Rafał Wiewiórowski. Shaping a Safer Digital Future: a New Strategy for a New Decade | European Data Protection Supervisor, January 2025.
- [708] Manish Raghavan, Solon Barocas, Jon M. Kleinberg, and Karen Levy. Mitigating bias in algorithmic employment screening: Evaluating claims and practices. *CoRR*, abs/1906.09208, 2019.
- [709] Lee Rainie, Cary Funk, Monica Anderson, and Alec Tyson. AI and Human Enhancement: Americans’ Openness is Tempered by a Range of Concerns. (March 2022):164, March 2022.
- [710] Roope Raisamo, Ismo Rakkolainen, Päivi Majaranta, Katri Salminen, Jussi Rantala, and Ahmed Farooq. Human augmentation: Past, present and future. *International Journal of Human-Computer Studies*, 131:131–143, 2019.
- [711] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

- [712] Neomi Rao. Three concepts of dignity in constitutional law. *Notre Dame L. Rev.*, 86:183, 2011.
- [713] Manuel Rausch and Michael Zehetleitner. A comparison between a visual analogue scale and a four point scale as measures of conscious experience of motion. *Consciousness and cognition*, 28:126–140, 2014.
- [714] John Rawls. A theory of justice. In *Applied ethics*, pages 21–29. Routledge, 2017.
- [715] Andrew G. Reece and Christopher M. Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):1–12, December 2017. Number: 1 Publisher: SpringerOpen.
- [716] Jeffrey H Reiman. Privacy, intimacy, and personhood. In *Privacy*, pages 23–41. Routledge, 2017.
- [717] Federal Reserve. Federal trade commission act section 5: Unfair or deceptive acts or practices. *Consumer Compliance Handbook, June*, 2008.
- [718] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, et al. Open problems in technical ai governance. *arXiv preprint arXiv:2407.14981*, 2024.
- [719] Lauren Rhue. Racial influence on automated perceptions of emotions. *Available at SSRN 3281765*, 2018.
- [720] Lauren Rhue. Affectively Mistaken? How Human Augmentation and Information Transparency Offset Algorithmic Failures in Emotion Recognition AI, November 2019.
- [721] Neil Richards. *Intellectual privacy: Rethinking civil liberties in the digital age*. Oxford University Press, USA, 2015.
- [722] Neil Richards. *Why Privacy Matters*. Oxford University Press, 2022.
- [723] Neil M Richards and Daniel J Solove. Prosser’s privacy law: A mixed legacy. *Calif. L. Rev.*, 98:1887, 2010.
- [724] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.
- [725] Lauren A Rivera. Go with your gut: Emotion and evaluation in job interviews. *American journal of sociology*, 120(5):1339–1389, 2015.
- [726] Georgios Rizos, Alice Baird, Max Elliott, and Björn Schuller. Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3502–3506. IEEE, 2020.

- [727] Ivan Robertson and Cary Cooper. *Well-being: Productivity and happiness at work*. Springer, 2011.
- [728] Ingrid Robeyns. Wellbeing, place and technology. *Wellbeing, Space and Society*, 1:100013, 2020.
- [729] Clara E Rodriguez, Michael H Miyawaki, and Grigoris Argeros. Latino racial reporting in the us: To be or not to be. *Sociology Compass*, 7(5):390–403, 2013.
- [730] Marc A. Rodwin. Patient Accountability and Quality of Care: Lessons From Medical Consumerism and the Patients’ Rights, Women’s Health and Disability Rights Movements. *American Journal of Law & Medicine*, 20(1-2):147–167, 1994. Publisher: Cambridge University Press.
- [731] Kat Roemmich and Nazanin Andalibi. Data subjects’ conceptualizations of and attitudes toward automatic emotion recognition-enabled wellbeing interventions on social media. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–34, 2021.
- [732] Kat Roemmich and Nazanin Andalibi. Emotion inferences in the workplace and healthcare: Workers’ and patients’ emotional privacy judgments and the relative influence of contextual, socio-demographic, and individual privacy belief factors. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2024. under review.
- [733] Kat Roemmich, Shanley Corvite, Cassidy Pyle, Nadia Karizat, and Nazanin Andalibi. Emotion ai use in us mental healthcare: Potentially unjust and techno-solutionist. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–46, 2024.
- [734] Kat Roemmich, Tillie Rosenberg, Serena Fan, and Nazanin Andalibi. Values in emotion artificial intelligence hiring services: Technosolutions to organizational problems. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–28, 2023.
- [735] Kat Roemmich, Florian Schaub, and Nazanin Andalibi. Emotion AI at Work: Implications for Workplace Surveillance, Emotional Labor, and Emotional Privacy. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, Hamburg Germany, April 2023. ACM.
- [736] Brishen Rogers. The law and political economy of workplace technological change. *Harv. CR-CLL Rev.*, 55:531, 2020.
- [737] Andrew J Rohm and George R Milne. Just what the doctor ordered: The role of information sensitivity and trust in reducing medical information privacy concern. *Journal of Business Research*, 57(9):1000–1011, 2004.
- [738] John Rooksby, Alistair Morrison, and Dave Murray-Rust. Student Perspectives on Digital Phenotyping: The Acceptability of Using Smartphone Data to Assess Mental Health. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Glasgow Scotland Uk, May 2019. ACM.

- [739] David Rosen and Aaron Santesso. Inviolate personality and the literary roots of the right to privacy. *Law & Literature*, 23(1):1–25, 2011.
- [740] David Rosen and Aaron Santesso. *The watchman in pieces: Surveillance, literature, and liberal personhood*. Yale University Press, 2013.
- [741] Beate Rössler. *Privacies: philosophical evaluations*. Stanford University Press, 2004.
- [742] Beate Rössler. *The value of privacy*. John Wiley & Sons, 2018.
- [743] Nicolas Rüsch, Matthias C Angermeyer, and Patrick W Corrigan. Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma. *European psychiatry*, 20(8):529–539, 2005.
- [744] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.
- [745] John Rust and Susan Golombok. *Modern psychometrics: The science of psychological assessment*. Routledge, 2014.
- [746] Farig Sadeque, Dongfang Xu, and Steven Bethard. Measuring the Latency of Depression Detection in Social Media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM ’18, pages 495–503, New York, NY, USA, February 2018. Association for Computing Machinery.
- [747] Hiromasa Sakurai and Yusuke Miyao. Evaluating intention detection capability of large language models in persuasive dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1635–1657, 2024.
- [748] Nahal Salimi, Bryan Gere, William Talley, and Bridget Irioogbe. College Students Mental Health Challenges: Concerns and Considerations in the COVID-19 Pandemic. *Journal of College Student Psychotherapy*, 0(0):1–13, February 2021. Publisher: Routledge _eprint: <https://doi.org/10.1080/87568225.2021.1890298>.
- [749] Nandita Sampath. CR’s Comments to the Office of Science and Technology Policy on AI-enabled Biometric Processing, January 2022.
- [750] Samiha Samrose, Wenyi Chu, Carolina He, Yuebai Gao, Syeda Sarah Shahrin, Zhen Bai, and Mohammed Ehsan Hoque. Visual cues for disrespectful conversation analysis. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 580–586. IEEE, 2019.
- [751] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

- [752] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 458–468, New York, NY, USA, 2020. Association for Computing Machinery.
- [753] Michael J Sandel. *What money can't buy: the moral limits of markets*. Brasenose College, Oxford, 1998.
- [754] Michael J Sandel. Market reasoning as moral reasoning: why economists should re-engage with political philosophy. *Journal of Economic Perspectives*, 27(4):121–140, 2013.
- [755] Franklin E Satterthwaite. Synthesis of variance. *Psychometrika*, 6(5):309–316, 1941.
- [756] Andrea Scarantino. The philosophy of emotions and its impact on affective science. *Handbook of emotions*, 4:3–48, 2016.
- [757] Oscar Schachter. Human dignity as a normative concept. *American Journal of International Law*, 77(4):848–854, 1983.
- [758] Maya Schenwar, Macaré Joe, and Alana Yu-lan Price. *Who do you serve, who do you protect?: police violence and resistance in the United States*. Haymarket Books, Chicago, Illinois, 2016. OCLC: 975049067.
- [759] Klaus R Scherer, Angela Schorr, and Tom Johnstone. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.
- [760] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *Proceedings of the ACM on Human-computer Interaction*, 2(CSCW):1–27, 2018.
- [761] Ulrich Schimmack and Alexander Grob. Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality*, 14(4):325–345, 2000.
- [762] Karen L Schmidt, Zara Ambadar, Jeffrey F Cohn, and L Ian Reed. Movement differences between deliberate and spontaneous facial expressions: *Zygomaticus major* action in smiling. *Journal of nonverbal behavior*, 30(1):37–52, 2006.
- [763] Marc Schmitt and Ivan Flechais. Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, 57(12):1–23, 2024.
- [764] Björn Schuller, Adria Mallol-Ragolta, Alejandro Peña Almansa, Iosif Tsangko, Mostafa M Amin, Anastasia Semertzidou, Lukas Christ, and Shahin Amiriparian. Affective computing has changed: The foundation model disruption. *arXiv preprint arXiv:2409.08907*, 2024.
- [765] Dagmar Schuller and Björn W Schuller. The age of artificial emotional intelligence. *Computer*, 51(9):38–46, 2018.

- [766] Jennifer C. H. Sebring. Towards a sociological understanding of medical gaslighting in western health care. *Sociology of Health & Illness*, 43(9):1951–1964, 2021. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9566.13367>.
- [767] Jacob Segal. A delight in doing’: Individuality and action in the political thought of hannah arendt. *New England Journal of Political Science*, 2(1):6, 2007.
- [768] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* ’19*, pages 59–68, Atlanta, GA, USA, 2019. ACM Press.
- [769] Evan Selinger and Woodrow Hartzog. Facebook’s emotional contagion study and the ethical problem of co-opted identity in mediated environments where users lack control. *Research Ethics*, 12(1):35–43, January 2016. Publisher: SAGE Publications Ltd.
- [770] Amartya Sen. Capability and well-being73. *The quality of life*, 30:270–293, 1993.
- [771] Amartya Sen. Human rights and capabilities. *Journal of human development*, 6(2):151–166, 2005.
- [772] Amartya Sen. Development as freedom (1999). *The globalization and development reader: Perspectives on development and global change*, 525, 2014.
- [773] Amartya Sen et al. *Equality of what?*, volume 1. na, 1979.
- [774] Chirag Shah, Ryen White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Nagu Rangan, et al. Using large language models to generate, validate, and apply user intent taxonomies. *ACM Transactions on the Web*, 2023.
- [775] Rusheb Shah, Soroush Pour, Arush Tagade, Stephen Casper, Javier Rando, et al. Scalable and transferable black-box jailbreaks for language models via persona modulation. *arXiv preprint arXiv:2311.03348*, 2023.
- [776] Manoj Kumar Sharma, Nisha John, and Maya Sahu. Influence of social media on mental health: a systematic review. *Current Opinion in Psychiatry*, Publish Ahead of Print, July 2020.
- [777] Kim Bartel Sheehan. Toward a typology of internet users and online privacy concerns. *The information society*, 18(1):21–32, 2002.
- [778] Nelson Shen, Lydia Sequeira, Michelle Pannor Silver, Abigail Carter-Langford, John Strauss, and David Wiljer. Patient privacy perspectives on health information exchange in a mental health context: qualitative study. *JMIR mental health*, 6(11):e13306, 2019.
- [779] Katie Shilton. Values and ethics in human-computer interaction. *Foundations and Trends® in Human–Computer Interaction*, 12(2), 2018.

- [780] Katie Shilton, Jes A Koepfler, and Kenneth R Fleischmann. Charting sociotechnical dimensions of values for design research. *The Information Society*, 29(5):259–271, 2013.
- [781] Katie Shilton, Jes A Koepfler, and Kenneth R Fleischmann. How to see values in social computing: methods for studying values dimensions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 426–435, 2014.
- [782] Kyle Shobe. Productivity driven by job satisfaction, physical work environment, management support and job autonomy. *Business and Economics Journal*, 9(2):1–9, 2018.
- [783] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.
- [784] Yan Shvartzshnaider, Vasisht Duddu, and John Lacalamita. Llm-ci: Assessing contextual integrity norms in language models. *arXiv preprint arXiv:2409.03735*, 2024.
- [785] Herbert A Simon. Motivational and emotional controls of cognition. *Psychological review*, 74(1):29, 1967.
- [786] Natasha Singer. In Screening for Suicide Risk, Facebook Takes On Tricky Public Health Role. *The New York Times*, December 2018.
- [787] Vivek K. Singh and Rishav R. Agarwal. Cooperative phoneotypes: exploring phone-based behavioral markers of cooperation. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp ’16*, pages 646–657, New York, NY, USA, September 2016. Association for Computing Machinery.
- [788] Scott Skinner-Thompson. *Privacy at the margins*. Cambridge University Press, Cambridge, United Kingdom ; New York, NY, 2021.
- [789] Mona Sloane, Emanuel Moss, and Rumman Chowdhury. A silicon valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns*, 3(2):100425, 2022.
- [790] P Slovic. The perception of risk. *earthscan*. London, UK, 2000.
- [791] H Jeff Smith, Sandra J Milberg, and Sandra J Burke. Information privacy: Measuring individuals’ concerns about organizational practices. *MIS quarterly*, pages 167–196, 1996.
- [792] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.
- [793] Daniel J. Solove. A Taxonomy of Privacy. *University of Pennsylvania Law Review*, 154(3):477, January 2006.
- [794] Daniel J Solove. *Nothing to hide: The false tradeoff between privacy and security*. Yale University Press, 2011.

- [795] Daniel J Solove. The myth of the privacy paradox. *Geo. Wash. L. Rev.*, 89:1, 2021.
- [796] Yongmin Song, Phuong Huy Tung, and Beonghwa Jeon. Trends in artificial emotional intelligence technology and application. In *2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD)*, pages 366–370. IEEE, 2022.
- [797] Jared Spataro and Corporate Vice President for Microsoft 365. Microsoft Viva: Empowering every employee for the new digital age, February 2021.
- [798] 89th Cong. 2nd Session Special Subcomm. on Invasion of Privacy, House Comm. on Government Operations. The computer and invasion of privacy, 1966.
- [799] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. How to do better with gender on surveys: a guide for hci researchers. *Interactions*, 26(4):62–65, 2019.
- [800] Barbara Stanley, Gonzalo Martínez-Alés, Ilana Gratch, Mina Rizk, Hanga Galfalvy, Tse-Hwei Choo, and J John Mann. Coping strategies that reduce suicidal ideation: An ecological momentary assessment study. *Journal of psychiatric research*, 133:32–37, 2021.
- [801] Luke Stark. The emotional context of information privacy. *The Information Society*, 32(1):14–27, 2016.
- [802] Luke Stark and Jesse Hoey. The Ethics of Emotion in Artificial Intelligence Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 782–793, Virtual Event Canada, March 2021. ACM.
- [803] Luke Stark and Jevan Hutson. Physiognomic artificial intelligence. *Available at SSRN 3927300*, 2021.
- [804] Luke Stark, Amanda Stanhaus, and Denise L. Anthony. “I Don’t Want Someone to Watch Me While I’m Working”: Gendered Views of Facial Recognition Technology in Workplace Surveillance. *Journal of the Association for Information Science and Technology*, 71(9):1074–1088, 2020. eprint: <https://onlinelibrary.wiley.com/doi/10.1002/asi.24342>.
- [805] Dirk D Steiner. Personnel selection across the globe. 2012.
- [806] Cynthia Kay Stevens and Amy L. Kristof. Making the right impression: A field study of applicant impression management during job interviews. *Journal of Applied Psychology*, 80:587–606, 1995.
- [807] Cynthia Kay Stevens and Myeong-Gu Seo. Job search and emotions. *The Oxford handbook of recruitment*, pages 126–138, 2014.
- [808] Lauren Stewart. Big data discrimination: Maintaining protection of individual privacy without disincentivizing businesses’ use of biometric data to enhance security. *BCL Rev.*, 60:349, 2019.
- [809] Catherine Stinson. The dark past of algorithms that associate appearance and criminality: Machine learning that links personality and physical traits warrants critical review. *American Scientist*, 109(1):26–30, 2021.

- [810] Christine Storm and Tom Storm. A taxonomic study of the vocabulary of emotions. *Journal of personality and social psychology*, 53(4):805, 1987.
- [811] Anselm Strauss. A social, vworld perspective. *Studies in symbolic interaction*, 1:119–128, 1978.
- [812] Harald Strömfelt, Yue Zhang, and Björn W Schuller. Emotion-augmented machine learning: overview of an emerging domain. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 305–312. IEEE, 2017.
- [813] R Subhashini and PR Niveditha. Analyzing and detecting employee’s emotion for amelioration of organizations. *Procedia Computer Science*, 48:530–536, 2015.
- [814] V Suchitra Mouly and Jayaram K Sankaran. On the study of settings marked by severe superior-subordinate conflict. *Organization Studies*, 18(2):175–192, 1997.
- [815] Derek H Suite, Robert La Bril, Annelle Primm, and Phyllis Harrison-Ross. Beyond Misdiagnosis, Misunderstanding and Mistrust: Relevance of the Historical Perspective in the Medical and Mental Health Treatment of People of Color. *JOURNAL OF THE NATIONAL MEDICAL ASSOCIATION*, 99(8), 2007.
- [816] Louise Sundararajan. Happiness donut: A confucian critique of positive psychology. *Journal of Theoretical and Philosophical Psychology*, 25(1):35, 2005.
- [817] Veikko Surakka and Jari K Hietanen. Facial and emotional reactions to duchenne and non-duchenne smiles. *International journal of psychophysiology*, 29(1):23–33, 1998.
- [818] Daniel Susser, Beate Roessler, and Helen Nissenbaum. Online manipulation: Hidden influences in a digital world. *Geo. L. Tech. Rev.*, 4:1, 2019.
- [819] Nandhini Swaminathan and David Danks. Application of the nist ai risk management framework to surveillance technology. *arXiv preprint arXiv:2403.15646*, 2024.
- [820] George Szmukler. Compulsion and “coercion” in mental health care. *World Psychiatry*, 14(3):259–261, October 2015.
- [821] Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0). NIST Special Publication NIST AI 100-1, National Institute of Standards and Technology, January 2023. NIST Pubs.
- [822] Ziying Tan, Xingyun Liu, Xiaoqian Liu, Qijin Cheng, and Tingshao Zhu. Designing Microblog Direct Messages to Engage Social Media Users With Suicide Ideation: Interview and Survey Study on Weibo. *Journal of Medical Internet Research*, 19(12):e381, December 2017.
- [823] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How well do my results generalize now? the external validity of online privacy and security surveys. In *Eighteenth symposium on usable privacy and security (SOUPS 2022)*, pages 367–385, 2022.

- [824] Janice Hopkins Tanne. Florida bans abortions after six weeks, leaving millions of women in southeastern us without care, 2024.
- [825] Andrew E Taslitz. The fourth amendment in the twenty-first century: Technology, privacy, and human emotions. *Law & Contemp. Probs.*, 65:125, 2002.
- [826] Gavin Tay and Bern Elliot. Maverick* Research: Emotion AI Will Become You Without Your Knowledge, November 2019.
- [827] Gavin Tay, Annette Zimmerman, and Bern Elliot. Maverick* Research: Emotional Wellness Will Rescue Your Organization and Distributed Workforce, April 2021.
- [828] Doris Teutsch, Philipp K Masur, and Sabine Trepte. Privacy in mediated and nonmediated interpersonal communication: How subjective concepts and situational perceptions influence behaviors. *Social Media+ Society*, 4(2):2056305118767134, 2018.
- [829] The American Journal of Managed Care. Vulnerable Populations: Who Are They? *The American Journal of Managed Care (AJMC)*, 12(13):348–352, November 2006. Publisher: MJH Life Sciences.
- [830] Kim Theodos and Scott Sittig. Health Information Privacy Laws in the Digital Age: HIPAA Doesn't Apply. *Perspectives in Health Information Management*, 18(Winter):11, 2021.
- [831] Hubert Thomas. Opinion 4/2015-towards a new digital ethics data, dignity and technology. 2015.
- [832] Marilyn D. Thomas, Nicholas P. Jewell, and Amani M. Allen. Black and unarmed: statistical interaction between age, perceived mental illness, and geographic region among males fatally shot by police using case-only design. *Annals of Epidemiology*, 53:42–49.e3, January 2021.
- [833] Laura K. Thompson, Margaret M. Sugg, and Jennifer R. Runkle. Adolescents in crisis: A geographic exploration of help-seeking behavior using data from Crisis Text Line. *Social Science & Medicine*, 215:69–79, October 2018.
- [834] Judith Jarvis Thomson. The right to privacy. *Philosophy & Public Affairs*, pages 295–314, 1975.
- [835] Michael Thomson. A capabilities approach to best interests assessments. *Legal Studies*, 41(2):276–293, 2021.
- [836] Robert Thornberg, Kathy Charmaz, et al. Grounded theory and theoretical coding. *The SAGE handbook of qualitative data analysis*, 5:153–69, 2014.
- [837] Somchanok Tivatansakul, Michiko Ohkura, Supadchaya Puangpontip, and Tiranee Achalakul. Emotional healthcare system: Emotion detection by facial expressions using japanese database. In *2014 6th computer science and electronic engineering conference (CEEC)*, pages 41–46. IEEE, 2014.

- [838] Jan Tolsdorf and Florian Dehling. In our employer we trust: mental models of office workers' privacy perceptions. In *International Conference on Financial Cryptography and Data Security*, pages 122–136. Springer, 2020.
- [839] John Torous, Mathew V Kiang, Jeanette Lorme, and Jukka-Pekka Onnela. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Mental Health*, 3(2), May 2016.
- [840] Kate E Toth and Carolyn S Dewa. Employee decision-making about disclosure of a mental disorder at work. *Journal of occupational rehabilitation*, 24:732–746, 2014.
- [841] Horst Treiblmaier and Peter Filzmoser. Benefits from using continuous rating scales in online survey research. In *ICIS*, 2011.
- [842] Sabine Trepte. The social media privacy model: Privacy and communication in the light of social media affordances. *Communication Theory*, 31(4):549–570, 2021.
- [843] Khiet P Truong, David A Van Leeuwen, and Franciska MG De Jong. Speech-based recognition of self-reported and observed emotion in a dimensional space. *Speech communication*, 54(9):1049–1063, 2012.
- [844] Thomas Tsiampalis and Demosthenes B Panagiotakos. Missing-data analysis: socio-demographic, clinical and lifestyle determinants of low response rate on self-reported psychological and nutrition related multi-item instruments in the context of the attica epidemiological study. *BMC Medical Research Methodology*, 20(1):1–13, 2020.
- [845] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. Recognizing Depression from Twitter Activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3187–3196, New York, NY, USA, April 2015. Association for Computing Machinery.
- [846] Mary P Tully, Kyle Bozentko, Sarah Clement, Amanda Hunn, Lamiece Hassan, Ruth Norris, Malcolm Oswald, and Niels Peek. Investigating the extent to which patients should control access to patient records for research: a deliberative process using citizens' juries. *Journal of medical Internet research*, 20(3):e7763, 2018.
- [847] Daniel W Turner III. Qualitative interview design: A practical guide for novice investigators. *The qualitative report*, 15(3):754, 2010.
- [848] Joseph Turow, Nora Draper, Mara Einstein, James F Hamilton, and Edward Timke. The voice catchers: How marketers listen in to exploit your feelings, your privacy, and your wallet. *Advertising & Society Quarterly*, 22(4), 2021.
- [849] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8):1301–1309, 2017.
- [850] Ozlem Ulgen. Ai and the crisis of the self: Protecting human dignity as status and respectful treatment. In *The Frontlines of Artificial Intelligence Ethics*, pages 9–33. Routledge, 2022.

- [851] Gaby Umbach. Futures in eu governance: Anticipatory governance, strategic foresight and eu better regulation. *European law journal*, 30(3):409–421, 2024.
- [852] European Union. Aims and values, 2024. Accessed 06-17-2025.
- [853] U.S. Department of Health and Human Services. Code of federal regulations, title 45: Public welfare, part 46: Protection of human subjects, section 46.104: Exempt research. <https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46#46.104>, 2018. Revised Common Rule.
- [854] Eric M Uslaner. Trust, civic engagement, and the internet. *Political Communication*, 21(2):223–242, 2004.
- [855] Ramyadarshni Vadivel, Sheikh Shoib, Sarah El Halabi, Samer El Hayek, Lamiaà Essam, Drita Gashi Bytyçi, Ruta Karaliuniene, Andre Luiz Schuh Teixeira, Sachin Nagendrappa, Rodrigo Ramalho, Ramdas Ransing, Victor Pereira-Sanchez, Chonnakarn Jatchavala, Frances Nkechi Adiukwu, and Ganesh Kudva Kundadak. Mental health in the post-COVID-19 era: challenges and the way forward. *General Psychiatry*, 34(1):e100424, February 2021.
- [856] André Calero Valdez and Martina Ziefle. The users' perspective on the privacy-utility trade-offs in health recommender systems. *International Journal of Human-Computer Studies*, 121:108–121, 2019.
- [857] Michel F Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170, 2006.
- [858] Madison Van Oort. The emotional labor of surveillance: Digital control in fast fashion retail. *Critical Sociology*, 45(7-8):1167–1179, 2019.
- [859] Jukka Varelius. The value of autonomy in medical ethics. *Medicine, Health Care and Philosophy*, 9:377–388, 2006.
- [860] William N Venables and Brian D Ripley. Random and mixed effects. In *Modern applied statistics with S*, pages 271–300. Springer, 2002.
- [861] Peter-Paul Verbeek and Pieter E Vermaas. Technological artifacts. 2012.
- [862] Anushree Verma. Hype Cycle for Sensing Technologies and Applications, 2020, July 2020.
- [863] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.
- [864] Catalin Voss, Jessey Schwartz, Jena Daniels, Aaron Kline, Nick Haber, Peter Washington, Qandeel Tariq, Thomas N Robinson, Manisha Desai, Jennifer M Phillips, et al. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. *JAMA pediatrics*, 173(5):446–454, 2019.

- [865] Rosalie Waelen and Michał Wieczorek. The Struggle for AI's Recognition: Understanding the Normative Implications of Gender Bias in AI with Honneth's Theory of Recognition. *Philosophy & Technology*, 35(2):53, June 2022.
- [866] Robin Wakefield. The influence of user affect in online information disclosure. *The Journal of Strategic Information Systems*, 22(2):157–174, 2013.
- [867] Ari Ezra Waldman. *Privacy as Trust: Information Privacy for an Information Age*. Cambridge University Press, 2018.
- [868] Ari Ezra Waldman. *Industry unbound: The inside story of privacy, data, and corporate power*. Cambridge University Press, 2021.
- [869] Lauren Walker. Belgian man dies by suicide following exchanges with chatbot.
- [870] Anthony FC Wallace and Margaret T Carson. Sharing and diversity in emotion terminology. *Ethos*, pages 1–29, 1973.
- [871] Michael Walzer. *Spheres of justice: A defense of pluralism and equality*. Basic books, 1984.
- [872] Michael Walzer. *Thick and thin: Moral argument at home and abroad*. University of Notre Dame Pess, 1994.
- [873] Jun Wan, Sergio Escalera, Gholamreza Anbarjafari, Hugo Jair Escalante, Xavier Baró, Isabelle Guyon, Meysam Madadi, Juri Allik, Jelena Gorbova, Chi Lin, et al. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3189–3197, 2017.
- [874] Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*, 2024.
- [875] Ying Zhu, Yong Sun, Yu Wang. Nokia mobile data challenge: Predicting semantic place and next place via mobile data. *Work*, 80(100):120, 2012.
- [876] Samuel D Warren and D Louis. Brandeis, the right to privacy, 4 harv. L. rev, 193(10.2307):1321160, 1890.
- [877] Alan S Waterman. The relevance of aristotle's conception of eudaimonia for the psychological study of happiness. *Theoretical & Philosophical Psychology*, 10(1):39, 1990.
- [878] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.
- [879] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*, 2023.

- [880] Suzanne P Weisband and Bruce A Reinig. Managing user perceptions of email privacy. *Communications of the ACM*, 38(12):40–47, 1995.
- [881] Glenn E Weisfeld and Stefan MM Goetz. Applying evolutionary thinking to the study of emotion. *Behavioral Sciences*, 3(3):388–407, 2013.
- [882] Robert S Weiss. *Learning from strangers: The art and method of qualitative interview studies*. Simon and Schuster, 1995.
- [883] Alan F Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1):166, 1968.
- [884] Alan F Westin and Danielle Maurici. *E-commerce & privacy: What net users want*. Privacy & American Business Hackensack, NJ, 1998.
- [885] Meredith Whittaker, Meryl Alper, Cynthia L Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, et al. Disability, bias, and ai. *AI Now Institute*, page 8, 2019.
- [886] Meredith Whittaker, Meryl Alper, Olin College, Liz Kaziunas, and Meredith Ringel Morris. Disability, Bias, and AI. Technical report, AINow, 2019.
- [887] Kyle Wiggers. Microsoft launches Viva, an AI-powered information hub for enterprises, February 2021.
- [888] S Elizabeth Wilborn. Revisiting the public/private distinction: Employee monitoring in the workplace. *Ga. L. Rev.*, 32:825, 1997.
- [889] Alex C Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T Iqbal, and Jaime Teevan. Supporting workplace detachment and reattachment with conversational intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.
- [890] John Williams. *Wordsworth: Romantic poetry and revolution politics*. Manchester University Press, 1989.
- [891] Langdon Winner. Do artifacts have politics? *Daedalus*, pages 121–136, 1980.
- [892] Donghyeon Won, Zachary C. Steinert-Threlkeld, and Jungseock Joo. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM ’17, page 786–794, New York, NY, USA, 2017. Association for Computing Machinery.
- [893] Pak-Hang Wong. Democratizing algorithmic fairness. *Philosophy & Technology*, 33(2):225–244, 2020.
- [894] Richmond Y. Wong, Deirdre K. Mulligan, and John Chuang. Using science fiction texts to surface user reflections on privacy. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp ’17, pages 213–216, New York, NY, USA, September 2017. Association for Computing Machinery.

- [895] Richmond Y. Wong, Ellen Van Wyk, and James Pierce. Real-Fictional Entanglements: Using Science Fiction and Design Fiction to Interrogate Sensing Technologies. June 2017.
- [896] Stephen Wood, Valerio Ghezzi, Claudio Barbaranelli, Cristina Di Tecco, Roberta Fida, Maria Luisa Farnese, Matteo Ronchetti, and Sergio Iavicoli. Assessing the risk of stress in organizations: getting the measure of organizational-level stressors. *Frontiers in Psychology*, 10:2776, 2019.
- [897] David Wright. Making privacy impact assessment more effective. *The Information Society*, 29(5):307–315, 2013.
- [898] James D Wright, Peter V Marsden, et al. Survey research and social science: History, current practice, and future prospects. *Handbook of survey research*, pages 3–26, 2010.
- [899] WSJ Staff. Inside tiktok’s algorithm: A wsj video investigation. <https://www.wsj.com/tech/tiktok-algorithm-video-investigation-11626877477>, July 2021. The Wall Street Journal. Investigation using automated accounts to reveal how TikTok uses watch time to infer user preferences.
- [900] Li Wu, Rong Huang, Zhe Wang, Jonathan Nimal Selvaraj, Liuqing Wei, Weiping Yang, and Jianxin Chen. Embodied emotion regulation: The influence of implicit emotional compatibility on creative thinking. *Frontiers in Psychology*, 11:1822, 2020.
- [901] Kim Wuyts and Wouter Joosen. Linddun privacy threat modeling: a tutorial. *CW Reports*, 2015.
- [902] Kim Wuyts, Laurens Sion, and Wouter Joosen. Linddun go: A lightweight approach to privacy threat modeling. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 302–309. IEEE, 2020.
- [903] Christopher Wylie. *Mindf* ck: Cambridge Analytica and the plot to break America*. Random House, 2019.
- [904] Heng Xu, Hao Wang, and Hock-Hai Teo. Predicting the usage of p2p sharing software: The role of trust and perceived risk. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 201a–201a. IEEE, 2005.
- [905] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*, pages 506–523. Springer, 2020.
- [906] Tong Xue, Surjya Ghosh, Gangyi Ding, Abdallah El Ali, and Pablo Cesar. Designing real-time, continuous emotion annotation techniques for 360 vr videos. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2020.
- [907] Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun Liu. Beyond scalar reward model: Learning generative judge from preference data. *arXiv preprint arXiv:2410.03742*, 2024.

- [908] Iris Marion Young. House and home: Feminist variations on a theme. In *Motherhood and space: Configurations of the maternal through politics, home, and the body*, pages 115–147. Springer, 2005.
- [909] Paul Thomas Young. Motivation and emotion: A survey of the determinants of human and animal activity. 1961.
- [910] Eman MG Younis, Someya Mohsen, Essam H Houssein, and Osman Ali Sadek Ibrahim. Machine learning for human emotion recognition: a comprehensive review. *Neural Computing and Applications*, 36(16):8901–8947, 2024.
- [911] Robert B Zajonc. Emotions. 1998.
- [912] Michalinos Zembylas. Emotion, resistance, and self-formation. *Educational theory*, 53(1), 2003.
- [913] Michalinos Zembylas. Emotions and teacher identity: A poststructural perspective. *Teachers and Teaching*, 9(3):213–238, 2003.
- [914] Colin A. Zestcott, Irene V. Blair, and Jeff Stone. Examining the presence, consequences, and reduction of implicit bias in health care: A narrative review. *Group Processes & Intergroup Relations*, 19(4):528–542, July 2016.
- [915] Biqiao Zhang and Emily Mower Provost. Automatic recognition of self-reported and perceived emotions. In *Multimodal Behavior Analysis in the Wild*, pages 443–470. Elsevier, 2019.
- [916] Dongsong Zhang, Jaewan Lim, Lina Zhou, and Alicia A Dahl. Breaking the data value-privacy paradox in mobile mental health systems through user-centered privacy protection: A web-based survey study. *JMIR Mental Health*, 8(12):e31633, 2021.
- [917] Huaizheng Zhang, Yong Luo, Qiming Ai, Yonggang Wen, and Han Hu. Look, read and feel: Benchmarking ads understanding with multimodal multitask learning. In *Proceedings of the 28th ACM international conference on multimedia*, pages 430–438, 2020.
- [918] Huangbin Zhang, Chong Zhao, Yu Zhang, Danlei Wang, and Haichao Yang. Deep latent emotion network for multi-task learning. *arXiv preprint arXiv:2104.08716*, 2021.
- [919] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nicheli. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59:103–126, 2020.
- [920] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2025.

- [921] Shikun Zhang, Yuanyuan Feng, Lujo Bauer, Lorrie Faith Cranor, Anupam Das, and Norman Sadeh. “did you know this camera tracks your mood?”: Understanding privacy expectations and preferences in the age of video analytics. *Proceedings on Privacy Enhancing Technologies*, 2021(2):282–304, 2021.
- [922] Shikun Zhang, Yuanyuan Feng, and Norman Sadeh. Facial recognition: Understanding privacy concerns and attitudes across increasingly diverse deployment scenarios. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 243–262. USENIX Association, August 2021.
- [923] Yifan Zhang, Ge Zhang, Yue Wu, Kangping Xu, and Quanquan Gu. Beyond bradley-terry models: A general preference model for language model alignment. In *Forty-second International Conference on Machine Learning*, 2025.
- [924] Jingya Zheng, Gaofeng Tao, Shuxin Qin, Dan Wang, and Zhongjun Ma. Intent-based multi-cloud storage management powered by a fine-tuned large language model. *IEEE Access*, 2025.
- [925] Kathryn Zickuhr. Workplace surveillance is becoming the new normal for us workers. *Washington Center for Equitable Growth*. <https://equitablegrowth.org/research-paper/workplace-surveillance-is-becoming-the-new-normal-for-us-workers/>. Institute for Research on Labor and Employment University of California, Berkeley, 2521:94720–5555, 2021.
- [926] Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, 2017.
- [927] Annuska Zolyomi and Jaime Snyder. Social-Emotional-Sensory Design Map for Affective Computing Informed by Neurodivergent Experiences. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):77:1–77:37, April 2021.
- [928] Shoshana Zuboff. The age of surveillance capitalism: The fight for a human future at the new frontier of power, edn. *PublicAffairs, New York*, 2019.