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

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 to 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 for which I cannot thank you enough.

To Emily Mower Provost: I admire your intellectual enthusiasm and care for your research, and I'm grateful you brought both qualities 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 the thoughtfulness you've given to each interaction while serving on my committee. 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. Funding for empirical work was sponsored by the University of Michigan Graduate Student Research Grant and National Science Foundation award numbers 2236674 and 2020872.

A special shout-out to Shanley Corvite, Nadia Karizat, and Cassidy Pyle—this work wouldn't be the same without you. I thank you not just as collaborators but as friends: your steady support while navigating the complexities of this research made surviving a PhD not just possible, but worthwhile.

To my dear friends and labmates Kaiwen Sun, Abraham Mhaidli, and Yixin Zou: whether it's a technical puzzle, a half-formed idea, personal advice, or just showing up with an open ear, I know I can rely on you to be there, and it means the world to me. Thank you. To Tanisha Afnan, Byron Lowens, Justin Petelka, Lu Xian, Jackie Hu, Allison McDonald, and the rest of the incredible spilab crew, 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—I love that about you—and yet you made sure it stayed lit for me: always a new book and 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 seeing others has provided a steady well of compassion and clarity when my own runs low. To my inquisitive bear, Orson: your big heart and boundless curiosity have kept me inspired and open to the wonder that keeps the intellect humble, honest, and alive. Your vast interests in how things work and why people do what they do provide me constant reminders 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 a return to those same texts years later while I finished this dissertation, as you began reading them in earnest yourself. It was a rare kind of experience that saw both the philosophers—and each other—anew. The unforeseen gift given by the chance to think and grow alongside you has provided me the clearest reminder that the best kind of 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 bear traces of your patience, your presence, and your love. The meaning of this dissertation, for me, begins and ends with you.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

Technologies that automatically detect, interpret, respond to, and interact with human emotion make bold promises about their potential to improve human life by *humanizing* technology. Yet in deployment, these promises confront the reality of lived experience and intersect fragile conditions of social life—power, agency, dignity. Technology's potential to enhance human flourishing is, by the same token, a potential to undermine it. Where the needle turns, this dissertation argues, depends on our answer to a single question:

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

Grounded in empirical studies of privacy perceptions, experiences, judgments, and harms, this dissertation develops the concept of *emotional privacy* and defends the need to protect it as a moral and political imperative in AI-mediated societies. Linking mixed-methods research on emotion AI deployments in social media, the workplace, and healthcare settings to broader philosophical debates on the human values of emotions and privacy, this dissertation shows that safeguarding emotional privacy demands more than regulating emotional information—it must secure our capacities to live a *truly human life* with agency, dignity, and meaning in interdependent social domains.

Part I lays the theoretical and technical foundations, surveying philosophical and scientific accounts of emotion alongside the mechanics of emotion AI. Through qualitative studies with emotion AI data subjects and a content analysis of commercial emotion AI vendors' promotional materials, Part II builds the empirical case for emotional privacy as a distinct interest. My findings illuminate how emotional data flows—when filtered through power-laden social dynamics, institutional designs, and capacities for personhood—carry context-dependent risks capable of reconfiguring what it means to be human.

Part III measures and evaluates these risks through the lens of Helen Nissenbaum's theory of privacy as Contextual Integrity (CI). It reports on a mixed-methods factorial vignette study that quantitatively measures emotional privacy judgments, supported by qualitative analyses of participants' perceived risks and benefits of emotion AI use cases in employer and healthcare

scenarios. My analysis validates CI's normative claim: privacy judgments are grounded in context-relative informational norms and teleological aims. Yet the results also surfaced a deeper, context-independent normative expectation behind participants' emotional privacy judgments: that agency and dignity be upheld, even as their realization proved contingent on specific implementations.

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). CA-CI retains CI's analytic strengths while anchoring its justificatory logic in a shared normative threshold: human dignity. Deepening CI's normative components, it formalizes *purpose* as a sixth constitutive parameter and treats the set of core capabilities Nussbaum specifies as 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—as fixed transmission principles. Under CA-CI, a data flow is presumptively appropriate only if its foreseeable impact preserves these dignity thresholds for every person; if it predictably erodes even one below bar, it fails the dignity test—unjustified unless and until it is redesigned with the essential constraints for maintaining dignity in place. Through three case studies, I demonstrate CA-CI's practical utility, showing how my normative governance framework surfaces dignity-eroding flows permitted or overlooked by purely procedural, classification-based governance approaches and guides their redesign to uphold the integrity of both the person and the context. Drawing a universal *moral minimum* line in the sand at respect for human dignity, CA-CI offers policymakers, regulators, technologists, and other system evaluators a principled roadmap for building and governing socio-technical systems that uphold and enhance the human capacities upon which a just and flourishing society depend.

CHAPTER 1

Introduction

Human emotions are fundamental components of the inner life: intelligent, embodied judgments that reveal what we value, believe, and strive to protect—often before we can fully articulate it ourselves [1]. The lens through which we see the world, emotions are the felt edge of practical reason and moral orientation: how we come to know what matters. The personal salience and epistemic richness that render emotions so valuable to human life also make them dangerous targets for extractive computation. In the age of emotion AI, the inner life becomes legible to intelligent machines, to the institutions that deploy them, and to the markets that trade in their outputs.

AI systems that detect, interpret, and respond to emotions and related dispositional cues—intention, belief, affective state—are rapidly becoming embedded in critical sectors such as social media, the workplace, and healthcare. Across domains, emotion AI promises significant social benefits: enabling more timely health interventions, supporting individuals in crisis, predicting relapse in chronic illness, and helping organizations better understand emotional needs and mental health demands [2, 3, 4, 5, 6, 7, 8, 9, 10]. Yet technology is neither a neutral actor nor a substitute for human intent—it is an amplifier of existing institutional aims and social structures [11]. The same systems that promise enhanced care, support, and safety can also entrench social stratification, manipulate vulnerability, and compromise autonomy. Technology deployments do not operate in a vacuum: their effects percolate through institutions with competing incentives and are shaped by design choices which encode assumptions about what is knowable, valuable, and permissible. A technology's potential may at once enhance flourishing or foreclose it, realize capacities for agency and dignity or erode them.

With such clear potential for both promise and peril, we need a principled framework for reasoning about the appropriate role of AI technologies in human life. Under what conditions are algorithmic interactions with human emotion justifiable? By what standards should we evaluate the promises these technologies offer and the risks they pose? Where should we draw the line between appropriate use and unacceptable harm?

Part empirical inquiry, part critique, part theory development, this dissertation argues that the justificatory boundary for technology lies in its impact on the core conditions that make human dignity possible. Using emotion AI as a diagnostic lens, I investigate the normative stakes that arise when emotions are inferred and acted upon in social media, healthcare, and workplace domains—sites of identity formation, relational interdependence, and material survival, with profound influence over emotional life. When emotion AI is deployed in these settings, their power to shape what people can *do* and *become* amplifies by reconfiguring the conditions through which individuals experience emotion, form beliefs, affiliate with others, and pursue personal visions of the good.

Part I surveys leading theories of emotion and how they have been made tractable for technical systems, followed by results from four empirical studies that examine the ethical and privacy implications of emotion AI technologies in social media, workplace, and healthcare contexts presented in Parts II and III. Part II draws on qualitative methods to center the perceptions and experiences of, and impacts to, emotion AI's data subjects. With findings that underscore the roles of *context* and *purpose* in evaluating emotion AI's stakes, I name and frame the normative insights surfaced as problems of *emotional privacy*. Part III confirms these insights through a cross-comparative factorial vignette survey study, applying mixed methods to empirically assess how people judge emotional privacy when emotion AI is integrated in labor and health contexts for specific purposes—locating the appropriateness of emotional data flows in the intuitive moral judgments of those directly impacted by the technology. Through the lens of Nissenbaum's Contextual Integrity (CI) [12], I measure how emotional privacy judgments vary within and across contexts and sampling frames, qualified by participants' benefit and risk perceptions. My findings emphasize that when AI systems infer, interact with, and shape the emotional dimensions of our lives, machines, institutions, and data markets gain access not merely to another category of sensitive information—they gain access to the constitutive components of dignity and moral personhood [1]. The results validate the normative basis for CI's theory of privacy [13]: workers and patients alike tend to judge emotion AI uses more favorably when aligned with a context's *telos*, or core purpose, and less favorably of uses that strain or disrupt it. Yet, the data also revealed a more fundamental 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.

To operationalize this insight in CI, Part IV moves from empirical findings to theory development. It traces the legal and philosophical trajectory of privacy as a dignitary interest and argues that the moral source of privacy lies not only in the integrity of a context, but in the integrity of the self—rooted in our capacities for agency and maintaining the moral boundaries necessary to live with dignity. These conditions are not derivative of context, but rather constitute what makes context morally and politically significant in the first place [14]. They are the conditions that make valued social domains both possible and legitimate: we cherish domains not simply because they function, but because their functioning supports human life in developing, sustaining, and inhabiting the capacities to think, feel, act, and relate—to become as whole persons across contexts [15].

Chapter 7 develops this insight to extend CI with a second justificatory line: recognizing flows as appropriate only when they uphold both respect for context *and* human dignity. This chapter introduces the Capabilities-Contextual Integrity (CA-CI) framework, which integrates Martha Nussbaum's Capabilities Approach (CA) [15] within CI by establishing concrete dignity thresholds as fixed transmission principles. Through three case studies, I demonstrate how CA-CI provides a more robust and morally coherent standard for normatively evaluating data flows, particularly where current governance procedures fall short. Chapter 8 concludes by reflecting on the future of emotional privacy amidst accelerating AI development, examining how emerging systems strain current regulatory paradigms and outlining key directions for future governance and research.

This dissertation is about privacy, but more fundamentally, it is about dignity. To insist on the moral and political significance of emotional privacy is not merely to protect personal data or mitigate technology-enabled harm. It is to safeguard the conditions under which people can *do* and *be* with personal integrity, to live with dignity in the face of mounting external pressures that coerce and distort our inner lives. As AI systems increasingly engage with the contours of human subjectivity to re-engineer how we live, relate, and aspire, the stakes of emotional privacy have never been higher. But we are not passive recipients of socio-technical fate. We are human agents capable of building and governing systems to serve human ends—not only in terms of aggregate benefit, but in the real and substantive opportunities available to each person to flourish. That future is not guaranteed, but it is within reach. This dissertation offers a framework for how we might begin.

Part I: Conceptual Foundations

Chapter 2 establishes why any serious governance of emotion AI must begin with the normative significance of both emotions and privacy for sustaining human dignity and agency. Drawing on philosophy and the affective sciences, it frames emotions as *evaluative judgments*—intelligent appraisals that organize our perceptions, guide action, and underpin human flourishing. It positions privacy, and especially emotional privacy, as a precondition for exercising this evaluative agency, showing how machine inference can distort or appropriate that function. The chapter also traces how competing theories of emotion are translated into technical assumptions inside emotion AI systems, revealing their normative load. Together, these foundations provide the intellectual runway for Parts II–IV, where my empirical studies and CA–CI framework take up the challenge of protecting emotional life in AI-mediated societies.

CHAPTER 2

When Human Emotions Meet Machines

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 [16, 17, 18, 19]. Today the debate spans virtually every field that studies human (and non-human) life: law, biology, sociology, linguistics, economics, psychology, neuroscience, anthropology, and computing, to name a few.

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. Philosopher Martha Nussbaum's conceptual analysis crystallizes this view: updating the Stoicist account of emotions with decades of empirical literature, she defines emotions as "intelligent responses to the perception of value"—affect-laden cognitions that disclose what we care about and why [1]. On this eudaimonistic account, emotions 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 [20], bias reasoning and memory formation [21], and energize goal pursuit [22]. In social domains, emotional expressions signal intentions, coordinate cooperation, and forge durable bonds [23]. Developmentally, emotions scaffold identity and moral learning [24]. At the societal level, they underwrite shared norms and institutional arrangements [25]. 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 [15]. 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 AI governance.

2.1.1 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. More than influencing the performance of emotion, social norms 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 casts *eudaimonia*, or flourishing, not as hedonic pleasure but as the fulfillment of one's inner potential, or *daimon* [26, 27]. The eudaimonistic account of happiness is not a feeling to be pursued directly, but a higher-order satisfaction that emerges from living in accordance with virtue. Modern positive psychology and Western social norms inherit its understanding that emotional maturity and self-cultivation are integral to moral excellence—enjoining individuals to "live their best life" and pursue happiness as both a personal and civic duty [28, 29, 30].

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 [31]. 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.

Indeed, the Western fixation on happiness as the *telos*, or ultimate aim, 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 [32, 33]. 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 [32].

Culturally 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. 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. Systems designed to nudge users toward happiness, calm, or engagement, for instance, 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 also in what they normalize.

Despite their points of divergence, a shared understanding lies behind cultural emotional differences: 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, when systems analyze behavioral signals to sense, infer, or interact with emotions, they operate at the heart of what it means to live well, and thus must be held to standards that protect that moral terrain.

2.1.2 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 [25, 34, 35]. Undergirding the development of relationships, social structures, and selfhood, emotions enable us to navigate moral meanings, construct identities, and form beliefs [24, 36]. Emotions are also motivational engines for intellectual curiosity and creative thought [37], with affective states shaping how individuals engage with ideas, learn from others, and act in the world [38].

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—trust, loyalty, shame, resentment—which legitimize, sustain, or disrupt them [25]. Emotions reflect and reinforce shared ideals, as Section 2.1.1 underscored—a function that enables individuals to pursue personal and communal visions of flourishing. Whether in family life, education, work, or civic participation, emotional experience and expression are central to the realization of meaningful, agentic lives.

2.1.3 Privacy

Less often acknowledged in the literature, emotional life depends upon the preservation of privacy. The salience of emotions in social connection necessitates 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 [39], 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 [40]. 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 [40]. The turbulence caused by breached emotional boundaries is often experienced as a privacy violation, adversely affecting relationships and producing psychological harms such as stress, shame, or loss of control [41].

Indeed, such harms can intensify when emotional information is exposed, inferred, or acted upon without consent. Luke Stark notes 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 [42]. Likewise, as Parts II and III emphasize, the particular sensitivity people ascribe to emotion data is not only because of what the information 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 [43, 44], and less likely when they experience anxiety or discomfort [45, 46]. While emotional interactions can foster support and intimacy, they also expose individuals to risk—especially when commodified for commercial gain and accessed by third parties, beyond their intended audience. Platforms may exploit these dynamics: known to intentionally foster trust to increase engagement and data richness [47, 48], emotional experiences transform into extractive assets.

Emotional privacy encompasses not only what we share, but what we reserve. To manage emotional expression—the ability to feel one thing and express another—is itself a privacy act [49]: a parent may reassure their child in a moment of crisis while privately feeling fear; a worker may express satisfaction while masking frustration. Such privacy practices rely on a basic assumption that what we share about how we feel is ours to control, and not automatically transparent to others. As Chapter 5 shows, emotion AI deployments can undermine 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 [50]. For instance, employees who can shape how their emotional state is perceived may enjoy greater job security or avoid punitive consequences [51, 52]. Privacy also enables emotional rest and reflection, providing a retreat from the performative demands of public life [53, 54]. When such functions are lost, individuals may face greater precarity and cumulative affective strain—eroding both wellbeing and dignity.

Making 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, emotional privacy is thus 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 [55, 56]. As a category for systematic study, the very term "emotion" emerged only in the late 19th century—overlapping with earlier concepts such as "passions" and "sentiments" [57]. With a lack of standardized terminology and conceptual coherence persistent within and across fields [58], scientific debates over the nature of emotion continue.

The social sciences generally group emotion theories into three categories: discrete, dimensional, and appraisal-based models [59, 60]. Pioneered in psychology by Magda Arnold and developed extensively by Richard Lazarus, 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 [61, 62]. By this view, emotions arise when an individual evaluates an event's novelty, goal relevance, agency, or norm compatibility [63]. Such evaluations can be swift and pre-reflective or slow and deliberative, but in either case emotions are intelligent commentaries on "what matters here": the same racing pulse might be appraised as righteous anger, stage fright, or competitive excitement depending on one's evaluative frame.

While rich in explanatory power, appraisal-based models remain underutilized in commercial emotion AI due to their contextual and cognitive complexity, and associated challenges with automated implementation [60, 64]. Nonetheless, appraisal theories 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 [1]. Though not typically cited in technical literature, it provides a crucial foundation for this dissertation's treatment of emotional privacy. Because appraisal models bridge embodied states and evaluative meaning, they foreshadow normative requirements for AI to account for moral context. I therefore return to Nussbaum's account in Chapter 7 to develop a normative model that centers emotional meaning, dignity, and human flourishing.

The two dominant meta-theoretical paradigms that permeate both contemporary scientific and applied understandings are discrete and dimensional models of emotion [65]. Their core distinction 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 certain categories of emotions are universal, evolutionarily ingrained, and expressed similarly across cultures. Dimensional models, by contrast, accommodate more flexible, socially and culturally embedded accounts of emotion, with affective states often treated as fluid, overlapping, and context-sensitive. Though these paradigms originate in scientific theory, they implicitly inform applied computing systems. Commercial systems often operationalize emotion within either a discrete or dimensional framework—assigning labels like "happy" or "angry," or estimating degrees of arousal or valence—thereby importing each model's ontological commitments into automated inference [66, 67].

This section organizes a non-exhaustive review of emotion theories around the discrete-dimensional divide, reflecting the dominant schema adopted in affective computing and emotion AI. Yet the boundaries between theoretical traditions are neither rigid nor mutually exclusive: many contemporary models draw from both categorical and dimensional frameworks, some from neither, and disciplinary affiliations do not map neatly onto specific paradigms. While understanding how these dominant theories shape the assumptions embedded in emotion AI systems is essential for evaluating their validity, scope, and implications, equally important is recognizing their common limitation: the presumption that machine-legible traces (e.g., behavioral or biophysiological signals) are necessary preconditions for an emotion to be detected—and by proxy, for it to be acknowledged as existing at all.

2.2.1 Discrete Biological Functions

Discrete, or "basic," emotion theories posit a small set of distinct, mutually exclusive, and biologically hardwired categories of emotion families that are universally expressed. The lineage begins with Darwin and is crystallized in Ekman's Basic Emotion Theory (BET), which identifies six core emotion categories—anger, fear, joy, sadness, disgust, surprise—each tied to dedicated neural programs and stereotyped expression patterns [68, 69].

Ekman's cross-cultural empirical studies, often finding above-chance recognition of posed expressions across cultures, were widely influential in establishing the idea that emotional categories are fixed and identifiable (e.g., see [70, 71, 72]). Discrete emotion theories, including Ekman's, are often paired with biological and physiological assumptions about emotional expression, which emphasize the embodied and communicative functions of affective states and their origins in evolutionary adaptation. These perspectives generally (though not necessarily) 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" [73].

Presuming the automatic physiological signature of a basic emotion can be reliably observed across individuals [74, 71, 75], BET provides an attractive off-the-shelf label set for emotion recognition systems [64]. Yet discrete algorithmic labels also inherit their underlying theory's limitations. First, recognition accuracy is typically higher within cultural groups than across them, indicating that local "display rules" shape perception-of-others' emotion. Second, similar expressions may carry divergent meanings: a smile in Tokyo, Lagos, or Toronto may express compliance, irony, or warmth, respectively. Ekman conceded these complexities, emphasizing in later work that not all emotions have unique facial markers, and prototypical cues like smiling span multiple categories [76].

Meta-analyses reinforce these limitations. Durán et al. report weak correlations between muscle movements and self-reported affect [77]. Barrett et al. highlight that emotion recognition training datasets—generally consisting of actors' staged photos labeled by third-party guesses, with little input from first-person reports or diverse cultural samples—produce brittle models that are prone to context errors and invalidly assumed to detect internal states [78]. Often critiqued for over-compression of affect, systems trained on categorical mappings risk reverse inference errors: assuming that expression x signals emotion y, regardless of context [79]. They are also prone to demographic and cultural bias [80, 81, 82, 83]. These concerns do not deny cross-cultural regularities in emotional expression—wide-eyed startle responses, noses wrinkled in disgust, voice tones softened to soothe, and the like still recur globally with striking frequency [84]—but they caution against assigning deterministic meanings to expressive cues.

Despite their limitations, discrete models remain appealing for their simplicity and compatibility with classification algorithms. In practice, this tractability helps explain why they have dominated commercial emotion AI: fixed emotion categories map neatly onto supervised learning labels. The result is a pipeline where the model's apparent precision rests on theoretical simplifications that are rarely made explicit to the end-users who rely on them to draw conclusions [78]. To situate their rise, the following sections briefly 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 naturalism. Darwin's *The Expression of the Emotions in Man and Animals* framed emotions as evolved adaptations, serving communicative and survival functions across species [85]. His comparative work on human and nonhuman animals is widely regarded as foundational to the scientific study of emotion [86, 87, 88]. Darwin influenced the field to view emotional expressions as biologically inherited adaptations: raised eyebrows, for example, both widen the visual field and communicate surprise to others. Less acknowledged in the universalist caricature, however, is Darwin's emphasis that an emotion's instantiation is contingent upon the subjective perception and social situation of the individual.

James–Lange theory. Building on Darwin, James and Lange argued that bodily changes elicited by an "exciting fact" are not epiphenomenal but constitutive of the felt character of emotion: we do not tremble because we are afraid; we are afraid because we feel ourselves tremble [89]. Properly understood, the James-Lange theory's claim is not one that wholly reduces emotion to physiology. Rather, it holds that physiological responses to stimuli form the immediate substrate of the felt aspect of emotion, and that awareness of these changes is a necessary constitutive element—what gives the emotion its "color." Their account left room for the object of emotion, the individual's habits, and the context at hand to shape which bodily changes occur and how they are interpreted.

The James-Lange theory laid the groundwork for emotion theories to treat emotional experience as a byproduct of physiological activation, including contemporary embodied and enactive views of emotion which retain the primacy of the body but reject the requirement of conscious awareness. On such accounts, emotions are enacted through lived, subjective bodily states: emotion is not something we have, but something we do [90, 91]. Many of today's affecting computing and emotion AI systems inherit assumptions from this lineage, operationalizing emotion as biophysiological activation that can be reliabily measured via signals such as heart rate, facial musculature, or brain activity.

Neuroscientific and embodied-functional models. The James–Lange view also informed neuroscientific and embodied-functional approaches, which investigate correlations between emotional experience, bodily processes, and neural dynamics [92, 93]. Cannon–Bard challenged James–Lange by proposing that physiological arousal and the conscious feeling of emotion occur simultaneously and independently [94], while later limbic system models, such as the Papez-MacLean circuit, distinguished affective experience (arising in cortical-limbic regions) from expression (governed by subcortical structures like the hypothalamus) [95]. Joseph LeDoux's work mapping the neural circuits through which sensory information is processed and evaluated emphasizes the amygdala's central role in detecting and responding to biologically salient threats. His studies distinguished rapid, subcortical "low road" pathways from slower, cortical "high road" routes, showing how both conscious and non-conscious elements contribute to the emotional experience [96, 97].

Martha Nussbaum's theoretical work on the structure of emotions elaborates on this lineage, treating fear—rooted in amygdala-based neural circuits—as a paradigmatic "primal" emotion, one that predates and is antecedent to the higher cognitive elaboration unique to humans, and that shares an ancient neural architecture across vertebrate species [98, 1]. In humans, later-evolving cognitive capacities enable complex evaluations that can transform primal emotional responses, giving emotions their distinct interpretive character [1, 99]. Yet as Lisa Feldman Barrett observes, the precise mapping of neural activation to discrete categories or affective dimensions remains elusive [73].

Functional accounts extend biological models of emotion to emphasize what emotions do: regulate arousal [100], guide attention [101], and facilitate communication and affiliation [102]. Herbert Simon, for instance, conceptualized emotion as an "interrupt" system: a fast, adaptive, and sometimes overactive cognitive override that re-arranges priorities in the face of environmental demands [101]. Simon's view echoes John Dewey's early pragmatist account, which synthesized Darwin's emphasis on expression and James's focus on bodily feeling by framing emotion as emerging from goal conflict or internal tension [103].

Overall, neuroscientific and embodied-functionalist perspectives view emotion as deeply intertwined with human cognition and reject the Cartesian legacy that treats emotion as antithetical to intelligence [23]—a stance that influenced early AI research and continues to inform debates about the normative role of emotions in moral and political life [104, 1].

2.2.2 Dimensions of Individual and Social Difference

Challenges to the discrete emotion template come from a broad family of theories that regard emotions as *continuous, context-sensitive, and socially embedded.* Rather than slotting emotions into fixed bins, these accounts complicate the "common view" that a facial configuration or heartbeat pattern maps cleanly onto a universal label [73, 78] by emphasizing emotional difference, gradation, and enculturation.

Indeed, empirical evidence shows substantial variation in how emotions are expressed and interpreted across individuals and cultures. For instance, the same facial expressions can correspond to different emotions depending on context [19]. Lisa Feldman Barrett refers to this as the "emotion paradox": despite strong intuitive beliefs that we can recognize discrete emotions when we see them, scientific findings struggle to support this common view, finding inconsistent patterns in emotional experience and expression—within and across individuals—that suggest a more nuanced concept of emotions [105].

Dimensional theories. Dimensional views derive from Wilhelm Wundt's theory of affect, which proposed that more fundamental feeling states—essential and irreducible components of consciousness—are located along three independent dimensions: pleasure—arousal—tension [106, 107]. Building on Wundt's early triad, contemporary dimensional models typically locate emotions as points in a two- or three-dimensional space: valence (pleasant ↔ unpleasant), arousal (activated ↔ calm), and sometimes dominance or control [108, 109].

Because points in a space are language-agnostic, dimensional models travel more easily across cultures and domains where emotion words do not align one-to-one [110]. For affective computing, the dimensional geometry captures a richer signal—emotional blends, shifts, and ambiguities that categorical schemes can over-compress and neglect. Yet dimensional models of emotion still require contextual interpretation to say *which* worry, thrill, or melancholy a vector actually signifies [60].

Psychological constructionism. Closely coupled constructionist theories view emotions not as fixed biological facts, but constructed interpretations based on context, culture, and learned associations [105]. Bodily fluctuations in valence and arousal become emotion categories like "anger" or "fear" only when interpreted through learned emotion concepts and situational cues [73]. Hence, there is no universal "anger circuit" waiting to be detected; emotions are assembled in the moment, emerging from more basic psychological components of core affect (characterized as a continuum along valence and arousal dimensions) and conceptual knowledge [111]. Constructionist accounts help explain why similar scowls in the same person can signal moral disgust in one setting and playful teasing in another [112]—and why predictions by automatic emotion detection systems, without access to individual cognitive appraisals, are inherently limited.

Brunswik's lens and interactionist views. Brunswik framed emotion perception as a probabilistic inference: observers sample noisy cues and apply their own priors [113]. Interactionist theorists extend this point to argue that emotion is co-constructed in situ through norms, roles, and reciprocal feedback [114]. 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.

2.2.3 Evaluative Appraisals: A Unified Theory of Emotions

Martha Nussbaum's theory of emotion unifies historical philosophical traditions with contemporary empirical insights. Rather than privileging one disciplinary vantage point, Nussbaum synthesizes these perspectives into a unified theory that defends emotions as a universally human phenomenon with intrinsic moral and political significance—rooted in embodiment, shaped by context, and 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—how we make sense of the *external* people, events, and ideas we encounter through our own *internal* lens, formed from our values, beliefs, attachments, and life plans. Fear, for example, 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: shaped by personal history, social context, and cultural norms, this interpretation necessarily evaluates the *object* of emotion in relation to our own *subjectivity*—who we are and what personally matters. Integral to our capacities for moral agency, emotions are *eudaimonistic*: structured by what we take to be good and indispensable to grasping what is at stake in living well, emotions animate the individual pursuit of a flourishing life [1]. Recognizing the nature and function of emotions as such enables ethical engagement with others, disclosing their humanity and our own [15].

Nussbaum's account contributes a feminist intervention to emotion theory and moral and political philosophy, confronting longstanding attempts to discredit emotional life as too irrational and epistemically unstable to warrant serious normative inquiry. Affirming the particularities of emotional life as normatively significant, she defends emotions as bearing not only instrumental worth, but also *intrinsic* moral and political value. Respect for emotional life, she insists, is a constitutive condition of dignity [15]—not because emotion is a fragile substrate, but because it is the substance that imbues us with inherent worth and renders us intelligible as agents striving for meaning. Normatively rich, empirically grounded, and theoretically rigorous, Nussbaum's theory provides unmatched analytical power for evaluating the moral and political stakes of emotion, privacy, and agency in AI-mediated societies.

2.3 What is Emotion AI?

Emotion AI refers to technologies that infer, model, and interact with human affect from observed data patterns across modalities such as facial expressions, vocal tone, body posture, physiological signals, and text. Two longstanding challenges in affect theory make this computational translation especially fraught: what Prinz describes as "the problem of parts" and "the problem of plenty" [115]. The former concerns which aspects of emotion—physiological, phenomenological, behavioral, mental, expressive—are treated as essential for detection, while the second refers to the surplus of competing definitions that resist straightforward integration. When a field lacking conceptual consensus becomes the basis for technical architectures deployed in everyday contexts, these unresolved debates harden into design choices with material social and political consequences.

From the earliest work in affective computing onward, critics have warned that these systems reify narrow assumptions about what emotions are, how they are expressed, and how they should be interpreted [64, 116, 117]. Boehner et al. emphasized the tendency to treat emotion as an internal, individual property, neglecting the interactional and socio-cultural dimensions that give emotional expressions meaning [116]. These critiques remain salient: as the empirical results in Part II demonstrate, commercial deployments continue to treat emotion as decontextualized "information" to be managed and disciplined, reproducing the same reductive pitfalls such systems set out to transcend [118, 119].

2.3.1 Early System Modules, Labels, and Goals

2.3.1.1 Emotion Classifiers

Until the early 2020s, commercial emotion AI systems were predominantly modular and modality-specific. Such pipelines typically separate preprocessing, feature extraction, and classification, frequently using support vector machines (SVMs), hidden Markov models (HMMs), or shallow neural networks [120].

The input modality strongly influences model choice: facial expression datasets, for example, are generally mapped to discrete emotion categories, while biosignals are modeled in dimensional affect spaces [121, 122]. Classifiers learn to map input features to emotion labels that follow suit: discrete (e.g., basic emotions), dimensional (e.g., valence–arousal–dominance), or hybrid (e.g., discrete categories with graded intensity or arousal) [60, 122]. Labels originate from human annotation, based either on self-reported emotions from participants, observer ratings of perceived emotion, or both. In practice, most commercial and academic datasets rely heavily on observer annotations [78, 123]. Stemming partly from the cost and intrusiveness of collecting self-reported emotions, this reliance embeds cultural and demographic norms in the data while biasing systems toward replicated social judgments. Disparities in accuracy, calibration, and intensity scaling have been documented across demographic subgroups [124].

2.3.1.2 System Model

Earlier emotion AI systems are often described in terms of three main "building blocks" [120]. *Emotion recognition* processes multimodal data to predict affective states based on training data patterns. Goals range from inferring the stimulus that elicits an emotional response (e.g., music [125], opinions [126, 127]) to predicting how an observer would perceive a target's expression and approximating the felt state of the target. Translating lived emotional cues into formalized data objects, commercial adoption of emotion recognition raised concerns about bias, normativity, and the reduction of nuanced affect to fixed categories [80, 81, 128, 129].

Emotion augmentation adds post-recognition layers—reasoning modules, goal-setting mechanisms, behavioral adaptation layers—to adjust system outputs in response to inferred affect, targeting objectives including modulating user engagement, personalizing content, influencing decisions, or adapting interaction style [130, 131]. With system adaptations often occurring without user awareness or control [132, 133, 134], new capacities for cognitive and emotional influence were introduced, intensifying manipulation concerns [135, 136, 137]. Because augmentation layers in earlier, narrower systems were largely rules-based [138], their logic was rigid yet traceable, making governance more feasible—albeit still constrained by cultural and scientific biases [139, 140]—than in emerging systems.

Emotion generation produces outwardly expressive behaviors or simulated affective states in machines. Applications outside of research or demonstration contexts remained uncommon [120] until generative AI systems entered the market.

Modular designs offer interpretability benefits, yet struggle with domain transfer and unimodalities [122]. As multimodal deep learning matured, the field has shifted toward unsupervised or self-supervised representation learning, more automated processes (e.g., feature extraction and selection), and more unified architectures that more tightly couple recognition and augmentation tasks, sometimes within a shared training framework.

2.3.2 Emerging Emotion AI Architectures

High-capacity models now ingest multimodal streams such as video, audio, and biometrics data alongside contextual signals like interaction history and environmental cues—increasingly capable of adapting outputs in real time. Transformer-based and contrastive-learning architectures can integrate facial, acoustic, and textual modalities end-to-end, while self-supervised and multimodal embedding techniques learn features directly from raw, unaligned data [141, 142, 143, 122]. Shared cross-modal embedding spaces enable transfer learning and multi-task optimization, allowing emotion recognition trained in one domain to steer behavior in another [144, 145].

Deep architectures can model latent affective states without mapping them to human-interpretable labels. Generative modules can then act on these latent vectors—producing affective speech, facial animation, or emotionally inflected dialogue—even when the underlying representation is not expressed as a labeled emotion [122, 146]. The ability to condition generation on opaque affective representations simultaneously increases adaptability and magnifies governance risks.

2.3.3 Governance and Safety Implications

As recognition, augmentation, and generation are increasingly integrated into unified pipelines and the representational space grows more opaque, the same latent features may drive multiple tasks—adaptation, persuasion, content generation—in parallel. This convergence makes it increasingly difficult to determine where and how patterns are learned, which features drive outputs, and when systems cross from adaptive personalization into covert influence [147, 148, 149].

Regulatory frameworks that categorize governance elements such as special data types or harm taxonomies will struggle to operationalize legislative demands in systems built on high-dimensional embeddings and end-to-end architectures. These designs complicate audit trails and transparency reporting, while generative modeling, domain generalization, learning from latent representations, and temporal representational drift all further hinder explainability. Without governance strategies adapted to these architectures, risks will continue to outpace the possibility for effective regulatory intervention.

The empirical studies presented in Parts II and III, conducted between 2020 and 2023, were situated at the transition point between earlier modular emotion AI designs and the emergence of deep learning-based and end-to-end architectures , together with more capable generative models. Through interviews, content analysis, and a mixed-methods factorial vignette survey, I examine how emotion AI was marketed, perceived, and experienced in real-world domains during this period of rapid technical change. While grounded in earlier architectures, the findings are relevant to newer systems, surfacing both the unique *emotional privacy* risks they pose and the limits of existing privacy and governance frameworks to identify and mitigate them. I address these gaps with a novel normative governance framework introduced in Part IV, and reflect on the future of emotional privacy and governance in next-generation AI systems in Part **??**.

Part II: The Empirical Case for Emotional Privacy

Across the studies presented in Part II, a recurring theme concerns how emotion AI deployments in both institutional and consumer settings facilitate emotion commodification—the exploitation of human affect for corporate gain.

In Chapter 3, interviews with social media users reveal predominantly negative attitudes toward emotion AI-enabled wellbeing interventions. Participants express deep skepticism that these systems could genuinely support users' wellbeing needs, and apprehension that interventions may manipulate emotional states for platform optimization. Underneath their unease lies concern about moral legitimacy: participants emphasized the irreducibility of human qualities—moral integrity, professional expertise, benevolence, shared humanity—in contexts of wellbeing support and care, doubting that machines could adequately replicate them. These results suggest that automated emotional interventions are perceived not only as untrustworthy and insufficient, but fundamentally misaligned with the forms of emotional reciprocity and moral support that people desire.

Chapter 4 turns to the workplace. A content analysis of promotional materials from emotion AI hiring services reveals 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, corporate affective preferences are crystallized 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.

Interviews with workers in Chapter 5 reveal emotion AI functioning as a tool of emotional surveillance and discipline, exposing workers' inner lives to managerial oversight in a context where consent is inescapably coercive. Privacy intrusions are described as essential functions of the system and viewed as primary drivers of harm. Participants report eroded emotional autonomy, intensified emotional labor, and suppressed worker dissent—particularly among those already marginalized by gender, race, or role. Echoing Elizabeth Anderson's critique of "tyranny" in the U.S. workplace [150], Chapter 5 illustrates how emotion AI deployments amplify structural asymmetries by narrowing individual agency over emotional life—extending unchecked institutional power to reorient labor norms and restrict opportunities for meaningful social participation. The findings surface *emotional privacy* as a distinct privacy interest—one that demands robust and targeted protections commensurate with the stakes introduced by emotion AI.

Uncovering a range of ethical and privacy risks, Part II underscores 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. Introducing emotional privacy as a critical lens for surfacing harmful data practices that expose, instrumentalize, and reshape the experience of human emotion, the results demonstrate the need for governance frameworks capable of identifying and mitigating its novel threats.

CHAPTER 3

Emotion AI on Social Media: Data Subjects' Conceptualizations of and Attitudes Toward Automatic Wellbeing Interventions[1]

3.1 Introduction

Social media platforms are distinctive contexts where people both share emotionally sensitive content and are emotionally affected by platform-mediated interactions [151]—dynamics that have long attracted interest from the companies themselves. In 2014, academic researchers and Facebook scientists collaborated on a large-scale experiment to test whether "emotional states could be transferred to others via emotional contagion" [2], deliberately influencing users' emotions by modifying content recommendations and analyzing their impact on emotional engagement. The public backlash was swift and severe, with critics condemning the study as a covert form of emotional manipulation [152, 153, 154]. Public commentary revealed a variety of concerns: being unknowingly manipulated, enrolled in research without informed consent, violations of normative expectations around data use, and a broader erosion of trust in Facebook's practices [155].

All the same, applications of emotion AI to social media have since expanded, in which emotionally expressive content is abundant and platforms are incentivized to model and modulate user affect. Advances in computing power and data availability have fueled the development of large-scale emotion recognition systems trained on user-generated content [156, 157]. These systems analyze emotional patterns to predict health risks [158], powering research in computational social science and clinical domains alike [4, 159, 160, 161, 162]. Digital phenotyping, which analyzes health and behavior from device data, shows growing interest in psychiatry and psychology for improving early detection and intervention of mental illness [3, 163, 164].

[1] This 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.

While academic researchers often frame these systems as supporting wellbeing, companies deploy them for consumer profiling and targeting, and governments explore their use for monitoring sentiment and assessing security risks [165, 166]. In the United States, inferred emotional data largely falls outside the scope of existing health privacy laws [167], enabling its use and sale to insurers, advertisers, and other third parties—often without users' awareness or consent [158, 168]. Concerns about threats to autonomy, privacy, and psychological safety have been raised widely [167, 169, 158, 170, 171, 172]. Yet this discourse largely excludes the perspectives of those directly targeted: the social media users whose emotional expressions train these systems and who may be subject to their influence.

This study centers those data subjects to explore their conceptualizations and attitudes toward emotion AI-enabled wellbeing interventions. We conducted semi-structured interviews with 13 adult social media users in the United States who had experienced both positive and negative meaningful events in the past year and reported sharing about them online. Participants generally expressed negative attitudes toward wellbeing-related interventions, shaped by unprompted comparisons to traditional, human-delivered support. Participants described four attributes essential to acceptable intervention: (1) helpfulness and authentic care; (2) personal and professional expertise; (3) moral integrity; and (4) benevolence grounded in shared humanity. These qualities, they emphasized, could only be embodied by human agents. AI-driven interventions, by contrast, were perceived as fundamentally incapable of possessing these qualities—eliciting mistrust, discomfort, and unease. Some participants acknowledged potential social value in automated wellbeing interventions at scale, citing mental health research, care access, and egregious harm prevention as possible benefits. Even so, they remained skeptical, pointing to potential harms such as retraumatization, health misinformation, surveillance overreach, and flawed predictions. Participants identified several conditions for acceptable use: constraints on data flow, inference accuracy, contextual sensitivity, and demonstrable benefit for the user.

Contributing empirical insights into how emotion AI on social media is perceived by those it implicates, these findings underscore an ethical gap where data practices that infer and influence user emotion are deployed without meaningful involvement from those most affected. When data subjects are treated as abstract "others"—objects of algorithmic prediction, judgment, and intervention whose wellbeing is assumed and engineered, rather than as persons with inherent dignity, needs, and values of their own—their capacity to live as agents of their own lives is preemptively constrained. These insights illuminate ongoing debates about the ethical and political conditions for socially responsible AI in emotionally intimate domains.

3.2 Background and Related Work

3.2.1 Emotion Inference and Wellbeing Interventions on Social Media

Social media platforms have become central sites for harvesting emotion-rich data [171, 173, 174, 175, 176, 177, 164], with researchers across disciplines exploring the promise of emotion inference for public health surveillance, early diagnosis, and real-time mental health intervention [178, 179, 180, 181, 182, 165, 183, 184]. Increasingly, platforms are considered not just as sources of emotional inference, but as potential delivery channels for wellbeing-related interventions with an emphasis on tailored interventions for specific conditions such as schizophrenia [185, 186], depression [187, 188, 162, 164], postpartum depression [189], and post-traumatic stress disorder [190]. Facebook's suicide prevention program remains a prominent example: using natural language processing models to flag posts suggestive of suicide risk and identify users in crisis, the system triggers messages to users, referrals to crisis services, or welfare checks by police following human review [191, 184, 192]. These interventions have drawn both praise for crisis outreach and criticism for privacy intrusions [193, 194, 195].

3.2.2 Human Values and Ethical Stakes

Scholars have raised ethical and privacy concerns about emotion AI on social media, highlighting risks to vulnerable populations [171, 196], lack of transparency and fairness [197, 198, 199, 200], and undermined user autonomy [197, 167]. Yet empirical engagement directly with data subjects remains rare. Few studies investigate how those whose emotional expressions power these systems perceive or experience them, particularly in the context of wellbeing-related interventions. The empirical work that is available, however, suggests a well of discomfort.

Ford et al. found that social media users were uneasy with emotion inferences for targeted mental health advertising, and even more so when human reviewers were involved [201]. Costello and Floegel found similar concerns among people with mental illness, who distrusted commercial platforms' motives and questioned whether their emotional data would be used responsibly [202]. These findings reflect a disjuncture between the aspirations of system designers and the lived concerns of users.

3.3 Methods

3.3.1 Recruitment

We conducted semi-structured interviews with 13 adult social media users in the United States, each lasting between 77 and 120 minutes (average: 106 minutes). Participants were recruited via a screening survey, with interviews conducted over voice or video call and transcribed for analysis. Calls for participation were shared via social media, professional networks, and Craigslist Houston and Detroit. Participants received a $30 honorarium. The study was approved by our institution's IRB.

Screening questions included inquiries regarding age, location, and social media activity. Only adult respondents residing in the U.S. who reported sharing both positive and negative personal experiences on social media in the past year were considered eligible. Out of 100 survey responses, we invited 20 individuals to interview; 13 completed informed consent and participated in interviews.

Interview invitations were extended iteratively based on survey responses and evolving sample composition, with the goal of capturing diverse identities and social media sharing experiences. Positive life events participants reported posting about include career accomplishments, educational attainment, and home ownership; negative events include job loss, health concerns, and relationship complications. To align with the study's focus on emotion-related interventions, participants' experiences sharing these life events online served as reference points for scenario-based probing during the interview, enabling participants to reflect on their attitudes and imaginaries regarding automated wellbeing interventions. In three cases where the participant was acquainted with the primary interviewer, a secondary researcher conducted the interview to preserve data integrity.

Participants ranged in age from 22 to 58 (average: 32.4 years). The sample included nine women, one man, and three participants who identified as gender-fluid, agender, or genderqueer. Racial/ethnic backgrounds included one Indian, two Asian, two Black, and eight white participants. Educational attainment varied: five held college degrees, six had graduate degrees, one had completed some college, and one had not completed high school. All participants were active social media users: eleven reported using Facebook regularly; other platforms included Instagram, Linkedin, Twitter, Tumblr, AO3, Reddit, Snapchat, Twitch, YouTube, and Discord.

3.3.2 Study Design

Interviews followed a semi-structured protocol to allow for exploration and flexibility. Scenarios are a well-established method in HCI and CSCW research for eliciting human values in contexts involving emerging or unfamiliar technologies [203, 204, 205, 206, 207, 208]. This method is particularly valuable in emotional contexts, where prior work suggests participants tend to respond with realism and emotional salience [209].

Interviews began with questions about participants' social media use, emotion sharing behaviors, and privacy expectations. When appropriate, interviews referenced information from the screening survey (e.g., "you had mentioned...") and encouraged participants to recall or revisit specific posts. No difficulties with recall were observed. The full interview protocol is provided in Appendix A.1.

Scenario prompts were presented via a randomized Google Doc tailored to each participant's preferred platform and emotional context. For both positive and negative experiences (as defined by the participant), the document included a variant of the following text:

> *"Think about a personal experience that brought out [positive/negative] emotions for you—maybe one we discussed earlier. Imagine you had shared this on [platform], explicitly sharing how you felt. Now imagine the platform used computational methods to detect those emotions at the time of posting."*

Participants were asked to describe which emotional experience they selected in order to establish emotional context for the scenario. If appropriate and time permitted, a second experience was discussed. These references typically drew on prior survey or interview disclosures, though participants occasionally introduced new examples.

Our interest was not in the emotional experiences themselves, but in participants' perceptions of the content they posted about those experiences as data inputs for emotional inference and automated intervention. Once the scenario was anchored, interviews probed both for general attitudes about emotion recognition practices on social media and specific downstream applications, including targeted advertisements and automatic wellbeing interventions. We defined the latter broadly as any automated action aimed at supporting a user's wellbeing or emotional state. For example, participants were asked how they would feel if a platform used emotion inferences "to intervene in some way to support [their] wellbeing or help [them] feel better." Following work on algorithmic folk theories emphasizing that how people perceive technology can shape their responses as much as how it functions [210, 211, 212], we kept prompts intentionally broad.

These choices allowed us to surface a wide range of values, assumptions, and imaginaries about emotion AI-enabled wellbeing interventions.

3.3.3 Data Analysis

Interviews were conducted and transcribed by the second author and another member of the research team. The first author led the data analysis presented in this chapter, using thematic analysis focused on participants' conceptualizations of and attitudes toward automatic wellbeing interventions. The process involved open coding, followed by higher-level grouping and axial coding to identify relationships and refine analytic categories [213]. Emerging codes and interpretive insights were discussed with the second author to support reflexivity.

Papers analyzing the same interview transcripts have reported results in separate publications initiated by the second author [214, 215], which pursued different research questions and analytic approaches. The first author was not involved in those publications.

3.3.4 Limitations

This study was not designed to produce representative or generalizable findings [216]. Rather, our goal was to develop interpretive insights into how social media users conceptualize and evaluate automatic wellbeing interventions enabled by emotion AI. In-depth interviews with a small, purposefully selected sample allowed us to explore these questions in rich detail. While our sample is demographically distinctive in some respects, we observed notable consistency across participants' narratives, lending confidence to the thematic validity of our findings.

Our participant pool skewed toward women and minority-gender individuals. Our sample also reflects sampling trends in HCI research on emerging technologies, where participants tend to hold college or graduate degrees and report high familiarity with digital platforms [217, 218]. While this may limit the transferability of our findings, it enabled us to engage in-depth with participants already accustomed to emotional expression in online settings.

Some degree of self-selection bias is likely, with several participants reporting they adjusted privacy settings or posting behaviors to manage visibility. While minor inaccuracies in recall are possible, they are unlikely to have interfered with the study's core analytic goal.

Importantly, our focus on social media users who share emotional content excludes those who do not—but who may still be subject to emotion recognition and intervention. Understanding the perspectives of less expressive users is a direction for future work. Future research should also examine emotion AI attitudes and perceptions among social media users across a wider range of educational backgrounds, racial and ethnic groups, age cohorts, cultural contexts, and geographic locations.

We particularly emphasize the need to engage social media users with mental health conditions. Prior work has shown that people with eating disorders and other psychiatric diagnoses use social media for emotional and social support [219, 220, 221, 222, 223]. Emotion analysis by platforms in these online communities can lead to what Feuston and Piper call the *coded gaze*—a reductive reading of vulnerable expression that can result in unwanted intervention. Future work should investigate whether and how users' attitudes toward automatic wellbeing interventions vary by diagnostic status, especially for interventions targeting their conditions directly.

3.4 Findings

We begin by examining participants' predominantly negative attitudes toward automatic wellbeing interventions on social media and the factors shaping those views. While some expressed ambivalence—viewing interventions as potentially beneficial for *others*—participants remained largely skeptical of the technology and uncomfortable with being personally subjected to it by platforms. They emphasized concerns about the harms such interventions might introduce and underscored the importance of individual control over whether one is targeted. We conclude our presentation of results by highlighting specific qualities on which participants' attitudes toward the technology depended.

3.4.1 What Only Humans Can Offer

Participants' negative attitudes stemmed from participant-initiated comparisons between wellbeing interventions traditionally delivered by humans and imagined future interventions enabled by emotion AI. A prevalent theme was the human versus AI dichotomy: participants questioned whether AI could replicate certain attributes they deemed essential to wellbeing-supportive roles: (1) helpfulness and authentic care; (2); personal and professional expertise; (3) moral integrity; and (4) benevolence grounded in shared humanity.

3.4.1.1 Helpfulness and Authentic Care

Participants doubted the ability of automatic wellbeing interventions to deliver meaningful help or genuine care. P1, who disclosed experience with mental illness, reflected on past experiences searching for suicide-related information on Google: *"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."* They later added: *"I don't know that a computer is able to serve the right information to help someone*," expressing skepticism that algorithms could provide appropriate support in moments of distress.

Echoing the need for authenticity, participants reported uncertainty that an automatic wellbeing intervention could offer support that felt genuine and personally helpful. As P5 remarked: *"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."*

Such reflections highlight how algorithmic interventions may feel impersonal, lacking the emotional resonance of support offered by caring, trained humans—particularly for individuals living with mental illness.

3.4.1.2 Personal and Professional Expertise

Participants expressed skepticism that AI-enabled wellbeing interventions could match either the professional or personal expertise of human support systems. Regarding professional training, participants emphasized the value of care from qualified human experts. As P3 noted: *"I don't know, a therapist went to grad school for it. They've studied the thing."* P10 similarly remarked: *"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."*

On personal experience, participants underscored the trust placed in friends and community members during times of crisis and struggled to imagine extending that trust to AI. As P3 reflected: *"They also have a certain amount of community attached. I don't feel like an AI could get there...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."* For many, the emotional needs these interventions aimed to address were already met by empathetic, personally connected humans. As such, participants questioned whether algorithmic alternatives were capable of fulfilling those roles and could ultimately be welcomed. This skepticism may reflect our sample's predominantly college-educated background, as individuals with higher education tend to have more robust support networks [224].

3.4.1.3 Moral Integrity

Skepticism toward automatic wellbeing interventions on social media emphasized platforms' presumed financial motivations. Unlike mental health practitioners who are held to ethical standards of conduct that engender trust, platforms were viewed as lacking moral commitments and driven instead by profit. This misalignment fueled cynicism that algorithmic interventions—and the platforms that deliver them—could be trusted. As P5 voiced: *"[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."* Similarly, P2 questioned the primacy of ethical over financial incentives: *"I'm just not so convinced that the financial incentives of the companies are such that ... ethical incentives would take priority."*

P9 echoed these concerns, stating: *"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."* While P9 acknowledged that support might feel helpful in moments of loneliness, they remained wary of underlying motivations: *"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 reflections highlight how perceptions of commodification—of emotion, vulnerability, and care—shaped participants' negative views toward emotion AI-enabled wellbeing interventions. For many, the involvement of profit-driven platforms fundamentally compromised the moral legitimacy of any wellbeing intervention they might offer.

3.4.1.4 Benevolence through Shared Humanity

Participants questioned whether automatic wellbeing interventions could ever be received with the same openness or trust as interventions offered by another human being. In contrast to benevolent disclosures and interventions that take place within trusted relationships (e.g., with a mental health professional or close friend), participants described algorithmic interventions as intrusive, unsettling, and lacking in interpersonal resonance. As P13 put it: *"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."*

Although some participants acknowledged the potential usefulness of automatic wellbeing interventions, they emphasized that such interventions lacked the mutual understanding and moral context that human interactions provide. Participants viewed human reciprocity as a core requirement for truly supportive care—something algorithms and the platforms that deploy them could not embody. For example, P3 expressed discomfort that algorithmic systems lack not only personhood, but the shared human experience that underpins moral resistance to unacceptable behaviors like manipulation: *"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 that authentic support requires emotional compassion and relational intent—qualities grounded in the shared element of humanity. As P1 noted: *"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 remarked: *"...it's good to be accurate [with recognizing and predicting emotions], but there's no* *humanity in it, right?"*

These reflections suggest that participants did not simply reject emotion AI-enabled wellbeing interventions on social media because of technical limitations or privacy concerns—they questioned their very legitimacy as care. Without shared humanity, algorithmic interventions were perceived to fundamentally lack the moral reciprocity and benevolence that underwrite the trust required for support to be experienced as genuinely helpful.

3.4.2 Emotion AI Uses for Social Good

Some participants differentiated between their personal discomfort with automatic wellbeing interventions and a cautious openness to their potential social value. With reservations, a minority of participants envisioned these technologies as a possible *social good*—especially when conceptualized as benefiting *others.* In these cases, participants imagined emotion AI interventions as useful in three domains: (1) supporting academic research; (2) increasing access to wellbeing support; and (3) preventing egregious harm. These potential use cases were not prompted by interviewers but emerged organically during discussion. Even when voicing tentative support for such uses, participants stressed that emotion data collection and subsequent interventions should be transparent and consensual.

3.4.2.1 Supporting Academic Research

Several participants expressed conditional support for automatic wellbeing interventions if they were developed and used by researchers. For example, P7 said *"I would want that to be used in research, and in mental health studies."* Participants conveyed greater trust in research-driven interventions than those led by platforms. As P3 suggested: *"I 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."'* Even participants otherwise resistant to emotion data collection made exceptions for academic uses. P5 noted *"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."*

These more favorable attitudes are consistent with prior work showing that some users—particularly those who are college-educated or already enrolled in research studies—are relatively comfortable with academic uses of social media data [225]. Our own sample, also predominantly college-educated, likely contributed to this pattern of greater trust in academic researchers.

Still, participants emphasized that such research should be transparent and voluntary—where individuals provide informed consent to both the emotion data collection and corresponding intervention. P5 added: *"I wouldn't be okay with that being used without anybody's knowledge* *because that's just shady."* P7 similarly stressed, *"I would want that information known to me as the user."* In short, participants distinguished between exploitative and ethically grounded uses of emotional inferences and related wellbeing interventions, with support contingent on awareness and meaningful consent.

3.4.2.2 Increasing Access to Wellbeing Support

Several participants suggested that automatic wellbeing interventions could benefit social media users without strong social support networks. For example, P8—while skeptical of emotion inferences used for advertising—responded positively to its use in acute moments of distress: *"That feels more like the social good side of it...[if] someone is having an acute moment then this platform can be used to actually provide resources that might help."* Participants recognized that wellbeing interventions might serve as a lifeline for vulnerable users. P6 said: *"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 echoed: *"You see kids on there and they might have a problem. They need help."*

Even in these cases, participants maintained a distance between perceived social benefit and personal acceptance: they acknowledged the potential for increased access to support but remained hesitant to welcome such interventions for themselves. Future work could explore how access to social support networks shapes attitudes toward emotion AI-enabled wellbeing interventions.

3.4.2.3 Egregious Harm Prevention

A commonly reported justification for automatic wellbeing interventions was the prevention of serious harm such as suicide, violence toward others, or domestic terrorism. Participants considered such purposes as legitimate contexts for intervention. As P11 remarked: *"I think it could be a very good thing...for people who are an immediate threat to themselves or others."* P8 likewise considered that if someone is *"searching about ways to commit suicide or ways to hurt someone...I feel like the social good [matters because] someone's bodily safety is at risk."*

P6 took this further, suggesting that such tools could help prevent school shootings by prompting timely intervention: *"we need to monitor posts a little more closely...if there's somebody who's vaguely talking about a school shooting or something...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."* Here, P6 invoked a public safety rationale to justify limited surveillance of social media to prevent acts of domestic terrorism such as school shootings, implying a tension between urgency and accuracy.

In summary, some participants imagined automatic wellbeing interventions on social media more positively when considering their collective benefit rather than personal impact. This shift, however—where social good supersedes individual concerns—was still contingent on meeting minimum standards like transparency and consent. Future research should examine these boundary conditions more closely, especially how trust and perceived legitimacy vary by personal and societal frames.

3.4.3 Emotion AI's Potential for Harm

When considering automatic wellbeing interventions on social media, most participants maintained negative attitudes—especially when reflecting on the potential impact on others. Chief among their concerns were the various ways these interventions might introduce harm and the need for both individual and external controls to mitigate those risks.

3.4.3.1 Potential Harms

Participants identified a broad range of possible harms, including risks of re-traumatization, the spread of inaccurate health information, inappropriate surveillance, and interventions based on faulty predictions.

Some expressed concern that the interventions themselves could provoke negative emotional reactions like anger, frustration, or distress in already vulnerable individuals. P10 offered a vivid example: *"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."'* This reaction highlights the risk of re-traumatization through impersonal or poorly timed interventions. Others worried that if emotion-based interventions became commonplace on social media, people might mistake them for credible medical advice. P7 warned: *"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."* Here, concern centered on misinformation and overreliance, particularly among those with less health literacy.

Participants also expressed unease about surveillance infrastructure and the risk that such data could be misused—especially by third parties such as parents, employers, or government entities. P3 questioned: *"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 recognized that emotional monitoring systems could enable surveillance overreach and invite ethically questionable re-purposing of the gathered information, particularly for vulnerable users like children.

Finally, the risk of inaccurate predictions—especially in high-stakes contexts—surfaced as a serious concern. P6 reflected: *"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."* In these cases, participants raised concern that the danger wasn't a lack of intervention, but the possibility of false positives leading to misidentification or punitive outcomes.

Taken together, these accounts show that participants imagined a broad spectrum of harms—both interpersonal and systemic—stemming from emotion AI-enabled wellbeing interventions. These perceived risks reinforce the need for robust safeguards and accountability mechanisms.

3.4.3.2 Individual and External Bounds

Agency and consent emerged as non-negotiable conditions for ethical deployment. Across the board, participants expressed discomfort with automatic wellbeing interventions that limited the choice architecture for individual control or lacked external regulation, particularly in consideration of the power differential between platforms and users. 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."* Participants like P2 stressed that any intervention perceived as coercive—regardless of intent—crossed a moral boundary.

Some participants were willing to cautiously support automatic wellbeing interventions under narrow conditions—particularly when oriented toward crisis support—but insisted that the interventions remain bounded and accountable. P8 reflected: *"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...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."* Here, the distinction between private discomfort and public crisis generates moral ambivalence: even when acknowledging potential benefits, participants emphasized the need for limitations, transparency, and individual agency.

In sum, participants upheld their skepticism even when imagining broader social applications. Whether targeting themselves or others, automatic wellbeing interventions were not seen as benign; for such tools to be considered acceptable, participants stressed the need for both individual and external constraints to reduce the likelihood of misuse and harm.

3.4.4 Conditional Trust and Acceptance

Although some participants rejected automatic wellbeing interventions categorically, others described specific qualities that might render such interventions more acceptable. We identified three conditions that shaped participants' trust and comfort: (1) accuracy; (2) contextual sensitivity; and (3) positive outcomes.

3.4.4.1 Accuracy

Participants emphasized the need for high accuracy. When the stakes of intervention are high, mislabeling could result in more than just an inconvenience—it could cause genuine harm. P3 cautioned that if someone says *"they want to die, that's not always accurate... so if people are at risk, for them to use that...will do more harm than good, that's my fear."* The possibility of false positives or oversimplified inferences generated deep concern among participants.

Related to accuracy was a desire for interventions to be relevant—to feel customized and appropriate to one's actual experience. P12 elaborated: *"[You might be able to learn something about yourself and about the condition too...As long as it's a credible source...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."*

For participants, accurate wellbeing interventions enabled by emotion AI entailed not only technical precision but relevant and actionable insights tailored to their specific personal conditions, experiences, and goals. Insufficiency in these qualities may risk not only ineffectiveness but potential harm.

3.4.4.2 Contextual Sensitivity

The acceptability of emotion AI-enabled wellbeing interventions hinged upon context. Interventions perceived as non-intrusive, delivered timely, and appropriate to the domain (e.g., physical health vs. mental health) were viewed more favorably. P7 illustrates this divide: *"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 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."* While some participants viewed any form of targeted health advertising as inappropriate, others like P7 expressed more nuanced viewpoints that considered ads-based interventions for less sensitive (e.g., non-mental health) conditions as potentially desirable. This example underscores the importance of contextual nuance: most participants rejected blanket justifications for intervention, and instead expected situational calibration and sensitivity to domain-specific norms as a condition for acceptance.

3.4.4.3 Positive Outcome

Participants' attitudes also hinged on the intervention's actual outcomes. Interventions that demonstrably improved wellbeing were described as capable of shifting negative attitudes toward cautious acceptance. P7 offered a pragmatic take: *"Because if it's successful and I feel better, then I feel like I can't be upset about it."*

This trust, however, was not unconditional. Participants wanted assurance that interventions would reliably deliver benefit and avoid unintentional harm. P10 proposed formal standards as both safeguard and assurance mechanism: *"As long as that support is...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."*

These reflections suggest that participants' openness to emotion AI-enabled wellbeing interventions depends not only on technical performance, personal relevance, contextual sensitivity, transparency, or consent. Such qualities may help earn initial trust, but it is positive, tangible impact—whether individuals actually benefit—that sustains it through the felt legitimacy of outcomes. Without that assurance, participants viewed automatic interventions as morally tenuous and socially unproven. Trust in emotion AI wellbeing interventions cannot not be presumed or measured by technical function alone—for legitimate acceptance, it must be earned through demonstrable, validated, and meaningful benefit.

3.5 Discussion and Conclusion

This study explored how social media users perceive and evaluate automatic wellbeing interventions powered by emotion AI. By centering the perspectives of those whose emotions are subject to mining, inference, and intervention, we illuminate how such systems are received not merely as tools for wellbeing, but as moral actors—embedded in power structures, laden with intent, and capable of profound human impact. Participants expressed deep skepticism toward these interventions not only due to concerns about technical performance or data protection, but because such systems were seen as fundamentally incapable of providing the relational, contextual, and moral depth that genuine care demands.

As our findings demonstrate, participants implicitly treated trust as a normative threshold, earned under narrow conditions of acceptability: demonstrable benefit, contextual sensitivity, and credible governance. While participants questioned whether platforms that commodify attention and affect could ever act with the benevolence, expertise, and moral integrity required to genuinely support users' emotional lives, some acknowledged the potential for emotion AI interventions to promote social goods—such as expanding access to crisis resources and supporting mental health research—if deployed with transparent intent, voluntary participation, and strict constraints on use.

Harms such as those stemming from misclassification, retraumatization, surveillance creep, and agency loss should be anticipated and preemptively mitigated—especially in social media contexts where consent, even when offered, operates as procedural theater within coarse-grained choice architectures and opaque data regimes.

Participants' views varied regarding what specific uses of emotional data were appropriate, underscoring the need to accommodate divergent thresholds of acceptability. Across these differences, fine-grained controls and meaningful transparency were seen as essential to preserving dignity, enabling users to retain moral agency even when systems infer, act upon, or intervene in their emotional lives. Our findings indicate that respect for divergent individual needs entails, at minimum: clearly communicating what is inferred, offering granular consent options for how emotional data may be used, providing feedback channels and access to redress, and embedding external safeguards to prevent misuse.

Ultimately, participant perceptions of emotion AI wellbeing interventions on social media reflect a demand for system legitimacy. Their skepticism was not rooted in technophobia or misunderstanding, but in a clear-eyed recognition that interventions targeting the intimate terrain of emotional experience must meet a higher moral standard of justification—one that begins not with what systems can do, but with what people have reason to accept.

CHAPTER 4

Emotion AI in Hiring: Values Underpinning Technosolutions to Labor Problems[1]

4.1 Introduction

Emotion AI hiring services are entering the commercial marketplace with growing force, offering organizations the promise of greater predictive power and control over employment outcomes [226, 227]. What are the implications of emotion AI in hiring for our socio-technical futures? Technologies are not neutral tools; they reflect and reinforce moral and political human values [228], shaping not just what *is* but what *should be* [229]. By examining how values are negotiated and materialized in technology development and deployment, we can better understand how systems impact societal priorities [230, 231]. Hiring platforms are key actors in the promotion and adoption of emotion AI. To surface the normative assumptions driving these deployments, we applied a values-based lens [230, 231] to the promotional materials of 229 emotion AI hiring services to ask:

> *What organizational problems do emotion AI hiring services claim to solve? How do they purport to solve them? And what values underlie these promoted uses of emotion AI?*

Our analysis yields four key insights:

1. **Problem Framing:** Emotion AI hiring services promote their technology as a solution to three core organizational challenges: hiring (in)accuracy, (mis)fit, and (in)authenticity.

2. **Legitimization Strategies:** These services legitimize their technosolutions by aligning them with dominant corporate ideals, presenting emotion AI as a rational response to inefficiencies in human judgment.

[1] This 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.

3. **Operational Mechanisms:** They claim to address these problems through two primary mechanisms: the automatic extraction of a candidate's *affective value*—enabling the algorithmic *commodification* of emotional labor—and the automatic exclusion of candidates based on informational asymmetries and inferred psycho-biological traits.

4. **Underlying Values:** Finally, we identify the core values promoted across these services as techno-omnipresence, techno-omnipotence, and techno-omniscience. Emotion AI is positioned as not just a tool, but a moral imperative—the only entity capable of truly resolving hiring's deepest challenges.

4.2 Background and Related Work

We begin by outlining the role of emotion in hiring, followed by a review of relevant literature on AI in organizations, workplace management technologies, and critiques of AI use in hiring.

4.2.1 Emotions in Hiring

Hiring is fundamentally interpersonal and emotionally charged [232]. Employers rely on a range of signals to assess candidates' human capital, social capital, and demographic characteristics [233, 234], which in turn shape their perceptions of interior traits such as competence, motivation, and emotional disposition [235, 232]. These perceptions are often informed by implicit or explicit stereotypes, group-level assumptions, and personal experience [232]. Rivera's concept of "emotional energy development" describes how the emotional energy interviewers feel toward candidates modulates hiring decisions—not only do employers seek qualified candidates, but they are drawn to those who excite them and with whom they imagine forming close personal and professional relationships [232]. In many cases, interviewers describe the emotional experience of the interaction as the most decisive factor in evaluating a candidate.

Emotional dynamics also influence candidates' behavior. Applicants attune themselves to interviewers' affective responses, using that feedback to navigate the hiring process, such as to negotiate salary offers [232]. Positive or negative emotional cues—excitement, boredom, tension—can tilt outcomes in either direction. Thus, emotional signaling is bidirectional and strategic: both employers and candidates use it to assess fit, express preferences, and exert influence over employment outcomes.

By automating or displacing these emotionally laden interactions, emotion AI hiring systems may fundamentally alter how emotions function during the hiring process. Their one-sided design and algorithmic opacity risk severing the mutual, adaptive feedback loop that characterizes conventional hiring encounters.

4.2.2 AI in Datafied Organizations

Organizations are increasingly adopting emotion AI as part of human capital and talent management strategies, aiming to automate or augment hiring decisions [226]. Emotion AI now appears across the recruitment pipeline: from algorithmic sourcing and matching [236], to automated screening [237], to fully integrated hiring platforms [238].

These data-driven systems rest on ideological claims about how work should be organized and which workers are desirable [239]. As Ajunwa et al. show, automated hiring platforms afford a "managerial frame" in which workers are rendered fungible—on-demand, interchangeable, and easily redeployed across job tasks and organizations [238]. Studying the perspective of technology providers is therefore essential for understanding how data-led practices are legitimated and spread.

Emotion AI in hiring must be situated within these broader logics of workforce datafication. Its claims to assess fit, authenticity, or potential are embedded in larger strategies of optimization, efficiency, and control. Yet these systems do not operate neutrally: workplace surveillance and data extraction disproportionately target and harm marginalized groups, reinforcing racialized and gendered hierarchies of visibility and risk [240, 241, 242]. As Fourcade and Healy note, datafied classifications reflect the values of dominant actors, encoding political and moral judgments about what kinds of behavior—and by extension, what kinds of people—are desirable [242].

Building on this work, this study examines how emotion AI in hiring reflects and reinforces these moral and political dynamics, bringing questions of power, legitimacy, and accountability into sharp relief.

4.2.3 Workplace Talent Management

The use of emotion AI to infer the emotions and affective states of employees and job candidates continues a long-standing organizational preoccupation with accessing workers' inner lives. The roots of personnel selection in industrial and organizational (I/O) psychology can be traced to early twentieth-century figures such as Walter Dill Scott and Hugo Münsterberg, whose work was shaped by Darwinian ideas of "survival of the fittest" [243].

By the early 1900s, U.S. employers were collaborating with I/O psychologists to extract psychological insights about workers—seeking indicators of loyalty, emotional stability, and interpersonal compatibility [244]. Instruments like the Minnesota Multiphasic Personality Inventory (MMPI) became widely used to evaluate candidates not only for personality traits (e.g., neuroticism), but also for conformity to health norms and gendered behavioral expectations [245, 246].

Despite longstanding concerns about the fairness, validity, and discriminatory potential of such assessments [247, 248], I/O psychology has remained deeply embedded in workplace decision-making. The integration of emotion AI into these systems reflects a contemporary digital trans-formation of this tradition—one in which affective data is harvested and analyzed to guide hiring, promotion, and personnel management decisions at scale [249].

4.2.4 Criticisms of Algorithmic Hiring

A widely cited *Harvard Business Review* report warns that AI hiring systems may infer protected attributes such as physical or mental disability in discriminatory ways, all while lacking scientific rigor or mechanisms to ensure fairness across protected groups [250]. The authors highlight a lack of "convincing hypotheses or defensible conclusions" around the validity or ethics of inferring psychological traits from physiological data for hiring purposes, and question whether existing U.S. laws are adequate to address the discriminatory risks these tools pose.

Much of the public discourse has focused on demographic bias. A 2018, *Reuters* investigation revealed that Amazon's internal talent systems systematically disadvantaged women, sparking renewed debate over how machine learning systems reflect and amplify existing inequalities [251]. This led to an explosion of technical research on mitigating algorithmic bias, especially through dataset diversification and model debiasing strategies [252, 253, 254, 255]. However, many scholars remain skeptical that technical solutions alone can resolve these structural harms [256, 257].

Lee distinguishes between *explicit* and *implicit* algorithmic biases, noting that while explicit discrimination may be addressed through legal or policy intervention, implicit forms are harder to detect, mitigate, and redress—particularly when rooted in broader social inequities [258]. Similarly, Nakamura critiques how AI systems trained on internal, non-representative data may perpetuate ableist assumptions and exclude disabled applicants. Crucially, this opacity operates as a feature rather than a bug for organizations, allowing for plausible deniability in the face of discriminatory outcomes [259].

In response to criticism, many AI hiring vendors claim to implement bias mitigation strategies and publish documentation describing their approaches. Yet empirical studies reveal troubling gaps in transparency. Raghavan et al. analyzed disclosures from pre-employment assessment vendors and found that most offer vague, unverifiable claims about fairness, dataset composition, and model validation [257]. When debiasing methods are discussed, they often involve simplistic interventions such as removing features correlated with protected categories—without addressing the deeper methodological demands of anti-discrimination law. The authors argue that voluntary, outcome-based debiasing is insufficient and call for policy frameworks to govern algorithmic fairness in hiring. Building on this, Sanchez-Monedero et al. show that U.S.-based vendor practices often fail to meet European legal standards, and suggest that the U.K.'s regulatory approach offers a useful model for addressing transparency concerns in hiring-related algorithmic discrimination [260].

Scholars have linked emotion AI applications to pseudoscientific traditions such as physiognomy and phrenology, warning that these technologies revive discredited methods of inferring character from appearance and are capable of reproducing structural discrimination at scale [261, 262, 263].

While these critiques are vital, there remains a need for systematic and empirical analysis of what emotion AI hiring vendors actually claim—and what values those claims encode. Adopting methods from prior vendor-facing studies [257, 239, 260], we address this gap in critical AI research to interrogate the moral and political assumptions embedded in emotion AI hiring systems.

4.3 Methods

Technological systems shape the possibilities and expectations of those who interact with them [229]. To understand the values underpinning the uses of emotion AI in hiring, we turned to the public claims made by emotion AI hiring vendors on their websites. Despite the challenges of studying opaque organizational systems, recent scholarship demonstrates that vendor-facing materials offer a rich window into industry practices, ideologies, and imagined futures [257, 260, 263].

We conducted an in-depth content analysis of 229 emotion AI hiring services, analyzing how these vendors describe the problems they aim to solve, the mechanisms they employ, and the values they promote. Following Shilton's work on sociotechnical systems, we treat content analysis as a well-suited method for identifying not only which values are present, but where they are located—whether embedded in technical mechanisms, operational goals, or public-facing narratives [264, 229]. Our approach remains analytically descriptive, focusing not on whether these values are inherently "good" or "bad" but on how they operate within the sociotechnical imaginaries advanced by emotion AI hiring vendors [265, 231]—laying critical groundwork for future work to evaluate the legitimacy and desirability of these values, particularly as they inform systems with profound consequences for human dignity, agency, and opportunity.

4.3.1 Data Collection

Data collection proceeded in three stages: (1) identifying commercially available emotion AI hiring services and their websites; (2) reviewing each site to determine inclusion eligibility; and (3) compiling website content for analysis.

Identifying Emotion AI Hiring Services

We began by consulting four platforms: Crunchbase, a startup directory referenced in prior AI hiring studies [257], and three enterprise software review sites—G2, TrustRadius, and Capterra. Our initial Crunchbase queries used keywords including *emotion recognition, affect recognition,* *emotion AI, emotional AI, emotion AI, emotional artificial intelligence, sentiment analysis, emotion detection, affect detection, and emotion analytics.* These searches returned limited results and failed to identify known vendors (e.g., Retorio), which had been previously cited as using biometric inferences in recruitment contexts [266].

To investigate further, we manually searched Crunchbase for several vendors known to employ emotion AI but found their listings used broader labels such as “Artificial Intelligence” or “Machine Learning.” For example, Retorio’s profile made no mention of emotion inference capabilities, despite its website describing “behavioral analytics AI” that “reveals soft skills” through “psychological science.”

This led us to a critical observation: vendors tended to avoid explicit reference to emotion AI, instead using vague or euphemistic language to describe affective inference—obfuscating their classification as emotion AI vendors. Across multiple vendor websites, we observed the consistent use of non-standard, non-technical terms to describe emotional or psychological assessment capabilities. As a result, search strategies relying on canonical emotion AI terms were insufficient to systematically identify relevant vendors.

Applying Inclusion Criteria

Given these challenges, we pivoted to a broader strategy. We first assembled a comprehensive list of commercial Human Resource software vendors from Crunchbase, G2, TrustRadius, and Capterra. We then manually reviewed each vendor’s website for two inclusion criteria: (1) the product marketed itself to hiring organizations as informing hiring decisions; and (2) it made claims to generate inferences about a candidate’s emotions or other affective phenomena.

To identify relevant vendors, we collected entries tagged under categories including *HR Analytics, Workforce Analytics, Employee Engagement, Employee Recognition, Performance Management, Recruiting Software, Talent Management* and *Talent Intelligence*. This yielded an initial dataset of 3,195 unique vendors.

Dataset Compilation

The dataset was divided among four researchers, each of whom manually reviewed assigned websites using the inclusion criteria above. Non-English sites were excluded. Between May and July 2021, this process produced a final dataset of 229 commercially available emotion AI hiring services.

Website content for each vendor was captured as PDF files using a browser extension and imported into a qualitative coding environment for analysis.

4.3.2 Data Analysis

Three team members analyzed the website content of the 229 emotion AI hiring services, focusing on the claims vendors made about their technologies.

Values can emerge in how technologies define a problem, propose solutions, and interact with stakeholder assumptions [231, 267]. Although values manifest throughout development and design processes [229], we follow a functionalist perspective in treating values as inferable from a system's intended use—its *practical end* [268, 269]. Because values are not directly observable, they must be interpreted from language, symbolism, and narrative framing [270]. Our analysis therefore centered on the language used in vendor claims, interpreting how emotion AI technologies are framed and justified in hiring contexts.

Given the discursive, power-laden nature of these claims, we adopted an interpretivist analytic approach grounded in constructivist and feminist epistemologies [271, 272]. The use of inter-rater reliability (IRR) metrics is inappropriate for this kind of qualitative analysis where codes are a part of the interpretive process rather than the empirical product [273]. Our goal was not to catalog claims at face value, but to interrogate how these claims operate within broader sociotechnical imaginaries. Vendor discourse reflects the perspectives of powerful actors designing systems for other institutions of power [267]—namely, hiring organizations. Coding without interpretation risks replicating these imbalances and legitimizing the very assumptions under critique [274, 273].

To ensure analytic depth and reflexivity, the first author conducted open coding on a randomized subset of sites, developing a preliminary codebook organized by the type of claim. This initial codebook was reviewed collaboratively by the team to build shared understanding. The remaining data were then divided among three researchers, who independently conducted close, line-by-line *in vivo* coding using language directly drawn from vendor claims. This choice served to both preserve semantic nuance and to minimize interpretive divergence during early-stage analysis [275].

The first author met with the research team weekly to discuss emergent concepts and document thematic patterns. As themes stabilized, we refined the codebook and conducted axial coding, grouping open codes into broader conceptual categories. Interpretive differences were reconciled during weekly discussions by collaboratively interpreting relationships among recurring themes and adjusting category boundaries to reflect agreement [276]. Once axial coding was complete, the first engaged in selective coding to delimit codes [276], consolidating the analysis around the core question of how emotion AI hiring services promote their technologies' intended use.

4.3.3 Limitations

Our analysis has several important limitations. First, identifying emotion AI hiring services was complicated by the vague and inconsistent language vendors use to describe their technologies.

Because our inclusion criteria relied on subjective interpretation of publicly available website claims, it is possible that some services were misclassified as using emotion AI. To mitigate this risk, we only included vendors whose websites explicitly referenced the measurement or inference of emotions or related affective phenomena. Nonetheless, some vendors may have been excluded if they were not listed on the review platforms we consulted or if their language did not meet our inclusion criteria.

Second, while our qualitative coding processes prioritized interpretive rigor, our reliance on subjective textual interpretation introduces some degree of analytic variability. However, our collaborative coding procedures and iterative codebook refinement were designed to ensure internal coherence and thematic saturation across the dataset.

Lastly, this study analyzes vendors' *promotional discourse*. The values we identify reflect the desired uses as articulated by vendors—not necessarily the values that emerge through real-world implementations. These claims are shaped by commercial incentives and by the presumed values of client organizations [267]. As such, they should be interpreted as expressions of aspirational or strategic positioning—not as direct evidence of use, impact, or user experience. Future work could build on this analysis by examining how vendors respond to critical scrutiny—whether by modifying claims, acknowledging limitations, or altering product design. Further research might also explore how hiring organizations interpret and implement these tools to offer a deeper understanding of the gap between vendor imaginaries and organizational practice.

4.4 Findings

Values in technology emerge not in abstraction, but in the practical ends technologies are designed to achieve: the problems they purport to solve and the higher-order goals to which those solutions contribute [268]. Our analysis reveals how emotion AI hiring services frame their offerings as technosolutions to three organizational problems: hiring (in)accuracy, (mis)fit, and (in)authenticity. For each, we unpack what the purported problem is, why it is framed as a problem, and for whom. We then interrogate these framings to reveal: (1) the corporate ideals that legitimize emotion AI as a means to address these problems; (2) the mechanisms by which emotion AI hiring services claim to solve them; and (3) the core values embedded in the uses of emotion AI they promote.

Across these cases, we find that emotion AI hiring services purport to create and extract what we term a candidate's *affective value*—transforming emotion, behavior, and inferred internal states into data points that can be evaluated and acted upon. This process enables the *affective commodification* of job candidates, whose inferred psycho-biological data becomes asymmetrically visible to employers and actionable in ways that reinforce organizational control. The services' legitimacy hinges on their alignment with corporate ideals including data-driven decision making, continuous improvement, precision, loyalty, and stability. Ultimately, we identify three interlinked values that undergird the role emotion AI is cast to play in hiring: *techno-omnipresence, techno-omnipotence,* and *techno-omniscience*—positioning emotion AI as the singular authority capable of solving hiring's most intractable problems.

4.4.1 Hiring (In)accuracy

The most prominent claim made by emotion AI hiring services is that their technology improves hiring *accuracy*. ZappyHire, for instance, claims that its AI video interview platform will *"Improve Your Hiring Accuracy by 72%"* through analysis of candidates' personal traits, enabling employers to *"make the right hiring decision with the right data points."* Such claims frame emotion AI as a technosolution to the problem of hiring inaccuracies.

According to these services, hiring accuracy is achieved when decisions are (1) objective, (2) unbiased, and (3) intelligent way. Each of these criteria is operationalized in specific and revealing ways.

Objective Hiring. Emotion AI hiring services claim objectivity not in the neutrality of their inferences, but in the automation and standardization of their evaluation processes. The argument is twofold: that automatic systems apply identical parameters to all candidates, and that these parameters themselves are intrinsically objective.

For example, HiredScore (partnered with the emotion AI provider Pymetrics) offers a *"highly-accurate candidate scoring"* to *"enable a future where hiring is efficient and fair"* by ensuring *"all people are evaluated the same for the same jobs."* Similarly, FaceCode claims its interview platform *"combines objective, standardized evaluation parameters with AI-based insights"* to deliver *"the most accurate and effortless interview reports ever...to help you make the right decisions."* Yet this commitment to objectivity is immediately undercut by platforms such as Eightfold.ai, which promise that *"every configuration and product feature"* can be optimized to suit the hiring organization's preferences. Such customization exposes the moral and political stakes of these ostensibly neutral parameters: as Bowker and Star argue, classification systems embed power [274]. When either employers or hiring platforms define what counts as "fit," "effort," or "engagement," and emotion AI enforces those definitions at scale, subjectivity is not eliminated—it is systematized.

Unbiased Hiring. Emotion AI vendors claim to eliminate bias—not through debiasing algorithms, but by removing human labor from the hiring process. Hiring decisions made by people are portrayed as subjective, error-prone, and driven by guesswork. Emotion AI, in contrast, is cast as a neutral evaluator.

Elevatus, for instance, urges employers to adopt its AI video interview service to *"start making decisions based on reliable data, rather than guesswork."* Similarly, the employee engagement platform Bob promises to help organizations *"base management decisions on evidence, not assumptions"* through its predictive analytics service which profiles individual employees for risk of burnout or attrition. iMocha claims its facial and voice analysis of pre-employment assessments identifies *"suspicious activity"* and scores candidates' "emotional intelligence" in a way that *"eliminates human error in grading,"* thereby ensuring evaluations are *"valid and reliable."* These narratives depict imperfect human decision-makers as the source of hiring inaccuracy—bias embodied. By displacing them, emotion AI is positioned as the path to fairness in hiring.

Intelligent Hiring. Finally, emotion AI hiring services bolster their objectivity and fairness claims with invocations of intelligence. Reejig, for example, describes its a talent management software which profiles and shortlists candidates based on inferred *"soft skills,"* describes its system as a *"mastermind"* that provides *"infinite intel"* to support unbiased succession planning. Here, intelligence refers not to human-like reasoning or skill, but to algorithmic capacity: the ability to analyze and correlate vast datasets with speed and precision.

The talent platform retrain.ai offers a similar pitch: *"accurate matching algorithms"* that *"generate useful, validated, unbiased and actionable workforce intelligence"* by combining data on people, jobs, and training programs. These systems claim to be able to see what humans cannot, and reveal that hidden knowledge to employers.

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

In promoting the use of emotion AI to achieve hiring accuracy, vendors appeal to corporate ideals of data-driven decision making and continuous improvement. More than mere marketing gloss, these ideals function as legitimating frameworks that render the technosolution both desirable and normatively imperative.

Jive, a people analytics and productivity platform that deploys continuous sentiment monitoring to maintain an *"ongoing, real-time read on employee morale and engagement,"* positions itself as a corrective to hiring (in)accuracy. It claims to understand the frustrations of decision-making based on *"hunches, vague statistics and hindsight,"* and invites organizations to instead puruse *"accurate, data-driven insight to guide your tactics, make timely corrections and better target your efforts for maximum impact."* Here, data-driven insight is framed as the ideal condition for strategic responsiveness. Similarly, iMocha emphasizes normative obligation: its emotion AI *"should be arranged for objectivity of scoring, and the elimination of personal judgment concerning correct answers."* Such language reflects how vendors recast their services not only as tools for accurate hiring, but as imperatives for achieving higher-order corporate ideals.

To make data-driven decisions about human capital, organizations must first render human traits measurable. Lattice, a people analytics provider that applies sentiment analysis to employee-generated data, promises to *"measure the health of your organization and take data-driven action to increase productivity, decrease employee turnover, and build a winning culture."* Organizational success here depends on the quantification of internal states like morale, motivation, or alignment—features historically resistant to measurement but increasingly framed as tractable under AI analysis. Human measurement thus becomes the precondition for human capital optimization.

JourneyFront, a vendor that infers candidate personality, values, satisfaction, and other internal qualities from pre-employment assessments and video interviews, illustrates how the ideal of continuous improvement underwrites this logic. Declaring itself the "World's Most Accurate Hiring Software," JourneyFront asserts: *"continuous improvement...if you can't measure it, you can't improve it."* Improvement is not a passive effect, but must be actively pursued through ongoing measurement, testing, and refinement. As the company explains, *"our process constantly tests, tracks, and makes changes that continuously improve your hiring process."* The imperative to measure, improve, and repeat is both strategic and normative. As Jive puts it succinctly: *"After all, if what you're doing isn't improving your results, why do it?"*

In short, emotion AI hiring services frame their offerings as necessary instruments for realizing data-driven, ever-optimizing organizational systems. The aspirational corporate ideals to which they appeal are cast as mandates—and by extension, the adoption of emotion AI both strategy and ethical obligation for responsible organizations.

4.4.1.2 Mechanism: Creating Affective Value and Affective Commodification

To advance in hiring processes mediated by emotion AI, candidates must possess what we refer to as *affective value:* the emotion-related data generated about job candidates that serves as a proxy for their worth to both the hiring platform and the organization. Emotion AI hiring services construct this value by automatically analyzing candidates' affective expressions and inferring traits such as engagement, enthusiasm, or emotional intelligence. Candidates whose affective profiles align with encoded expectations are rewarded by progressing through the hiring funnel, while those who do not are excluded.

For instance, JourneyFront promotes its *"auto-score"* feature, which ranks candidates based on inferred emotional and affective traits, enabling organizations to *"automatically filter qualified candidates"* and *"save time and know where to focus efforts."* Similarly, Jabri, a video interview provider that measures candidates' *"emotional and social aptitudes like interpersonal skills, communication skills, and personality traits"* invites organizations to use its *"game-changing analytics"* to *"discover their character,"* and *"review all critical personality skills important to [the] organization."* In such cases, vendors promise that if organizations *"measure what matters"*—

namely, the candidate's affective value—they can automate accurate hiring by advancing those deemed valuable and excluding those who are not.

This process depends on algorithmic assessments governed by opaque and proprietary rulesets which encode what count as desireable emotional expression and assign value accordingly. In doing so, emotion AI hiring services introduce hidden evaluative standards into the hiring process—standards that candidates cannot see, interpret, or contest. Affective value thus becomes a gatekeeping function: a threshold candidates must unknowingly satisfy to be considered for employment.

Once assigned, this affective value is commodified. Candidates are selected or rejected based not only on qualifications but on the emotional and psychological traits extracted and quantified by the system. Employment decisions—who is offered a job, who is deemed worthy of investment—are made, in part, on the basis of a candidate's fit with these algorithmically defined emotional profiles. In this way, emotion AI hiring services transform candidates' internal states into tradable assets—commodities in a labor market increasingly governed by affective metrics.

4.4.1.3 Core Value: Techno-omnipresence

The promise that emotion AI can solve the purported problem of hiring (in)accuracy rests on a claim to ubiquitous reach: the ability to be present and generate insights from anywhere, even the once private interiority of a candidate's emotional life, by analyzing vast data streams in ways that are framed as objective, intelligent, and unbiased. In doing so, emotion AI hiring services position themselves as essential and superior substitutes for human judgment, which is depicted as too inherently limited, fallible, and bound to reliably make accurate employment decisions.

Legitimized by corporate ideals of data-driven decision making and continuous improvement, this displacement of human discretion is framed as more than a competitive advantage—it is a moral imperative. Emotion AI "should" be implemented, vendors suggest, precisely because it offers the capacity to be present where humans cannot. This reflects what we term *techno-omnipresence:* the conviction that emotion AI can—and ought to—be everywhere, reaching into interior domains previously inaccessible while simultaneously replacing human presence in evaluative roles. Emotion AI's expansive scope is both functional and redemptive, cast as correcting the inherent subjectivity of human judgment.

QPage, an *"AI Mock Interview Machine,"* exemplifies these beliefs. The platform asserts: *"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."* Here, human discernment is portrayed as naive and insufficient, while emotion AI is cast as transcending cognitive limits *"beyond ordinary comprehension."* This rhetorical move reveals the deeper value animating emotion AI's deployment to hiring contexts: a belief in its divine-like superiority, made manifest in its presumed omnipresence.

By positioning themselves as the only viable path to hiring accuracy, emotion AI hiring services reinforce a techno-omnipresent logic that both legitimizes the removal of human decision-makers and naturalizes the mechanisms through which affect is quantified, commodified, and rendered governable in the hiring process.

4.4.2 Hiring (Mis)fit

Emotion AI hiring services claim that "fit" is achieved when there is alignment between a candidate and the organization across values, beliefs, character, and culture. A "misfit," conversely, is a candidate whose traits do not align. Misfits are framed as impediments to corporate efficiency—sources of waste, hassle, and dysfunction in the hiring process.

HRPuls, a pre-employment assessment provider offering automated psychometrics services, claims to detect both *"conscious and unconscious motives."* By *"identifying motivation and values through cultural fit analysis,"* HRPuls claims organizations can ensure that *"talent matches the company's values."* To achieve this alignment, organizations must first quantify candidates' internal traits, and emotion AI hiring services position themselves as the only viable means to do so, given their alleged ability to infer values, beliefs, and character from affect-laden data. For example, Equalture, a vendor offering *"neuroscientific gamification,"* claims its technology can *"hire the best-fits without bias."* This requires first "validating" company culture by analyzing current employees, and then assessing whether candidates objectively match that culture. Self-assessments by applicants or hiring teams, Equalture contends, *"will never be objective."* Here, automated fit assessments are not just useful but normatively imperative. Equalture promotes the "principle of hiring for culture fit" as the *"#1 rising star in recruitment,"* signaling its elevation to a hiring virtue.

This elevation is legitimized through the claim that hiring (mis)fits threaten organizational efficiency. Ducknowl, a video interview platform that uses AI to measure soft skills, promises more efficient hiring by filtering out candidates who look good on paper but *"won't fit well in an organization"* yielding *"quick and hassle-free hiring results."* HireOnboard similarly claims to *"eliminate applicant mischiefs"* by inferring candidates' cognitive and personality traits to automatically evaluate *"culture fit."* The implication is clear: hiring misfits waste resources—resources that emotion AI can preserve by eliminating the problem at its source.

These services promise that adopting emotion AI enables hiring managers to efficiently identify and recruit only *"fits,"* excluding misfits with minimal effort. Humantic, an *"AI with Emotional Intelligence,"* claims it can *"convert 30% more top candidates"* through bot-mediated assessments of candidate communication and interviews, promising *"data-driven candidate shortlists that take* *zero effort"* by judging *"culture fit without taking a test."* Similarly, Logi-Serve deploys interactive simulations to infer personality and aptitude in order to *"predict future performance"* and *"identify top performers."* These offerings frame hiring for fit not only as a solution to misfit-induced inefficiencies, but as a method of organizational optimization—delivering instant assessments that require little to no human resources from the organization itself.

Altogether, emotion AI hiring services present themselves as the only credible solution to the problem of hiring (mis)fits. By claiming to objectively assess the emotional and cultural alignment of job candidates, they promise organizations the ability to automate fit determination and thereby advance the imperative to hire only those who conform—efficiently, scalably, and invisibly.

4.4.2.1 Corporate Ideals: Precision and Loyalty

To solve the problem of hiring (mis)fits, emotion AI hiring services claim to offer precision at scale: the ability to algorithmically detect the affective attributes that determine a job candidate's fitness. HireOnboard, for instance, promises that its AI assessments will *"find the perfect fit for the job"* by measuring personality and cognitive ability, underscoring a promise of absolute precision. While the notion of "fit" is framed as value-neutral and objective, claims about what comprises such a fit (e.g., employee loyalty) reveal its alignment with corporate ideals.

HRPuls, for example, asserts that its AI-based psychometric and cultural fit assessments identify candidates whose *"motivation and values"* match those of the company, yielding *"higher productivity, satisfaction, and loyalty to the company."* Here, loyalty is no longer a variable to be observed but an outcome to be engineered—precisely measured and optimized through the emotional calibration of the workforce. JourneyFront similarly asserts, *"When a person is working on things they are interested in they are more engaged, work hard, and stay at jobs longer,"* concluding that *"measuring a person's interests is a must"* to achieve accurate job fit. Loyalty, by this narrative, is neither earned nor negotiated, but pre-measured, pre-selected, and pre-packaged.

KQ analytics makes this logic explicit, communicating to hiring organizations that its services allow them to remain *"focused on building a high-performance organization that lives and breathes your company's values."* The affective traits assessed through emotion AI hiring services are thus valued not just for their match to current organizational culture, but for their predictive loyalty to that culture. Fit, then, becomes a proxy for compliant, devoted labor—a means of selecting workers who will internalize and sustain organizational ideals without resistance or deviation.

4.4.2.2 Mechanism: Information and Psycho-biological Exclusion

To solve the purported problem of hiring (mis)fits, emotion AI hiring services identify "perfect" hiring fits by excluding (mis)fits. RecruitPack, a predictive hiring software that claims to read candidates' psychometric attributes and automatically rank them to *"pick 'A-player' candidates,"* illustrates this logic. Its system moves *"forward only those with desired attitudes and culture fit,"* promising to *"identify misfits in attitudes and values at the time of application."* Misfit exclusion is thus cast as both necessary for hiring fits and desirable in its own right, sparing organizations wasted time and resources. As RecruitPack bluntly states: *"you can eliminate [misfits] early and focus on the best candidates."*

We identify two distinct mechanisms through which exclusion operates: (1) information exclusion, whereby candidates are denied visibility into the inferences used to evaluate them, and (2) psycho-biological exclusion, whereby vendors frame fit as determined by immutable, biologically grounded attributes.

Information Exclusion. In conventional, human-based assessments of fit, evaluation is mutual: candidates and employers exchange information through dynamic interactions such as interviews, each assessing whether values, culture, and expectations align. Emotion AI hiring services replace this two-way process with an automated, one-sided assessment that excludes candidates from participation.

While some vendors advertise candidate "developmental insights" or "coaching tools," the most consequential inferences—the ones that determine rankings, shortlists, or automatic rejection—are not disclosed. Candidates therefore cannot contest or contextualize the judgments made about them. ZappyHire, for example, promises employers they can assess candidates *"even before [they] speak to them,"* offering robotic interviews and predictive scoring as tools to identify only those who *"matter."* Here, exclusion is not an unintended side effect but a marketed benefit: candidates need not participate in defining fit, because employers are furnished with unseen, proprietary evaluations.

The result is an intensified information asymmetry. Employers gain additional, hidden data points with which to judge candidates, while candidates lose both the opportunity to assess organizational fit themselves and the chance to challenge erroneous or biased inferences. Exclusion by design thus consolidates organizational power and further disadvantages jobseekers.

Psycho-biological Exclusion. A second mechanism frames fit as the manifestation of immutable, psycho-biological attributes. Vendors frequently use the language of genetics and evolution to imply that fit is innate rather than situational or socially negotiated.

HireMojo's *"Job Genome Project"* makes this explicit, drawing a direct analogy between hiring and genetic determinism. It claims that *"historical, analytic and prescriptive analytics combined with machine learning"* yield novel answers to workforce problems, reimagining job fit as biologically quantifiable. Similarly, Jive suggests that its sentiment tracking services improve *"culture* *fit while employees thrive naturally,"* implying that dispositional alignment is both innate and observable. HRPuls likewise claims its psychometric pre-employment assessments *"select talents that really fit"* by enabling the capacity to *"determine values and corporate cultural competence by means of complex algorithms, evolutionary processes and computer linguistics"*—a narrative that naturalizes fit as an evolutionary inevitability.

To bolster these claims, vendors appeal to scientific authority. Good&Co, for instance, promotes its *"Proprietary Psychometric Algorithm (PPA)"* as *"steeped in decades of research into career and individual differences literature on psycho-biological frameworks of personality"* and *"rooted in neuroscience and behavioral genetics."* While these services are not literally sequencing DNA, they strategically deploy biological metaphors and psychometric proxies to suggest immutability. This rhetorical move lends legitimacy to exclusionary judgments by casting them as grounded in science rather than preference.

Such appeals echo a controversial lineage of psychometrics long used to naturalize social hierarchies and legitimize racist, sexist, and classist discrimination [277, 278]. By invoking psycho-biological determinism, emotion AI hiring services recast exclusion as not only efficient but *natural*, embedding discriminatory logics in technical systems while insulating them with the aura of scientific rigor.

4.4.2.3 Core Value: Techno-omnipotence

The value underpinning emotion AI hiring services' response to the problem of hiring (mis)fit is what we term *techno-omnipotence*: the belief that emotion AI technology can—and should—possess the power to determine hiring "fits" and exclude "misfits." By design, these services remove candidates' agency to decide whether a job is a fit for themselves, replacing mutual evaluation with automated determinations. As Good&Co puts it, their *"intelligent, scientifically derived, and probability-driven algorithms [will] match jobseekers with the culture that's right for them."* In effect, the authority to judge fit is transferred wholesale to algorithmic systems.

This belief in techno-omnipotence is reinforced by vendor rhetoric that imbues emotion AI with quasi-divine power. "AI-powered" services routinely appeal to etchnical superiority to justify the surrender of evaluative hiring authority. Jabri, for instance, invites organizations to *"use the power of Jabri's digital video interview to discover their character."* Here, discovery is framed as revelation—an act beyond human discernment—positioning emotion AI as uniquely capable of exposing a candidate's character and determining their fitness for job opportunities.

For hiring organizations, the attraction lies not only in outsourcing evaluation but also in sharing in the power bestowed by these systems. Vendors promise that, by adopting their platforms, organizations themselves gain enhanced control over the workforce. Eightfold's *"talent intelligence"* platform promises clients the ability to harness *"deep-learning AI to help each person understand* *their career potential, and each enterprise understand the potential of their workforce."* RecruitPack similarly assures employers that its *"unique blend of psychometric tools"* enables them to *"quickly identify those with the can-do skills, will-do attitudes, and the fit-to characteristics for your role and your organisation,"* and to *"secure [the best applicants] before your competitors do."*

Such claims illustrate how the rhetoric of power serves a dual function: it legitimizes emotion AI's authority to define fit while promising organizations novel capacities for surveillance, control, and domination over labor. By surrendering to the supposed superior power of emotion AI, employers are told they will in turn wield greater power—able to "discover" candidates' inner character and "secure" loyal and compliant employees as means to gain or maintain competitive dominance. In this way, techno-omnipotence displaces human judgment to reconfigure the dynamics of organizational control, embedding authority in systems that promise mastery over the affective and biological dimensions of the workforce.

4.4.3 Hiring (In)authenticity

According to emotion AI hiring service claims, truth in hiring is achieved when organizations gain access to a candidate's interiority—when they can fully and deeply authenticate who a candidate really is. Vendors claim their systems extract deeper insights about an individual's disposition than human observation can alone, offering full visibility into the authentic self. QPage, for example, advertises automated psychometric assessments that *"verify"* the *"deeper truth"* about candidates. Similarly, Reejeig promises organizations *"full skills visibility"* through *"informed and accurate"* data-driven profiling to *"power talent planning."* In such cases, authenticity is equated with the completeness of algorithmically assembled profiles of a candidate's disposition, claimed to reveal the *"full truth"* of a person's inner life.

To establish authenticity as a hiring problem, vendors frame candidates as untrustworthy, inauthentic, and prone to deception. Equalture, a provider of *"neuro-scientific gamification"* assessments, asserts: *"of course intelligence isn't something you can fake; personality, however, is one of the easiest things to fake."* The company explains why candidates might engage in such fakery: *"it's indeed not smart to do, but you just want that job."* Here, candidate self-presentation is dismissed as "fake," justifying the use of emotion AI to penetrate surface impressions and uncover a candidate's "real" personality.

By adopting emotion AI hiring services, organizations are promised protection against deception and inauthenticity. Ducknowl, for example, claims its technology allows employers to *"find the genuine candidate"* and *"avoid bait-and-switch situations."* Idwall likewise promotes its *"face match technology"* as an automated solution to verify whether candidates are *"really who they say* *they are"*—an algorithmic verification that not only matches a person's face for biometric identity, but claims to read it to reveal the truth of who that person is beneath the surface—extending identity authentication from 1:1 person matching into the domain of one's interior disposition.

These claims position authenticity as a scarce and fragile quality that only emotion AI can reliably secure. By framing candidates as inherently deception as positioning themselves as arbiters of truth, emotion AI hiring services promise to solve the problem of inauthenticity by delivering employers quick, scientific, and supposedly objective access to the candidate's "real" inner self.

4.4.3.1 Corporate Ideals: Stability

Vendors appeal to the corporate ideal of stability to legitimize emotion AI as a solution to the purported problem of hiring (in)authenticity. By framing candidates as fake and untrustworthy, they position their technologies' insights into interiority as a way to mitigate uncertainty and enable stable, dependable hiring decisions.

FaceCode claims its self-described *"most intelligent coding interview platform"* allows employers to make *"truly informed hiring decisions thanks to automated interview summaries with AI-based behavioral insights."* Similarly, *"rich profiles with deep insights,"* promising 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."* Such claims promise organizations stability: by truly "understanding" candidates, organizations can secure hiring decisions in predictive, data-driven knowledge rather than uncertain human judgment.

Vendors tend to explicitly frame their services as a means to *"mitigate the risk"* associated with inauthentic hires. Retorio claims its video interview platform *"reveals hidden soft skills and traits,...measures personality, and predicts future potential,"* positioning measures of the worker's inner self as a basis for workforce stability. QPage dismisses conventional interviews as *"rarely predictive of success on the job,"* offering its AI *"Mock Interview Machine"* as a more reliable predictor of talent outcomes.

Across these examples, the appeal to stability rests on a simple equation: authentic candidates are predictable candidates, and predictable candidates produce stable organization. RecruitPack makes this logic explicit, promising adoptive organizations will avoid *"those 'bad hires' who look good at interviews but under-perform on the job."* Emotion AI hiring services thus present themselves as a bulwark against labor uncertainty, offering the assurance of organizational predictability and stability by claiming to reveal the *"whole truth"* of who candidates authentically are.

4.4.3.2 Mechanism: Extraction of Affective Value

The mechanism by which emotion AI hiring services claims to solve hiring (in)authenticity is *extraction*: drawing information about a candidates' interiority beyond what they choose to disclosure, and applying an affective valuation to assess whether candidates *"really are"* of value to the organization.

The information extracted is not impartial but oriented toward organizational utility. What emerges is a candidate's *affective value,* where their worth is measured by how well their inner traits—emotional and psychological traits, beliefs, values, and personality—align with the organization's goals. Reejig, for instance, promises that by using emotion AI to *"extract meaning"* from candidates, employers can *"create their workforce of the future,"* illustrating how extracted traits are positioned as novel resources for organizational strategy. Similarly, eLamp, which claims to infer *"critical"* and *"soft skills"* from *"any document,"* asserts: *"getting to know one's employees better enables [organizations] to make decisions that are anticipated and better adapted to operational demand."* Here, knowing candidates *better* is not about mutual understanding but about extracting organizationally relevant value to ensure stable and anticipatory decision-making about human resources.

Even vendors under public scrutiny employ this framing. HireVue, which ended its facial recognition-based emotion AI in 2021 following high-profile criticism [279], continues to generate inferences from speech and text. At the time of data collection, its website urged organizations to *"Go Beyond Resumes"* to reveal *"what really matters"* about candidates. By extracting unseen affective traits, HireVue promised organizations the ability to *"engage with the highest quality candidates first."*

Across these examples, emotion AI hiring services frame extraction as the key to truth and stability: by uncovering "what really matters," they claim to identify candidates whose affective value is highest and, therefore, most worthy of organizational investment.

4.4.3.3 Core Value: Techno-omniscience

The core value underpinning emotion AI hiring services' solution to hiring(in)authenticity is a belief in *techno-omniscience*: the belief that emotion AI possesses the capacity for all-knowing intelligence, uniquely capable of revealing who a person truly is, and that its superior knowledge ought to be used to secure authenticity in hiring.

This value rests on two linked assumptions: first, that a person's interior states and traits constitute their authentic, complete self; and second, that employers have a legitimate interest in accessing that authentic self to determine candidacy. In this framing, authenticity can only be gleaned by transgressing "beyond" what a candidate willingly shares. Instead, emotion AI claims the authority to penetrate beyond self-presentation, asserting a unique and supreme capacity to truly know the individual.

Adoreboard exemplifies this logic through its *"Emotics"* platform, which classifies *"over 24 emotions from any text"* and promises to solve the problem of hiring (in)authenticity *"by revealing the 'Unknown Unknowns' of...Employee Emotions"* to deliver actionable "business answers." Here, the implicit claim is that only emotion AI can transform hidden, otherwise inaccessible truths into usable knowledge—an epistemic authority unavailable to human judgment.

In this way, emotion AI hiring services' technosolution to (in)authenticity presupposes and promotes a belief in techno-omniscience, whereby AI does not merely assist human evaluation but transcends it, supremely capable of extracting and knowing the authentic self in full.

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."* [280].

Our findings demonstrate how emotion AI hiring services instantiate what Birhane terms the rationalist "God's eye view" paradigm: the belief that data science uniquely accesses objective and universal knowledge by isolating reason from human complexity, interdependence, and emotion [104]. Vendors discursively cast their technology in divine terms—omnipresent, omnipotent, and omniscient—while framing human interior states as isolable, measurable, and immutable, In doing so, they perpetuate a rationalist epistemology that treats knowledge of persons as static and objective [281]. This logic rationalizes algorithmic exclusion and dehumanization as the natural byproducts of objectivity, legitimating outcomes as productive social effects [280, 282, 283].

Unlike Cartesian traditions that sought to purge emotion from reason [284], emotion AI in hiring inverts this logic: vendors claim that only by accessing and quantifying dispositional qualities can organizations reach objective truth about the humans they employ. Emotion is refashioned as the missing piece to the rationalist puzzle—not the *obstacle* to objectivity, but its *key*. Whether this represents a genuine reconfiguration of rationalist epistemology or merely its expansion into new domains, the effect is the same: legitimizing AI's incursion into domains once regarded as irreducibly human.

As Birhane observes, the "God's eye view" shields practitioners from confronting ethics by treating technology as value-neutral. Our analysis shows how this shield extends to adopting organizations. By framing (in)accuracy, (mis)fit, and (in)authenticity as technical problems solved through affective extraction, commodification, and exclusion, emotion AI hiring vendors obscure how these systems function as disciplinary technologies that entrench inequality and exploit workers' affective lives.

4.5.1 Implications for Design and Policy

Our findings highlight how emotion AI hiring services unfairly constrain candidates' participation in hiring, exacerbate informational asymmetries, and commodify affective expression. We outline two avenues of intervention: (1) design for perceptible fairness and (2) enforcing fairness.

4.5.1.1 Design for Perceptible Fairness

Current systems systematically obscure what inferences are drawn about candidates, depriving them of the ability to contest or correct errors. Aligning with Fair Information Practice Principles (FIPPs) [285], vendors should: (1) grant candidates access to the information generated about them, and (2) provide opportunities for correction and deletion. While transparency alone cannot counteract deeper harms of affective commodification, such measures would make hiring processes more contestable [286] and allow candidates to better assess whether a job is fit for them.

4.5.2 Enforcing Fairness

Transparency requires enforcement to be meaningful. Under Section 5(a) of the FTC Act (15 USC §45), the Federal Trade Commission (FTC) prohibits unfair or deceptive acts in commerce [287]. In 2021, the FTC announced enforcement priorities targeting emerging technologies that reinforce asymmetries of power [288]. Our findings indicate that emotion AI hiring service claims may meet both the unfairness and deception thresholds under Section 5.

Unfairness. The FTC considers a practice unfair if it causes or is likely to cause substantial consumer injury that consumers cannot reasonably avoid and that is not outweighed by countervailing benefits [287].

Emotion AI hiring services' mechanisms of informational and psycho-biological exclusion likely meet this standard. First, the injury is substantial: candidates may be excluded from employment opportunities through proxy discrimination or misclassification, and subjected to economic and reputational harms based on unverifiable and error-prone affective predictions. By depriving candidates of the ability to negotiate their "emotional capital" or to mutually determine job fit [232, 289], control over the labor market is further consolidated in employers. Candidates deemed to lack *affective value* are excluded outright, while those selected are normatively molded into *"loyal"* fits that *“live and breathe”* company values. The emotional and dispositional dimensions of workers’ lives are exploited and subordinated to organizational ends.

Second, these harms are not reasonably avoidable. Jobseekers are rarely aware when emotion AI is in use, cannot access or contest the information generated about them, and cannot opt out without forgoing employment altogether.

Third, claimed benefits are unsubstantiated. Vendors frequently promise increased hiring accuracy, fit, and authenticity, yet our analysis shows these claims rest on vague assertions, pseudoscientific logics, and misleading appeals to objectivity—without evidence that these purported benefits outweigh the significant and unavoidable harms imposed on candidates.

Deception. The FTC defines deception as a misrepresentation or omission likely to mislead a reasonable consumer, where the misleading interpretation is reasonable and the information is material [287].

Our findings indicate emotion AI hiring services’ marketing practices satisfy this standard. Vendors routinely misrepresent their systems as unbiased and objective, suggesting that replacing human judgment with machine analysis eliminates discrimination. In reality, such systems may inherit and amplify algorithmic bias [257, 80, 290], yet marketing materials leave such discrimination largely unaddressed—obscuring corporate responsibility for ensuring fair hiring practices [291].

Further, vendors imply that their technology can accurately discern candidates’ internal emotional states and traits, despite extensive evidence that emotion recognition algorithms trained on others’ perceptions of external expressions tend to misalign with self-reported emotions, and lack external validity that warrants their use in high-stakes contexts such as hiring [78, 128]. Without substantiation, vendors reinforce deception by invoking pseudoscientific physiognomic logics—echoing biological determinism, personality genomics, eugenic rhetoric—to legitimize automated exclusion on the basis of inferred psycho-biological traits presumed immutable and predictive of job performance. These representations mislead hiring organizations into believing that their practices are scientifically grounded, materially shaping employment decisions with life-altering consequences for candidates.

4.6 Conclusion

This study interrogated claims made by 229 emotion AI hiring services to reveal how they construct emotion AI as a technosolution to three hiring problems: hiring (in)accuracy, (mis)fit, and (in)authenticity. We found that:

1. The mechanisms by which emotion AI services claim to solve these problems—extracting and commodifying affective value and automatically excluding candidates on the basis of inferred psycho-biological traits—unfairly intensify information asymmetries and deceptively exploit job candidates.

2. These solutions are legitimized by alignment with corporate ideals such as data-driven decision making, continuous improvement, precision, loyalty, and stability.

3. Vendors promote techno-omnipresence, techno-omnipotence, and techno-omniscience, casting emotion AI as a divine epistemic authority that employers should harness to resolve hiring challenges, enforce loyalty, and police authenticity.

By unpacking these claims, we show how emotion AI hiring services discipline workers' affective lives, disguise exploitation as objectivity, and perpetuate pseudoscientific hierarchies under the banner of fairness. Recognizing these practices as unfair and deceptive is a necessary step toward policy findings that constrain their harms.

More broadly, our findings demonstrate how AI in hiring reflects the ethics and values not of the people, but of the powerful. The fundamental normative question, then, is what values—and whose values—will govern the socio-technical futures we choose to create.

CHAPTER 5

Emotion AI at Work: Emotional Privacy, Surveillance, and Autonomy[1]

5.1 Introduction

Workplace surveillance is expanding to include automatic monitoring of worker emotion, mood, affect, and related constructs [292]. 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 [5, 293, 294, 6, 295, 296, 7, 297, 298]. By one industry estimate, 50% of U.S. employers will use emotion AI to monitor their employees' mental wellbeing by 2024 [5].

Commercially available emotion AI-enabled enterprise systems feature diverse capabilities. 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 [299]. 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 [300]. More obtrusive forms of emotion monitoring include wearables that use bio-sensors and physiological signals that

[1] This 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.

aim to infer employees' affective and emotional states in real-time, which may be implemented to influence worker behavior [301, 302]. Despite the increasing commercial availability and adoption of emotion AI in the workplace [299, 303, 5, 301], claims that emotion AI improves organizational outcomes [5, 293] are not scientifically well-established.

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 [304, 240, 241], and workers' exposure to and interaction with emotion AI-enabled workplace monitoring may occur regularly [292]. 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" [293]. 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 [305, 306, 104]. To this end, we conducted semi-structured interviews with U.S. 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. Workers perceive emotion AI to violate their *emotional privacy,* a term we introduce to describe privacy over one's emotions.

2. 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* [53].

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 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 not all emotion AI-enabled harms we identify in this work can be mitigated through either technical or policy solutions.

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 [244]. 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 [245, 246], 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 [246].

Though the emergence of these increasingly invasive surveillance practices were met with public concern over employee privacy [307, 308, 309, 310, 311], U.S. 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 [304]. 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 [311].

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 [130, 312, 313]. However, worker emotions are influenced by the workplace context [314], and emotion recognition technologies fail to adequately account for the contingency of emotions to the workplace context [315, 316]. 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 [316].

Given the expansion of workplace monitoring practices to include the detection and monitoring of workers' affective phenomena (i.e., emotion, mood, and core affect [317]), as well as the contextual sensitivity of these constructs to the workplace [318], 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 [319]. 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 [5, 293].

Workers' affective phenomena and organizational outcomes are mutually constitutive. As drivers of human behavior and decision-making [320, 321, 322], workers' emotions, moods, and affects influence organizational outcomes and events including sales [323], productivity [324, 325, 326], workplace violence [327], and insider threats [328]. 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 [5, 293, 294, 6, 295, 296, 7, 297, 298].

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 [130], enhance communication with audiences during online presentations [329], and develop post-meeting feedback systems to improve meeting effectiveness and inclusivity [330].

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 employers 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 [5, 293], 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 [331], and expose workers to unwanted monitoring and productivity management practices that risk employee privacy [332]. 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 [333]. 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 [334].

To ensure fair treatment of workers, emotion AI technologies should be used fairly and ethically [293]. Fair and ethical use of emotion AI may include commitments by actors deploying emotion AI systems that it is meaningfully consented to [335]; that its information is transparent and contestable [293, 336]; and that its use does not widen power asymmetries, such as those already present between workers and their employers [238, 336]. Yet, so far, the use of emotion AI in workplaces remains largely unconstrained and unregulated, and in the modern U.S. workplace, the growing adoption of emotion AI-enabled workplace surveillance is predicted to become the new norm [292]. 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

U.S. [292] 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 [337, 338], even if they are aware of the practice [304]. Indeed, U.S. workers are provided with insufficient privacy protection in the workplace [337], and are thus particularly vulnerable to privacy harms posed by ubiquitous workplace monitoring that has expanded to surveillance of workers' emotion and affect [292]. 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 [339], 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 [340]. 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 [340]. 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 [12]. 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) [41], 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 5.3.1), 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 U.S. both with *(n=6)* and without *(n=9)* cognizant experience subject to emotion AI in their workplace.

5.3.1 Study Design and 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 [341]. 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 [341, 342, 343]. 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 Pc and Pn, 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 (e.g., *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., [344]), 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 [345, 346], 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/Latino	customer service representative
Pc3	man	25-34	Latino/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	Latino	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" [347, 348], 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 [349, 350]. 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 [350]. 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 [350].

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 [351]. 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 [352]. General perceptions were strongly connected to privacy theories of contextual integrity [12], privacy regulation [339] and communication privacy management [340]. Perceived impacts codes were related to the sociological concept of emotional labor [53]. Perceived consequences codes closely resembled the typology of privacy harms introduced by Citron and Solove [41]; 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 and speculative methods are sometimes criticized for their findings' construct validity and generalizability to real-life experience. As described in Section 5.3.1, 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 [353], 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 [354]. In describing participants' boundary management processes [340, 339] 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 [53] 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 [12].

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 [355], 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 [339, 340] 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 [40] 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 [53], 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 [53] not only to comply with perceived expectations of their emotional expression, but also as an impression management strategy [356] 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 [53]. 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 [357]. 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. Pc1, 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, Pc1 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, Pn14, 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. Pn14 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 [41], 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 [41].

Psychological Harm. Psychological harms refer to negative mental responses experienced as a result of privacy violations [41]. 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 [41]. 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 inferences 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 [41]. 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 [358, 359]).

Economic Harm. Economic harms are the result of privacy violations that lead to monetary loss [41]. 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 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 [41]. 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 [360]. 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 [318]. 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 [80, 361], 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 [362].

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 [41]. 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 groups in 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" [41]. 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 disparately 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 [53] that are only partially recognized by the privacy harms taxonomy [41].

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 [53, 363]. 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 [41], 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" [53].

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 [364]. While emotion AI can amplify this discrimination harm [41], 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 [41] in that they may disproportionally 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 disproportionally 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 [41], 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 [360].

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 [41] 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 [5]. 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 [365, 150, 53] and amplifies the adverse consequences workers may experience from emotional labor enforcement and privacy intrusion. Even in the increasingly privacy-invasive modern workplace [150, 292], 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 Distinguishing 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 [366] and were followed by an impressive breadth of scholarship that largely focused upon emotional labor's adverse effects [367]. 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"' [53].

Hochschild depicts the interiority that remains deep inside workers as an "inner jewel" that evades the gaze of even the most authoritative employer [53]. 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 [368]. The breadth of scholarship aiming to improve the algorithmic detection of "fake" and "genuine" emotions [369, 370, 371, 372] 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 [41]. 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 [78, 128, 80, 290], 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* [41] and Richards' *intellectual privacy* [373], 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 [374]. 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.

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 [375, 150]. Yet, U.S. federal law does not currently limit or address the general surveillance of workers [375], barring public employees who enjoy constitutional privacy protection against their government employers [376]. As such, available legal avenues for workers regarding employer surveillance fall under a patchwork of state legislation and common law privacy torts [376], though both have proven woefully inadequate to protect against and remedy privacy harms workers endure in the workplace [376, 375, 377], and do not cover emotional privacy. Of note, the California Consumer Privacy Act (CCPA) mostly exempted employers from compliance under its "workforce data exemption" [378], though its successor as of 2023—the California Privacy Rights Act (CPRA)—extends protection to all personal information, including employee data [379], 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 [375, 380]. 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 [381], in effect normalizing extensive and invasive employee surveillance and silencing its legal objections [381, 375]. 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 [375], 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 U.S., most focus on remedying particular harms [337]. Exceptions include a few state statutes that limit the surveillance itself (i.e., video surveillance with audio [382]) and restrict the collection of certain types of employee data (i.e., biometric data [383]). 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 [384]. 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 [384]; 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," [385] particularly at the hands of anthropomorphized, emotionally intelligent AIs [386, 387, 388]. These open questions pose significant barriers to the application of enforceable regulatory frameworks to mitigate, prevent, and remedy potential harms from emotion AI [384, 389], a matter of increasingly pressing public concern [117, 390].

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 [384]. 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 [292]. 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 [391]. 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 [392], uses differential privacy, de-identification, and aggregation [392] to ensure identifiable data is visible only to the employee, while providing "privacy-protected" wellbeing-related insights to management [392, 393]. 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

[394].

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) [395] and ethical impact assessments [396]. Given the acceleration of privacy laws and regulation, prudent organizations that handle personal data will adopt data protection and privacy risk minimization standards [397]. 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 [215]. 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 [398] 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 [399, 400], and to that end we contribute *emotional privacy* as a lens to identify and address the harms posed by technologies that infer and interact with 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 [12, 13]. 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., [401, 402]) 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 Factors[1]

6.1 Introduction

Emotion AI deployments in both workplace and healthcare settings promise similar aims across contexts—improved safety, performance, and wellbeing [403, 296, 404, 405]. Alongside these transformative promises are substantial privacy and ethical trade-offs.

Understanding when and how privacy is transgressed is essential to grasping the social impacts of emerging technologies and mitigating their harms [41, 406]. 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 by context, purpose, 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.

[1] This 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).

Theoretically and empirically grounded in the understanding that privacy norms are interdependently bound by contextual variables [401, 407], vary by socio-demographic and individual privacy belief factors [408, 409, 410, 411, 412], and may differ between dominant (i.e., nationally representative) and minoritized perspectives [413], 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 these questions, 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 [12]. 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 [414, 415]—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 [413], 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 +2–5 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 [416]).

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 [417].

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 personal information generated by emotion AI 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

[413]. 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 −.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 [13]. When emotion AI purposes 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 [418, 334, 419] and healthcare [420, 421, 422, 423], 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 [424, 306]. Second, there is a conceptual gap: emotional privacy remains empirically under-theorized and thus difficult to operationalize [418], 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 [12]. 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 [12, 425]. 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 may be used for opaque or poorly justified ends.

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 [426, 427, 428, 429]. 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 [419, 241, 430, 431]. Speech data collected via continuous microphones, such as those in smart speakers, elicit fears of ambient surveillance and constant monitoring [432]. Text-based inputs, including monitored emails and messages, prompt concerns about intrusions into private communication and intent inference [433]. In emotion AI, such data are not shared directly but processed to infer emotional states and output task-relevant information. Under the theory of contextual integrity, these input modalities form part of the broader "sender" of emotional information [434].

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 [435]. Similarly, Zhang et al. showed that inferences about mental health from mobile data triggered privacy concerns that varied by data source and contextual framing [423]. 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 [436, 437, 316, 315, 362, 438, 439, 440]. 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 [409] or advancing a broader social good [441, 442, 435]. 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 [375]. Similarly, in healthcare, optimistic narratives about digital tools can obscure underlying privacy threats [443]. 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 [423].

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 [444, 445]. 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 [12]. 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 [409], race/ethnicity [409], and gender [410, 411]. Although the relationship between privacy perceptions and socio-demographic status remains understudied [408], research in privacy and HCI suggests that identity-based characteristics influence both one's exposure to surveillance and sensitivity to its harms [411, 446, 447].

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 [448]. Similarly, Bhatia and Breaux found that individuals with doctorate degrees reported lower concern about sharing personal information than those with lower educational attainment [409].

6.2.2.2 Race/Ethnicity

Research shows that Black and Latino populations are afforded less privacy in U.S. society, in part due to the long-standing normalization of racialized surveillance practices [449, 450]. This structural disparity may contribute to privacy resignation and a diminished perception of risk, despite disproportionately high levels of vulnerability to privacy intrusions [446]. 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 [80, 451, 452].

6.2.2.3 Gender

Gendered surveillance also shapes privacy experiences and concerns. Women, who face disproportionate exposure to gender-based harassment [453, 240] and workplace monitoring [241], consistently report heightened privacy concerns in these contexts [454, 455, 456, 457, 458]. 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 [459, 460], as well as a heightened reliance on safe, supportive, and affirming technology-mediated interactions [218, 461, 462].

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 [404], individuals with mental illness also face heightened risks from stigmatization, disability discrimination, and inaccurate inferences [405]. 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 [202, 420].

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 [463, 464, 465]. 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 [421]. 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 [423]. 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 [466]. These technologies often prioritize the extraction of emotionally legible signals, despite weak epistemological foundations, and risk medicalizing affective variance [467, 64]. 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 [437, 9].

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 6.3.3.

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 [412, 468, 469, 47].

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 [470, 471]. While early scales captured privacy concern as a generalized construct [412, 472, 473, 474], a growing body of work shows these concerns are highly sensitive to context [475, 476, 477]. Consequently, general privacy concerns have limited utility in explaining context-specific privacy preferences and outcomes [478]. Still, general concern measures like the Internet User's Information Privacy Concerns (IUIPC) scale remain widely used—both as controls in privacy perception studies [412] and as predictors of related constructs such as privacy decision-making [479] and expectations [402].

6.2.3.2 Perceived Risk

Closely related to privacy concern is perceived privacy risk [480, 409]. While sometimes conceptualized as a global construct [481, 482], perceived risk can also be framed in terms of specific potential harms [409]. 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 [483, 484]. 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 [485]. 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 [47, 486]. Trust in a specific institution can influence both general and context-specific privacy judgments [486], often by reducing the perceived risk of information misuse [487, 488]. At the same time, individuals' baseline privacy dispositions may precede and shape their levels of institutional trust [479, 486]. 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 [489], while Shen et al. showed that patients' willingness to share health information was similarly influenced by trust in healthcare organizations [421]. 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 [469, 490]. Sensitivity is closely linked to privacy risk: more sensitive data is seen as riskier and requiring stronger protections [412, 491]. Prior work suggests that emotional information and related data such as mental health status are widely regarded as sensitive, particularly in commercial and surveillance contexts [202, 492, 493, 362, 418]. 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 [435]. 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 6.3.2.

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 [12, 494, 495, 496], 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.

6.3.1 Theoretical Frameworks

Two theoretical frameworks for privacy underlie our study design: (1) Nissenbaum's *contextual integrity* [12], which defines privacy "as respecting the appropriate norms of information flow for a given context" [401]; 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 [413].

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) [401]. Together, these parameters "predict a complex dependency between privacy judgments on the one hand, and the values for all five parameters on the other" [407]. 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 [12, 409, 407], 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, 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 exclusion and oppression [497, 498, 499, 500, 497, 501], they propose that privacy theory and research move "beyond norms" to center

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	– 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 Emotional Privacy Norm—Fixed CI Parameters. Adapted from Martin and Nissenbaum, 2015 [401]. **Factorial Vignette Condition*

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 [413].

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 [502, 503]. 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 [504, 505]. 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.

6.3.2 Factorial Vignette 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 [506]. 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 [401]. 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 [507, 508, 509]. Certainly, how we measure privacy also has important societal implications, as public policy often relies upon conceptualizations of privacy as employed in research [401, 480] 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 [401].

6.3.2.1 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 [401, 413]. Grounded in contextual integrity theory (Section 6.3.1), 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 [510, 511, 512], while avoiding common limitations of ordinal (e.g., Likert-type) scales such as clustering and data loss [513, 514, 515, 516, 517, 518].

Although no consensus exists on the optimal dependent variable for privacy perceptions research [12, 494, 495, 496], the appropriate measure depends on the construct of interest. Studies focused on *behavioral privacy* often use willingness-to-use (e.g., [519, 409]). 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. [507].

Studies eliciting normative privacy judgments commonly use either participants' comfort levels (e.g., [520, 521]) or judgments of acceptability (e.g., [407]). 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 [409]. 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 [434]. 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.

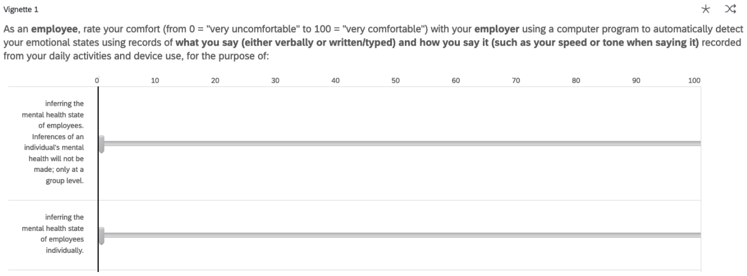
Participants saw only the text relevant to the given context *("as an employee..."* or *"as a patient...")* crossed with the assigned data input, followed by a separate VAS slider for each of the 14 purposes. This design minimized cognitive load and improved response efficiency. Although

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 Vignette Variables and Levels by Contextual Factor.
**CI Parameter: Data Subject*;
***CI 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 [522] and healthcare [523] 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



Vignette 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.

inferring the mental health state of employees individually.

Figure 6.1: Presentation of Vignettes. (Partial Example)

more specific mental health inferences at both individual and group levels [524, 404]; detecting or preventing self-harm or harm to others [362, 437, 438, 525]; and enabling early detection of mental and neurological illness with the goal of improving mental health support, safety monitoring, and research [437, 9, 315, 362, 525, 404].

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.

6.3.2.2 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.

6.3.2.3 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 [344, 526, 527, 528], 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 [529, 389]. These items adapted the Internet Users' Information Privacy Concerns (IUIPC) scale [412] to our specific contexts of employment and healthcare. Full item wording for the socio-demographic and privacy belief measures appears in Appendices C.4 and C.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.

6.3.2.4 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 [424]; 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.

6.3.3 Recruitment and Data Collection

6.3.3.1 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 6.3.1, this sampling strategy allowed us to investigate how privacy judgments vary both within and between socially dominant and minoritized perspectives.

6.3.3.2 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.

6.3.3.3 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 [530]. 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
	Latino/a	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 Descriptive Sample Statistics by Socio-demographic Level

6.3.4 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 who failed the attention check. We additionally removed one low quality (i.e., same answer for every question without justification in the open-ended questions) submission and 12 duplicate submissions. For those who had one incomplete and one complete submission, we preserved the complete submission and discarded the incomplete one; for those who 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 [531, 532].

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 C.4). Due to race/ethnicity mapping and value differences between our pre-screener and Prolific's categories described in Section 6.3.3.2, 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 Latino/a 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 U.S. reporting of Latino/a racial categories as white [533], we coded these participants' race/ethnicity as Latino/a. 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 Analysis 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' [301] and patients' [534] 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 [535]. 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 [536, 535].

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 [537], 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 [538, 537]. 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 [538].

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 it 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 [539], as it is generally inappropriate to use traditional p-values to assess the significance of fixed effects in mixed effect models [540].

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 [541]; the ICC for the representative and minoritized models was .72 and .67 respectively, indicating fair to good reliability [542]. For all models, we plotted the residuals to the quantiles of the standard normal distribution to confirm that the normality assumption was met [543]; although not all variables were normally distributed, the linear mixed effect analysis we employed is suitable for both normal and non-normal variables [544, 545].

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 6.3.2.2.

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) [273] 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 [273], 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 of Codebook Themes—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 of Codebook Themes—Healthcare Context

6.3.5 Limitations

6.3.5.1 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 [409]. 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 6.3.2.1). 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 [401].

6.3.5.2 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.

6.3.5.3 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 [546], 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.

6.3.5.4 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 [547, 538]. We interpret these conservatively and report confidence intervals to avoid dichotomous significance testing [547, 538]. Where appropriate, we discuss notable patterns that may have theoretical significance [547, 540].

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 healthcare 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 [413, 12], we highlight these variations to underscore the multi-dimensional nature of emotional privacy judgments. Our analytic framework 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—Employment Context

	Representative (n=300)	Minoritized (n=385)	Z-Test (comparison)
(Intercept)	36.64 (6.44)***	34.24 (6.08)***	0.27
Contextual Factors			
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)$^{\cdot}$	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)$^{\cdot}$	1.92 (0.87)*	-0.19
(12) auto. intervention – acute support	1.66 (0.93)$^{\cdot}$	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
Socio-demographic Factors			
Race/Ethnicity (baseline: white)			
Asian	-3.15 (4.31)	-8.05 (3.55)*	0.88
Black	5.64 (3.19)$^{\cdot}$	7.38 (2.79)**	-0.41
Latino/a	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
Privacy Beliefs			
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	

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$. **Bold** Z-scores indicate statistical significance between samples at $p < 0.05$.

Table 6.7: Regression Results—Employment Context

Regression Results—Healthcare Context

	Representative	Minoritized	Z-Test
	(n=300)	(n=385)	(comparison)
(Intercept)	28.91 (8.21)***	26.96 (6.72)***	0.18
Contextual Factors			
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)$^{\cdot}$	0.23 (0.95)	-1.57
Socio-demographic Factors			
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
Latino/a	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
Individual Privacy Beliefs			
general privacy concerns	-0.04 (0.08)	-0.08 (0.07)	0.35
emotion data sensitivity	-0.10 (0.05)$^{\cdot}$	-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	

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$. **Bold** Z-scores indicate statistical significance between samples at $p < 0.05$.

Table 6.8: Regression Results—Healthcare Context

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 the results, complemented visually by the coefficient plot in Figure 6.2.

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*

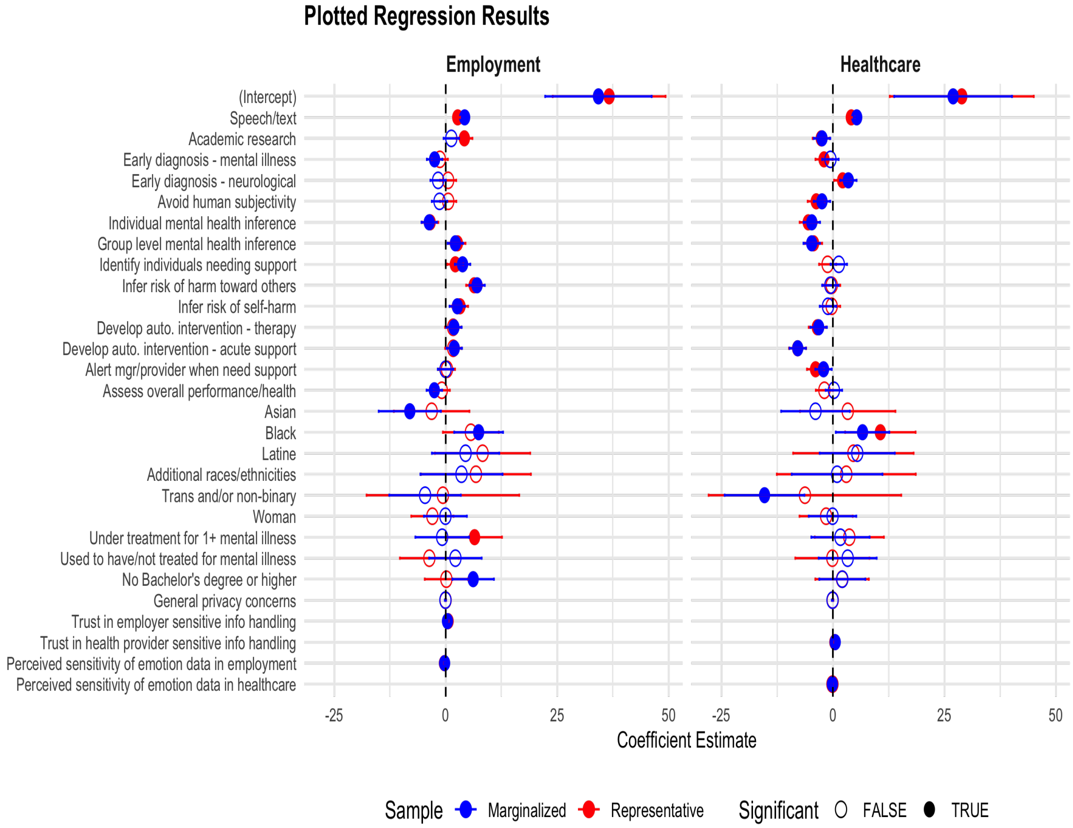


Figure 6.2: Coefficient Plot with Error Bars. Each point represents the tested independent variable; its position on the x-axis indicates the estimated effect size on reported comfort. Filled circles signify statistically significant relationships; open circles represent non-significant relationships; 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.

> *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).

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 qualitative findings confirm heightened concerns about emotion AI's potential to undermine autonomy, care access, and the patient-provider relationship.

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.

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.

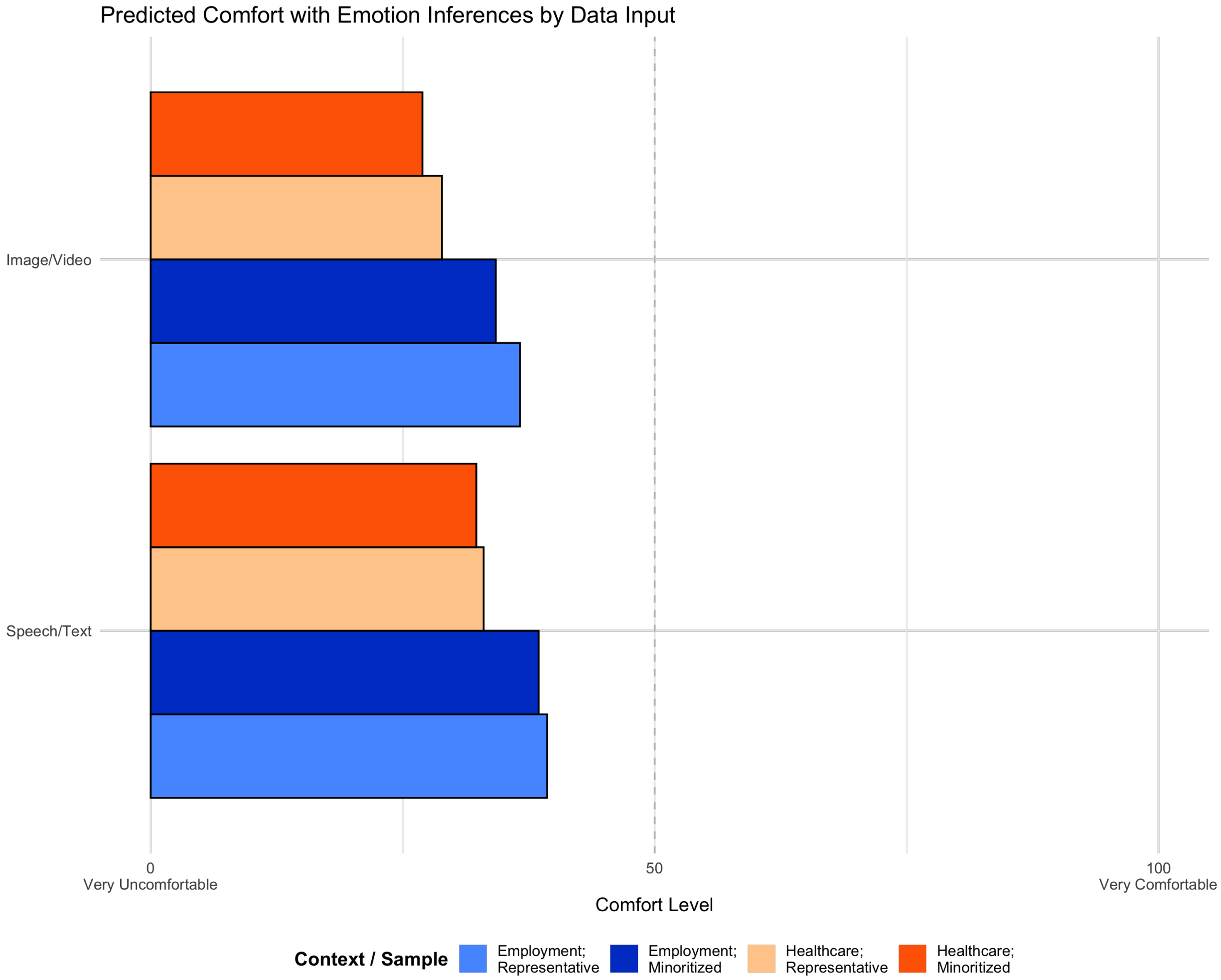


Figure 6.3: Predicted Comfort Levels by Data Input Type. This figure illustrates the predicted comfort levels by combining the data type variable coefficients to each mixed-effects regression model intercept, derived by analyzing respondent comfort on a scale from 1 (very uncomfortable) to 100 (very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

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 [419].

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 6.3). 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.

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 [409].

To contextualize these findings, we draw on qualitative analyses of perceived benefits and risks voiced by study participants, drawn from their open-ended responses—further explicated in Section 6.5.

Figure 6.4 visualizes predicted comfort levels for each purpose.

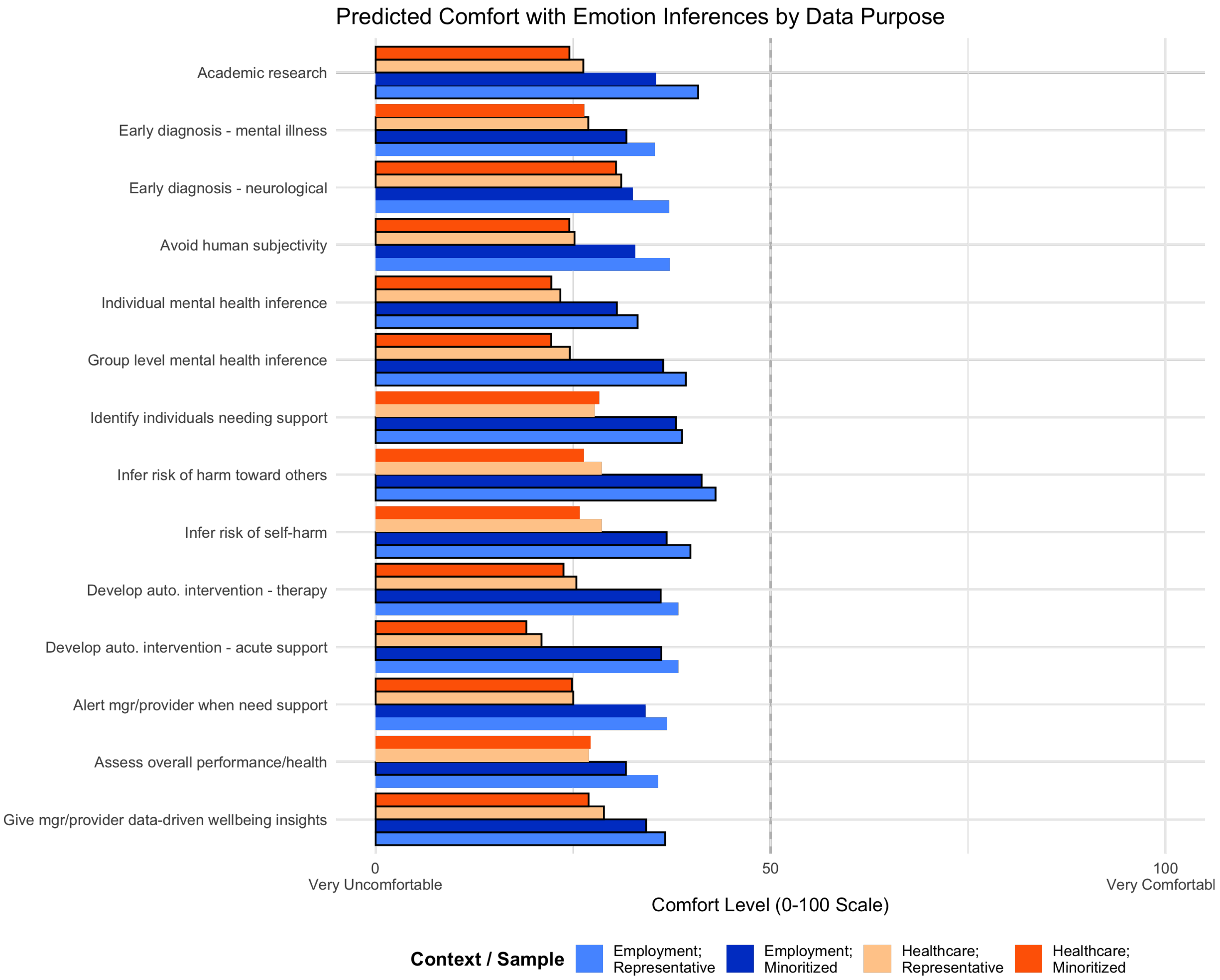


Figure 6.4: Predicted Comfort Levels by Purpose. This figure illustrates the predicted comfort levels by combining the purpose variable coefficients to each mixed-effects regression model intercept, derived by analyzing respondent comfort on a scale from 1 (very uncomfortable) to 100 (very comfortable).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

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 [303, 548, 549]. We examined how using emotion inferences to detect mental illnesses and neurological disorders earlier than otherwise possible influenced participants' 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 6.4, 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.

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. 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 6.2) 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.

Employee and Patient Assessments. Scholars and technologists have proposed automatic emotion inferences as a potentially objective method to reduce bias in both employee [550] and patient [142] 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 6.4).

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 suggest this discomfort may also reflect concerns among minoritized participants that emotional surveillance could impair work performance or lead to negative employment outcomes.

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.

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 6.4 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, qualitative responses 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. 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 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. 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 6.4, 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 qualitative results 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 [362]. 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 [551].

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 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.

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 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. 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 6.4.

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 qualitative results offer 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. 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.

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 6.4).

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. 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.

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 6.3.2.3 and justified in Section 6.2.2.

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 Latino/a 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 [80, 451, 452]—this group may attribute greater potential benefits to emotion inferences or perceive lower risk. Our qualitative analyses 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,

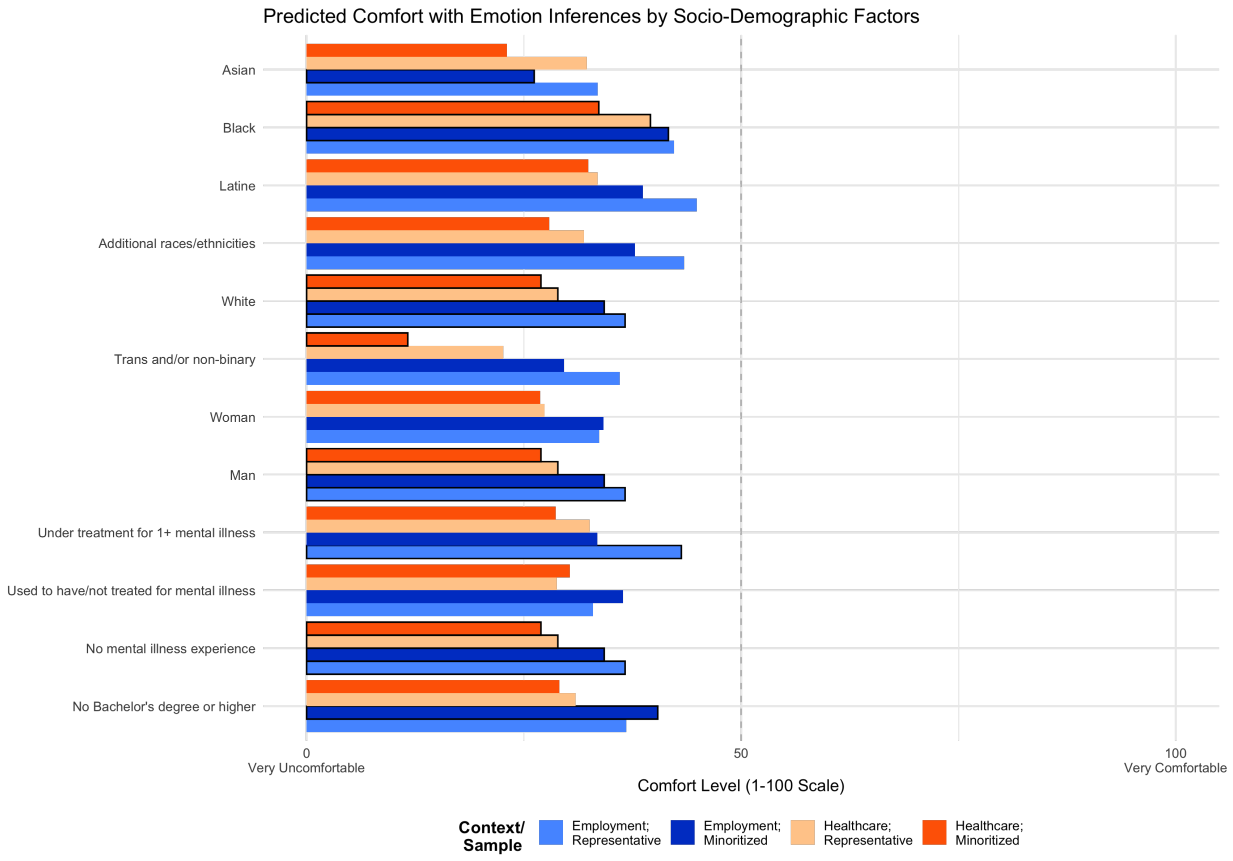


Figure 6.5: Predicted Comfort Levels by Socio-demographics. This figure illustrates the predicted comfort levels by combining the socio-demographic variable coefficients to each mixed-effects regression model intercept, derived by analyzing respondent comfort on a scale from 1 (very uncomfortable) to 100 (very comfortable). Bars with black borders indicate statistically significant results.

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 [241], 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 sociodemographic 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 6.2 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.

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.

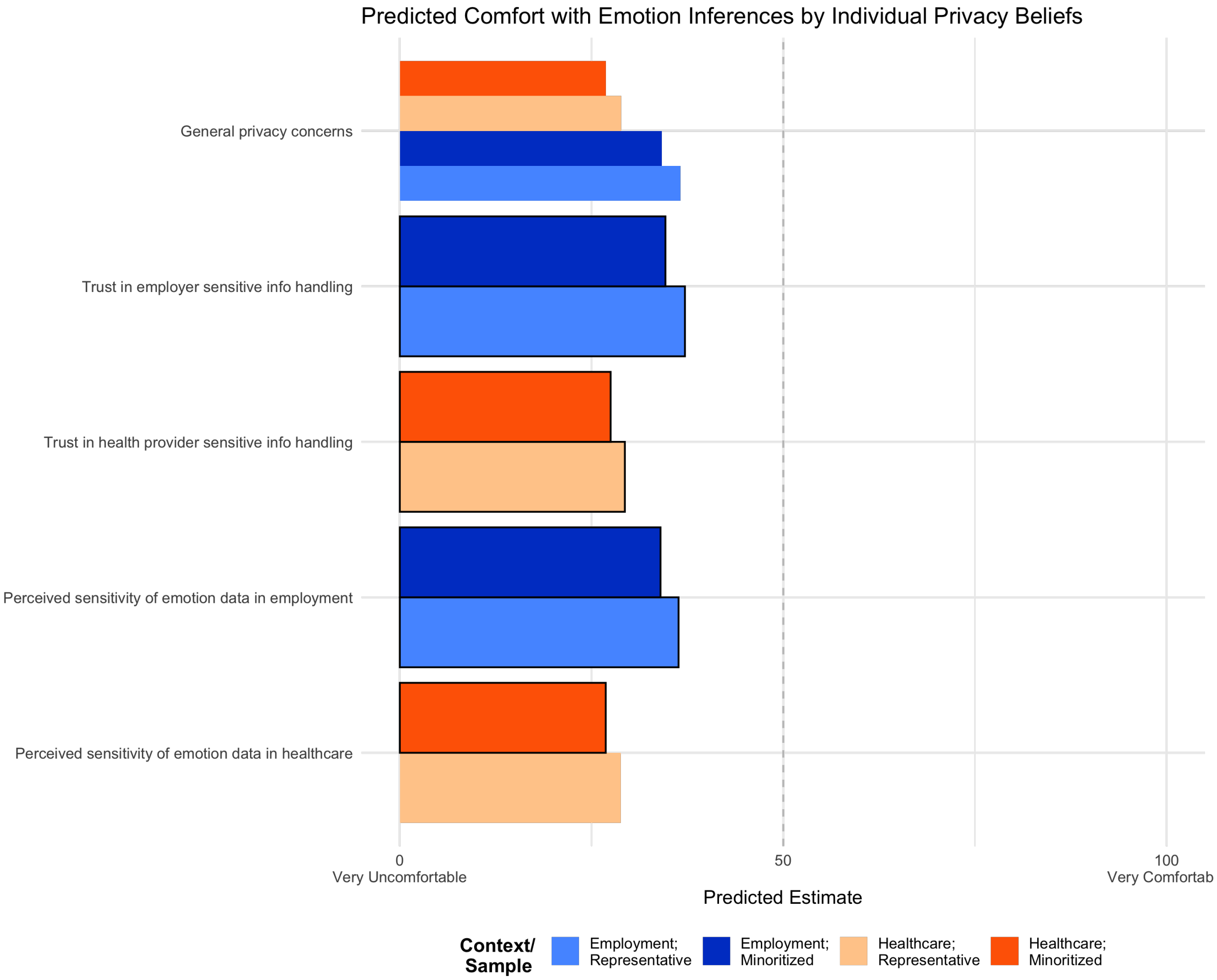


Figure 6.6: Predicted Comfort Levels by Individual Privacy Beliefs. This figure illustrates the predicted comfort levels by combining the individual privacy belief variable coefficients with each mixed-effects regression model intercept, derived by analyzing respondent comfort on a scale from 1 (very uncomfortable) to 100 (very comfortable). Bars with black borders indicate statistically significant results.

6.4.3.2 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 [529, 389]—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 6.7 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 6.8. 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 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.

In addition, our analysis examined whether and how participants' perceived sensitivity of emo-

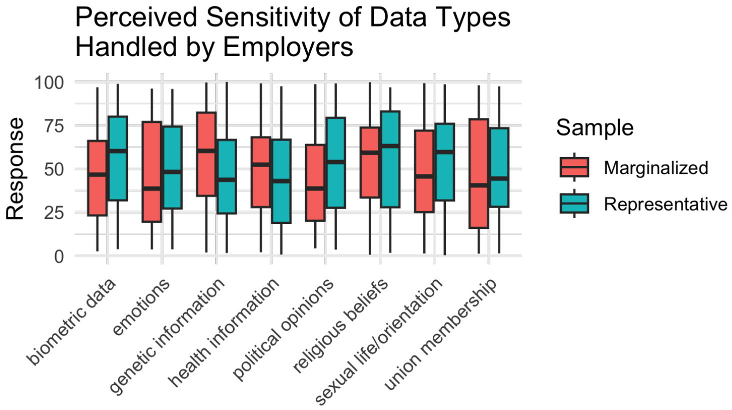


Figure 6.7: Perceived Data Sensitivity Comparisons—Employment Context. Box plots show the distribution of sensitivity ratings for each data type by sample, on a scale from 1 (not sensitive) to 100 (extremely sensitive).

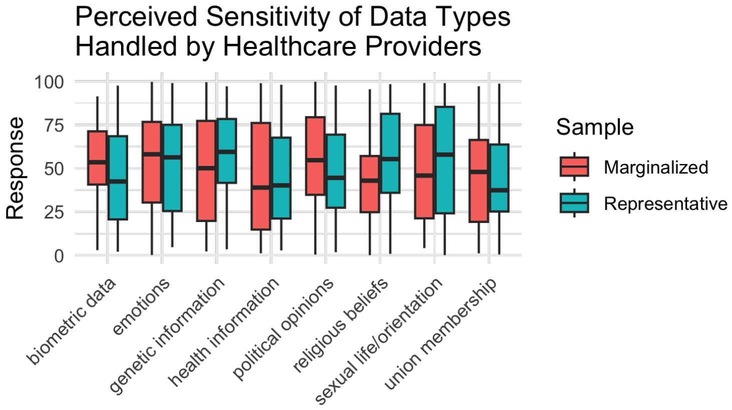


Figure 6.8: Perceived Data Sensitivity Comparisons—Healthcare Context. Box plots show the distribution of sensitivity ratings for each data type by sample, on a scale from 1 (not sensitive) to 100 (extremely sensitive).

tional 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, critically examining its implications and challenging assumptions that AI can resolve complex social problems without perpetuating existing patterns of injustice and discrimination [104]. This approach aligns with scholarship advocating that the voices of those subjected to technologies should participate in guiding their design, use, and regulation [552, 553]. Understanding data subjects' perceptions of emotion AI reveals whether its applications serve their interests in context, and highlights implications for policy, regulation, and decision-making.

The qualitative analysis applies this lens to examine how people perceive the integration of emotion AI in two contexts: the workplace and healthcare. After rating their comfort with various potentially beneficial applications in each setting, participants described in open-ended responses the benefits and risks they anticipated. While vignettes were framed to elicit possible positive outcomes, participants were not primed to consider risks, harms, or unintended consequences.

These findings highlight the importance of examining the contextual and consequences of emotion AI, particularly its entanglement with existing power dynamics. Centering the perspectives of workers and patients, the analysis suggests that deployment may reinforce the existing inequities and injustices in labor and healthcare it is deployed to remedy. The results underscore the need for normative frameworks that protect emotional privacy and address individual vulnerabilities, contributing to a more just and equitable integration of such systems by ensuring the agency and dignity of those subject to them.

6.5.1 Emotion AI in the Workplace

After answering survey vignettes about their comfort with various workplace applications of emotion AI, participants reflected in open-ended responses on the benefits and risks they anticipated.[2]

While some imagined potential advantages implied by the vignette purposes such as improved wellbeing support or bias reduction, the dominant sentiment was caution. Many feared that emotion AI would intensify existing workplace inequities, enabling intrusive surveillance, undermining autonomy, and amplifying bias. Such outcomes were seen as incompatible with workplaces that respect individual dignity and agency.

The following sections examine how participants anticipated emotion AI could shape the structural conditions of their work through (1) contextual impacts on safety, performance, and employment status and (2) the relational dynamics essential to preserving dignity and agency, including wellbeing, the experience of bias and stigma, and the degree of choice available to workers.

6.5.1.1 No Perceived Benefit

Roughly one-third of participants reported anticipating no benefits at all to emotion AI's application to the workplace. Even after responding to vignettes that suggested possible positive applications, these participants explicitly stated there was "none" or "no benefit" to them. Instead, they raised concerns that such systems could harm their wellbeing, erode the quality of their work environment, jeopardize employment, and create or amplify bias and stigma. Many framed these risks as incompatible with workplaces that respect individual dignity and agency.

Participants also expressed deep distrust in employers' use of emotion AI and anticipated responding either by conforming to its expectations or refusing participation altogether. As P103 explained, *"I don't [see a benefit]. Computers and such have not advanced enough to take the place of people in these things and every person's expressions and things are more subtle."* P86 similarly noted, *"I don't think anything could be beneficial from the intrusion of employers into employees' personal health."* Such perspectives reflect skepticism toward both the technological capability of emotion AI and the motives behind its workplace deployment. Even among those who acknowledged potential benefits, most went on to describe a range of risks they feared would compromise their autonomy and wellbeing.

[2] This 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.

6.5.1.2 Contextual Impacts

Workplace Safety. Several participants saw value in the ability to identify potential harm to oneself or others. P164 remarked, *"Personally, I think the most beneficial uses of these programs would be to detect potentially harmful/violent behavior,"* while P77 noted it could *"infer if employees could harm themselves and others."* P391 added, *"If it's able to see if a coworker has the potential to hurt themselves or another person I would want to know. It would prevent injury or in some cases death."* For these participants, emotion AI's capacity to recognize signs of danger was framed as a way to protect lives and, by extension, strengthen the workplace environment.

Performance Management. Some participants viewed emotion AI as a tool that could prompt constructive employer responses, particularly when employees were overworked or experiencing distress. P148 explained, *"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 wellbeing."* Similarly, P245 saw the potential for such systems to *"help [her] employer understand or at least be obligated to comply with offering support or lenience."* These perspectives suggest that, in theory, emotion AI could help secure accommodations or adjustments that respect employees' wellbeing, potentially benefiting both workers and employers. As P120 put it, using emotion AI to improve productivity could *"be a win-win."*

Work Status. Others worried that close monitoring would have the opposite effect, undermining performance and eroding morale. P339 argued, *"Monitoring employees' attention or work with close scrutiny has been proven to lower productivity."* P42 described a more personal impact: *"Lower self-esteem and bring embarrassment from being called out by an employer."* For these participants, the prospect of constant surveillance raised concerns not only about effectiveness but also about the precarity of their work status—preserving dignity, self-respect, and the ability to work without undue pressure or stigma.

Participants also expressed concern that emotion AI could create or intensify existing power imbalances between workers and employers, enabling high-stakes decisions with serious consequences. P10 worried it could *"give a little too much power or authority to the employers,"* while P115 feared it would give them *"more access to personal, private data on their employees,"* granting employers leverage that could be misused. P44 described a broader structural concern: *"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."* His remarks underscore the belief that even nominal safeguards could be ineffective, leaving workers vulnerable.

Some anticipated that such power could be used to justify unjust employment decisions. P245, who has a felony conviction, worried a system might *"block people by accidentally saying they're dangerous,"* reinforcing existing barriers to equal opportunity. P15 feared being fired or denied raises if emotion AI flagged her as *"not capable enough"* or *"not working enough."* P50 believed the technology could be used *"to fire employees struggling with mental health issues or to hold them back from receiving promotions or raises,"* or even to force unpaid leave. These accounts stand in stark contrast to earlier visions of supportive workload management, instead highlighting the risk that emotion AI could be used to legitimize exclusion, demotion, or termination.

Taken together, these perspectives show that while some workers saw potential for emotion AI to improve safety and support, many feared it would instead amplify existing inequities and weaken protections for fair treatment in the workplace.

6.5.1.3 Relational Impacts

Worker Wellbeing. Responding to vignettes portraying emotion AI as a tool for early health detection, identifying those in need of support, and providing data-driven insight into employees' wellbeing, some participants saw potential to improve both health and workplace experience—if employers used the technology in ways that genuinely demonstrated care. Anticipated benefits included detecting health conditions early, intervening before they worsened, and fostering greater understanding of mental health in the workplace.

P314 remarked, *"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."* Similarly, P27 suggested that earlier detection could *"lead to much more positive health outcomes and a higher quality of life for those nearing retirement age."* For some, early detection also offered a degree of self-determination; as P57 explained, *"If a system sees I'm having say... an OCD relapse before I do, I could take action earlier to stop it."*

Others emphasized the value of increasing both employer and self-awareness. P3 described potential benefits *"if employers are fair and understanding of their worker's mental health and the importance of providing accommodations when needed."* P230 saw value for personal coping strategies—*"I could find more ways to cope with my bipolar depression, while at home and work"*—while also imagining workplace-level gains in understanding: *"I feel like it's important for the workplace to understand mental health and also disabilities, and I feel like this is amazing and so beneficial."*

Yet even among those who imagined such benefits, concerns about harm were never far from the surface. Participants worried that employers could use emotion AI–generated data in ways that failed to demonstrate care, instead introducing risks that would damage wellbeing. Some feared inaccurate assessments leading to misdiagnoses or unjust assumptions. P36 said, *"I have a problem* *with the possibility of incorrectly assessing individuals. It would be a hell of a thing to get 5150'd into a psych ward just because a computer thought you needed it,"* highlighting the potential for such errors to harm both wellbeing and autonomy. Similarly, P84 warned that emotion AI *"could easily misdiagnose my condition or make it seem as if I had a poor work experience even if that is not the case,"* pointing to the risk of reputational and material consequences from flawed inferences.

Many saw emotion data as deeply personal and inappropriate for workplace monitoring. P96 described the risk of overreach: *"Gather sensitive information the employee wishes to be kept private, and they [emotion AI] just generally overstep boundaries."* For some, the very act of being analyzed would cause distress. P363 explained, *"The awareness that I am being analyzed would ironically have a negative effect on my mental health."* Several participants also challenged the relevance of such monitoring to the workplace at all. P62 noted, *"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."* P56 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."* Such remarks frame emotion AI as contextually inappropriate—violating norms about what information is relevant to share in a work setting—and undermining the boundaries necessary for maintaining dignity and agency in work contexts.

Bias and Stigma. Responding to scenarios in which emotion AI could be used to assess employees more objectively, some participants saw potential for these systems to promote fairness—if they were designed to address known limitations and account for individual differences. They imagined benefits such as reducing bias in decisions, countering human subjectivity, and lowering the stigma around mental health disclosure. Yet even among those who imagined such benefits, concerns about harm were never far from the surface, with many fearing that in practice, emotion AI would entrench existing prejudice, perpetuate stigma, and erode the conditions necessary for fair and respectful treatment at work.

Some participants believed carefully de-biased and technically reliable systems could help. P180 thought emotion AI could be beneficial *"assuming the software adjusts for those differences [across different identities] and still outputs correct and reliable data."* P228 said it might benefit them *"if it avoids employers' human biases in making judgments so as to be more objective,"* and P194 thought it could help *"in getting past the common problems of discrimination."* Others imagined a role in reducing the stigma surrounding mental health disclosure, allowing needs to be communicated indirectly. P28 valued being able *"to speak about my wellbeing not directly,"* while P311 noted it could *"allow access to care that normally has a negative stigma attached to it without having to put yourself [out there]."*

But participants more often described risks—particularly from inaccurate or misleading inferences. P87 warned that *“all current AI systems depend heavily on their training material...very difficult to provide a suitable large training set to cover the full gamut of human emotions.”* P94 worried that *“a system [might] see them make a weird facial expression”* and conclude they were suicidal. P328 pointed out that even accurate systems could cause distress if they surfaced emotional states at the wrong time, removing a worker’s control over when and how to address them.

Concerns about bias extended beyond inaccuracy to entrenched forms of discrimination. P95 cautioned against systems *“not programmed properly to consider race and culture,”* and described how her experience as a *“poor/black/elderly/woman”* already limited access to fair treatment. P7 questioned *“who is deciding what expressions ‘look violent”’* and warned of racial and gender bias *“particularly against POC, women, and trans individuals.”* P42 feared such systems could be used as *“evidence”* to justify unfair treatment of minorities or women by *“blaming it on mental instability.”*

Participants also linked these risks to mental health stigma. P358 stated, *“Mental health is stigmatized enough without allowing employers access to a computer program that thinks it can figure out mental health,”* while P234 warned, *“there’s a nasty stigma...you can be subjected to employee discrimination even though it’s against the law. I can’t afford to take that chance.”* Participants shared concern that emotion AI threatened to expose sensitive conditions in ways that could invite judgment, exclusion, or career harm.

Finally, some worried about overreliance, particularly when systems operated without meaningful human oversight. P103 warned it *“will be relied upon too much,”* P318 doubted that *“trusting employers to do the right thing”* was realistic, and P283 expressed that *“computers cannot do what a human is able to do.”* These concerns reflect a broader skepticism toward the premise that algorithmic inference can—or should—be trusted in workplaces as a means to improve worker wellbeing, especially when fair and equitable treatment is at stake.

Worker Autonomy. Participants described how they might respond if emotion AI were implemented in their workplace, with most imagining two broad reactions: altering their feelings or behavior to conform to the system’s expectations, or rejecting the technology altogether. These anticipated responses underscore the centrality of agency—whether retained, negotiated, or lost—when workers are subjected to emotion AI.

Some envisioned conforming as a strategy to avoid negative consequences, even at the cost of authenticity. P360 said it *“would cause me to act differently than I normally do at work,”* while P272 described feeling that *“you could not be yourself and roll your eyes at your supervisor or co-worker. . . you would have a constant feeling that Big Brother is watching.”* P185 anticipated they 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,”* highlighting the link between emotional labor and the preservation of privacy surfaced in Chapter 5. For P71, masking neurodivergent behaviors would require *"a massive amount of energy... which would make me very distracted and unproductive,"* with the added fear they might *"constantly be flagged by the software"* if unsuccessful. These accounts reflect concern that conformity to algorithmically encoded norms could erode both wellbeing and self-determination, especially for disabled or otherwise marginalized workers.

Others anticipated outright refusal, grounded in distrust of the technology's accuracy and its implications for privacy. P26 stated simply, *"I would not trust such a system,"* while P12 said, *"I do not trust a computer program to accurately and benevolently diagnose and/or treat mental health issues."* For some, refusal meant rejecting the conditions altogether. P124 asserted, *"I will never accept employment with any organization that uses them,"* P155 described it as *"an invasion of privacy that I would never agree to,"* and P292 said that its adoption would signal *"it is time to find a new employer."*

While refusal was imagined as a way to preserve autonomy, participants recognized that rejecting a job, quitting, or withholding consent is not an option equally available to all. Many would lack the power to act on their objections, especially if they were unaware that emotion AI was being used. This disparity in choice underscores the unequal distribution of agency in workplaces adopting such technologies and highlights how emotion AI's presence could constrain both the range and the reality of workers' self-determination.

6.5.2 Emotion AI in Healthcare

After rating their comfort with various healthcare applications of emotion AI, participants described in open-ended responses the benefits and risks they anticipated if such systems were integrated into healthcare. [3]

Many imagined ways in which emotion AI could enhance care—improving assessment accuracy, personalizing treatment, and facilitating the disclosure of sensitive information—particularly when human interaction alone might be insufficient. Others saw potential for the technology to create space for earlier intervention and expanded access to support.

Yet these possibilities were consistently tempered by concerns. Participants worried that emotion AI could misinterpret their emotional states, reinforce existing disparities in access and quality of care, or be used in ways that undermine trust in the clinician–patient relationship. Such concerns reflected a deeper unease about how emotion AI might shift control over self-presentation, diagnosis, and treatment away from patients—limiting their ability to direct their own care and constraining

[3] This 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.

their agency in managing their health.

The findings that follow consider both (1) contextual impacts—including how emotion AI could affect the accuracy and fairness of clinical decision-making, shape the prevention of harm, and alter access to and quality of care—and (2) relational impacts, such as its influence on patient voice, dignity, agency, and the experience of bias or inequity in healthcare interactions. Together, these accounts reveal how the integration of emotion AI into mental health services could either strengthen or undermine the conditions necessary for equitable, respectful, and patient-centered care.

6.5.2.1 Contextual Impacts

Assessments, Diagnoses, and Treatments. Responding to vignettes that described emotion AI in healthcare as a tool to assess mental health status, identify patients in need of wellbeing support, diagnose illness earlier than otherwise possible, and avoid human subjectivity in evaluation, some participants saw potential for improving mental healthcare. They imagined systems that could detect patterns human providers might overlook, facilitate earlier diagnoses, and help ensure that treatment decisions were based on more complete information.

P57 observed that *"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 echoed this, noting that *"sometimes doctors are busy writing notes or [are] distracted... the system could help detect things the doctor didn't notice."* These accounts linked the promise of emotion AI to the reality of overworked or inattentive providers, suggesting that additional insight could improve accuracy and responsiveness.

Others emphasized the potential to shorten long diagnostic delays. P7 reflected on a late ADHD diagnosis, saying that *"making resources so that people can get diagnoses and properly treated faster would have helped [them],"* while P33 imagined it would be *"nice for the program to notice a particular health concern before [she] did to facilitate faster treatment."* These comments align with broader ideals of early detection and timely intervention, though participants often framed these benefits as contingent on the system's ability to operate without reproducing existing errors or inequities.

Yet many also feared that emotion AI could worsen the very problems it aims to solve. Concerns centered on inaccurate inferences, misdiagnosis, and the erosion of patient voice. P321 warned that *"culturally, expression can vary depending on many factors, which might lead to inaccurate readings,"*reflecting doubt that emotion AI could account for complex variation in emotional expression across cultures and identities. Others noted that providers might accept flawed inferences at face value, sidelining patients' own perspectives. P81 explained, *"Sometimes a system could tell you that you are at risk of something when you are really operating at a safe level... Their own* *impressions should come first before being labeled."*

These accounts reveal how inaccurate or biased outputs could shape treatment decisions in ways that disempower patients, replacing dialogue with algorithmic judgment. For participants, the risk was not only faulty diagnoses but also the displacement of their lived experience from the center of care. Patient perspectives anticipate that integrating emotion AI into assessments, diagnoses, and treatment planning could undermine trust, diminish agency, and entrench the very inadequacies it purports to address.

Harm Mitigation. Responding to vignettes describing emotion AI as a tool to identify patients at risk of harming themselves or others and to alert providers for intervention, some participants saw potential benefits.

They imagined that, in specific situations, such monitoring could help avert crises or connect individuals to timely support. P322 thought it could work *"by possibly monitoring a dangerous person's social media, or offering links and hotlines when someone is in need of immediate support."* P279 acknowledged it might *"help severely mentally ill people who need monitoring to stay safe,"* though they also warned it could be *"way too invasive"* in most other contexts. These perspectives suggest a narrow conditional acceptance—potential value if the system truly improves safety without excessive intrusion.

Yet many participants feared that harm prevention efforts could misfire, with serious consequences. Several doubted emotion AI's ability to distinguish between genuine risk and ordinary expressions of emotion. P135 worried that *"I can be mad about something and the system may interpret that I will hurt someone... I could involuntarily be subjected to unnecessary help or even restraint."* P193 added that the system *"may not recognize"* genuine self-harm intent but *"could falsely flag someone,"* and that *"reporting this to health services could be detrimental to the patient."* For these participants, inaccurate inferences risked triggering harmful interventions—such as police involvement or involuntary commitment—that could erode trust, damage wellbeing, and disproportionately harm minoritized communities.

Participants also described how constant surveillance for risk could itself have adverse effects. P360 said the idea of such monitoring *"makes [them] uncomfortable,"* P373 anticipated it *"may cause... anxiety,"* and P282 believed it could *"lower self-esteem, frighten, put on the defensive, or otherwise make matters worse."* These reactions point to the paradox of a system intended to protect wellbeing instead producing fear, hypervigilance, or diminished self-worth.

Across these accounts, participants framed harm prevention as a high-stakes domain where the costs of error—false positives, missed detections, intrusive interventions—could outweigh the intended safety benefits. Underlying these concerns was a belief that the power to define and act on perceived risk should not be ceded entirely to automated systems, especially in matters so closely tied to individual dignity, autonomy, and the right to self-determination in care.

Quality Care Access. Responding to vignettes that described emotion AI in healthcare as a tool to assess patients' overall health, provide deeper insight into patients' mental health, and provide automatic interventions, some participants could imagine such systems enhancing treatment quality and expanding access to needed services—provided the technology was used to broaden access to quality care.

Participants envisioned emotion AI inferences could help providers identify gaps in care, tailor interventions, and better coordinate support across clinical teams. P33, for example, thought it *"could be nice for the program to notice a particular health concern before [they] did to facilitate faster treatment."* P7 reflected that better diagnostic tools *"would have helped [them]. . . so that people can get diagnoses and properly treated faster."* For these participants, the technology's value depended on it being used to open pathways to care rather than to restrict them.

However, many feared that granting providers, insurers, or other parties access to emotion AI–generated inferences could lead to harmful misuse, ultimately reducing access to quality mental healthcare. P368 warned that such data *"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."* P50 anticipated *"the quality of care [for their multiple neurological and mental health conditions] would decrease dramatically if this technology was put in place by health providers to cut costs."* These perspectives reflect concerns that emotion AI could function as another triage mechanism for rationing care, entrenching disparities rather than alleviating them.

Participants also raised the risk that emotion AI could diminish the human element of mental healthcare by reducing or replacing direct provider–patient interaction. P8 cautioned against *"removing much more of the human from human medicine,"* arguing that such changes might benefit only those who create and profit from the technology. P152 warned that *"relying too heavily on computer-assisted programs can lead to poor healthcare"* by tempting providers to *"step too far back from the process."* For some, this concern was tied to specific technologies like chatbots. P242 feared that *"chatting with a chatbot for mental health support. . . could cause me to feel isolated, invisible and lead to depression or self-harm."* Similarly, P50 emphasized that *"psychological healing also takes place primarily within human relationships, not AI chatbots,"* stressing that human beings *"are better at reading one another than a computer can ever be."* These comments highlight a belief that mental healthcare quality depends on interpersonal connection—something participants saw as impossible to replicate in fully automated interactions.

Others worried about more direct barriers to treatment. P117 believed emotion AI *"could be misused to limit access to certain treatments or services,"* while P51 suspected it *"could be misused by health insurance companies to decrease client support and increase costs for clients."* For some, these risks were embedded in a broader critique of healthcare's profit-driven incentives. P159 put it bluntly: *"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."*

Across these accounts, participants expressed deep skepticism that emotion AI in healthcare would be implemented in ways that genuinely improve patient outcomes. Instead, they anticipated that its outputs could be weaponized to justify cost-cutting, exclusion, or targeted marketing while eroding the human relationships, trust, and interpersonal care essential to patient dignity, agency, and wellbeing.

6.5.2.2 Relational Impacts

Patient Voice. Vignettes described emotion AI in mental healthcare as a means to identify moments when patients might need emotional support, respond through an intelligent computer program, or even conduct therapy. For some, this raised the possibility of improving one of the most persistent shortcomings of mental healthcare: the loss or dismissal of patient voices in clinical interactions. Others, however, feared that rather than amplifying their perspectives, emotion AI would further marginalize them by replacing patient self-reports with automated inference as a primary basis for care decision-making.

Some participants saw a role for emotion AI in legitimating patient disclosures and facilitating more open communication with providers. P242 believed such a system *"could be beneficial if they actually provide emotional support. . . 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."* For patients who had experienced being ignored, gaslit, or disbelieved—especially those from minoritized communities—emotion AI's promise lay in *"backing up"* their accounts, giving weight to self-reported symptoms that might otherwise be discounted.

Yet participants more often worried that emotion AI could weaken their voice in care. P113 warned it would harm her ‘*‘if doctors place complete confidence in software and discount the information [she] may tell them if it doesn't support [the] software."* P87 foresaw *"a danger. . . if such systems become widespread, it will become very difficult to refute their diagnoses."* In these accounts, emotion AI was not a tool for listening but a mechanism for replacing human judgment with machine authority—making patients' own narratives harder to assert or defend.

Overall, participants underscored that their agency in mental healthcare depends on being heard, believed, and actively involved in decisions about their own care. Emotion AI was seen as capable of either reinforcing or eroding that agency: it could create new openings for disclosure when used in genuinely supportive ways, or it could entrench patterns of dismissal and disempowerment when its inferences are privileged over patient accounts.

Provider Bias. Vignettes described emotion AI in healthcare as a way to supplement clinical judgment with more "objective" insights, potentially reducing the influence of subjective provider judgment in mental health assessments, diagnoses, and treatments.

Some participants saw this as a meaningful opportunity to counter the prejudice, stereotyping, and inattentiveness they had experienced in traditional care. P334 reflected, *"I'm an adult ADHD person and would have benefited GREATLY from technology such as this... the path to my diagnosis was arduous and oftentimes hindered by non-objective professionals."* For P95, whose past encounters with human doctors had been marred by bias, *"a program that's able to... give an unbiased evaluation, couldn't be any worse."* These views framed emotion AI as a potential corrective—if it could in practice produce accurate, unbiased inferences and be used to inform, rather than replace, human care.

More often, however, participants anticipated that emotion AI could amplify rather than mitigate bias. They worried that providers might accept its outputs uncritically, using them to legitimate their own prejudices or act on the system's embedded stereotypes. P331 warned it *"could be biased or based on stereotypes that could lead to incorrect information and harm by falsely associating traits with someone."* P7, concerned about race and gender bias, described how such systems might label people as *"unhealthy looking"* for failing to match normative appearance or behavior, with specific risks for disabled people, non-native speakers, and those whose cultural or gender expression diverged from the system's expectations.

Some underscored that algorithmic bias compounds existing inequities in care. P306 cautioned, *"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."* For them, emotion AI's inability to account for multiple, intersecting identities—and the opacity of how such systems are built—posed a risk of deepening disparities under the guise of technical objectivity. Others noted that human oversight was no safeguard if the providers themselves held biases or lacked the skill to recognize flawed outputs. As P335 put it, *"It depends on the provider and their potential biases... at the end of the day."*

Participants largely rejected the idea that automation alone could solve the problem of bias in mental healthcare. While some imagined emotion AI as a way to counter subjectivity, far more feared it would entrench or magnify the very inequities it promised to address—undermining trust, fairness, and the conditions for dignified, patient-centered care.

Data Access. Vignettes presented emotion AI in mental healthcare as a way to deepen providers' and researchers' understanding of mental health through data-driven insights, with potential to improve care and advance research. Some participants welcomed this prospect, describing it as an opportunity to address the persistent problem of providers lacking a full grasp of patients' conditions. P253 believed it *"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."* P285 similarly saw value if *"the researcher could help practitioner better understand some aspect of their practice."* These views framed emotion AI as a supplemental tool that—if implemented with care—could enhance understanding of mental health and inform better treatment.

Alongside these possibilities, participants voiced deep concerns about how emotion AI inferences could be accessed and used by actors beyond the clinical relationship, particularly insurance companies and third parties, in ways that could restrict care, raise costs, or exploit sensitive data. P368 warned that it *"could cause healthcare providers to create biases about their clients and even drop them from their system altogether,"* while P50 feared it could be used *"to cut costs"* at the expense of quality care. P159 doubted the technology's deployment would be motivated by patient benefit at all, seeing it instead as *"another profit center. . . without our knowledge or consent."* Such accounts reveal how perceived profit motives and structural incentives in healthcare shaped skepticism toward emotion AI, even when participants could imagine potential benefits.

Privacy was a central thread in these concerns. P92 explained, *"Sometimes we don't want to reveal things about ourselves. This would make me feel very vulnerable and exposed,"* highlighting fears of losing control over whether and how emotional states are disclosed. Others questioned basic governance: P230 asked, *"How will the data be held? Will it be deleted afterward? If sent in for research, how many others will witness my data?"* These questions reflected a broader unease with opaque and potentially unregulated data flows. Participants also worried about compelled disclosure or leakage: P149 wondered if readings could *"ever be turned over to, or subpoenaed, by law enforcement,"* while P265 feared recordings *"could be used for facial recognition beyond the supposed purpose"* and *"once it gets out into the open you are at a loss."*

Several participants anticipated that such risks could extend beyond the patient to others in their environment. As P149 noted, *"There's also the bystander issue. . . how would such audio or video recording ensure that other people's privacy in my residence was protected?"* This underscored the possibility of *"collateral"* privacy violations falling outside traditional safeguards such as HIPAA.

Across these accounts, participants recognized that emotion AI could, in theory, strengthen mental healthcare by expanding knowledge and improving provider insight. Yet they overwhelmingly feared that the same data could be used to deny services, increase costs, or leak beyond its intended context—violating privacy, undermining trust, and eroding patient autonomy. For many, the risks of losing control over sensitive emotional information outweighed the promised gains, especially in a high-stakes domain where confidentiality and trust are foundational to care.

6.6 Discussion

This study examined *emotional privacy* judgments—how people evaluate the appropriateness of inferred emotional information flows—using contextual integrity (CI) [12] as both an empirical measure and a normative lens. CI holds that privacy is preserved when information flows align with entrenched social expectations about the roles, purposes, and constraints of a given context, and is violated when these expectations are breached. In domains like employment and healthcare, these contextual norms are tightly bound to the very conditions that enable human dignity and agency. Work and care are not simply transactional arenas; they are capabilities-bearing contexts in which fair opportunity, health, and self-determination are constituted. As the findings underscore, flows of emotional information that reinforce these contextual ends may enhance dignity and agency; flows that undermine them may erode the same.

By fixing the information type parameter to emotion inference and varying the purpose of use and data source across factorial vignette scenarios, the study empirically tested CI's central claim: that appropriateness judgments hinge not just on what information is collected, but why and for whom it is used. The quantitative results support CI's normative heuristic—participants judged emotion AI use as more appropriate when its purpose clearly advanced the core goals of the context (e.g., preventing imminent harm in healthcare, ensuring workplace safety), and less appropriate when purposes were misaligned (e.g., individual profiling in hiring, insurance risk scoring). These patterns held across both nationally representative and purposive samples, but with important divergences: participants from minoritized groups were more likely to perceive both greater benefits and greater risks, underscoring that normative privacy judgments can diverge even among individuals occupying the same contextual role.

The qualitative findings sharpen this picture. While participants could imagine benefits—earlier detection of health conditions, targeted support for people in distress, reduced stigma around mental health—they more often described these benefits as fragile, contingent on trust, and easily outweighed by harms. Across both employment and healthcare contexts, participants expressed concern that emotion AI would expand institutional surveillance in ways that reduced their ability to control how they are seen, interpreted, and treated. Many feared that the very institutions charged with enabling their flourishing would instead use emotional inferences to intensify power asymmetries, limit opportunities, or pre-emptively constrain their choices. In short, the same contexts that give structure and meaning to our lives—through dignified work and equitable care—were seen as vulnerable to being reshaped in ways that diminish the agency and respect they are meant to confer.

Together, these results illuminate both where participants draw the normative line on emotion AI and why—revealing contextual and identity-based vulnerabilities that shape emotional privacy judgments, yet converge on a shared expectation: to be treated as a dignified human being with inherent worth. These vulnerabilities tend to be most acute for minoritized groups, whose needs and concerns often diverge from those of socially dominant groups. Relying solely on socially dominant "internal standards of justice" [554] to define appropriate flows risks reinforcing systemic injustices and silencing dissenting perspectives [12, 413, 555, 425]. Aligning the design, deployment, and governance of emotion AI with the unmet needs and persistent concerns of those most vulnerable to technological impact strengthens protections for everyone.

Building on these insights, the sections that follow highlight implications for anticipatory governance, purpose binding, protections for emotional information, and designing for contextual vulnerability and emotional privacy.

6.6.1 Anticipatory Governance

Our results have significant policy relevance. Notably, the patterns in our study mirror many elements of the EU AI Act and its clarified application guidelines [556]. We observe striking alignment between public intuitions and the regulatory architecture: areas that are regulated as high or unacceptable risk in the AI Act also emerged as scenarios for which participants indicated higher discomfort. 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 drew in our survey closely matches the EU's regulatory reasoning, with privacy judgments surfaced in our study providing regulatory legitimacy to the Act.

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.

6.6.2 Bounding Inference Purpose

Contextual integrity does not formally model purpose—treating purpose as implicitly constrained within a context's goals and optionally encoded as a transmission principle rather than a standalone parameter. Our results however confirm that purpose is a decisive driver of emotional privacy judgments—often operating as the hinge between perceived acceptability and unacceptability.

In the workplace, performance-scoring inferences—a common managerial practice designed to boost productivity [430, 292, 375]—elicited *lower* comfort, reflecting skepticism toward uses tied to surveillance and evaluation. Yet sharing workers' emotion inferences with academic researchers—an extraneous purpose that does not obviously advance workplace goals—*raised* comfort levels, perhaps because it felt less threatening to workers' agency or day-to-day autonomy.

A similar pattern surfaced in healthcare. Automating interventions or diagnosing mental illness—purposes tightly coupled to clinical objectives—*lowered* comfort—as the qualitative findings emphasize, such uses are perceived to override patient-initiated disclosure and undermine interpretive agency. By contrast, neurological disorder screening, another clinical use, *increased* comfort—underscoring that even within the same domain, purposes vary in how they are normatively received.

These divergences cannot be explained by contextual integrity's five canonical parameters or by contextual goals alone. Instead, the explicit *purpose* of the inference—interacting with the type of information, actors involved, and the transmission principles governing the flow—emerged as central to participants' emotional privacy judgments. As Nissenbaum has noted, purpose's salience in privacy evaluation has grown alongside evolving technologies and data practices [557], and recognizing purpose as a constitutive contextual parameter may be a "necessary antidote" [558]. These findings support extending contextual integrity to formally incorporate purpose, enabling more precise governance through purpose limitation and inference minimization rules.

Drawing on the quantitative patterns and complementary qualitative results from this study, we propose a narrowly scoped set of permissible purposes for emotion inference, paired with prohibitions on purposes likely to undermine dignity, agency, or trust:

- **Purpose binding:** Mirroring the EU AI Act's risk-based, context-specific approach [556], 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, allowing only narrowly justified, socially valuable uses of emotion AI while protecting emotional privacy and reinforcing the social ends these contexts are meant to serve.

6.6.3 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 [384].

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, 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 [412, 487, 401].

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 [559] 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 [419]. Current U.S. regulation typically limits protections to biometric identification [560]. 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.

6.6.4 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 [416] and preserving patient autonomy and dignity in healthcare [417]. 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 [561].

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 [562]—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 [556].
- **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.

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

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

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 Emotion Inference Purpose Level Groupings. We grouped results into higher level themes to aid interpretation.

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 [71, 68, 84], 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 [1]. 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 [558]. 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 [558]. 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 [13]. Functioning as teleological benchmarks, these contextual purposes give the context its normative structure and social legitimacy [558].

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" [558]. 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 [557, 558]. 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" [558]. 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 [15]. 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 [517], solitude [563], control [564], boundary management [565], or contextual appropriateness [12], 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 [12, 425], 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 sociotechnical 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 [12, 558]. 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 [558]. 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 [12, 13, 558, 14].

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 [15]. 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 [566, 50], the private and the public [567, 568, 563], and the internal and external domains of personhood [565, 564, 569, 570, 571, 572]. 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* [569, 570, 573, 564]: 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 [564], personal well-being and development [574], and civil society [575]. Intrinsic accounts establish privacy as a first-order moral and political value: one that justifies constraints on surveillance, exposure, and coercive conformity [576, 577, 569], even in the absence of measurable harm [578]. 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" [577]. 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 [573, 373, 453]. 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* [577]. 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 [579]. 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 [579]. 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 [580]. 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 [579, 581].

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 [579]. Poets like Wordsworth cast solitude as necessary for preserving the inner self against the overstimulation and moral conformity of modern life [580]. 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 [582].

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 [580]. 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 [579], 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" [583].

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 [246, 373, 584, 585, 586]), 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 [579]. That this myth endures reveals how easily privacy's spiritual and ethical dimensions can be reduced to questions of taste, sensitivity, or celebrity control [575]—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 [587].

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" [588]. Yet the ensuing period of "vigorous growth and experimentation" in privacy tort law [586] was eventually blunted by a narrowing of scope [579]: 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 [579]. 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 [589].

The decisive shift came in 1960, when Prosser, aiming to formalize and stabilize the emergent tort doctrine [588], codified four privacy torts: intrusion upon seclusion, public disclosure of private facts, false light, and appropriation of name or likeness [590]. 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 [588]. This fit well with Holmesian jurisprudence, which rejected law's moral aspirations in favor of optimizing social behavior [591].

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 [586]. 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 [588]. Prosser's categories made no reference to the *inviolate personality*; they offered courts actionable compensable categories, not philosophical grounding [579].

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 [592], and for Alan Westin to redefine privacy as control over information flows [564]—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 [589]. 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 [41]. 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 [588]. 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 [564]. 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 [593, 574]. 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 [575]. 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. "Functionalism underemphasizes the close and special connections moralists have stressed between and among privacy, personhood, and fitness for social participation and contribution" [569]. 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 [418]—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 [12].

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 [12]. CI's normative emphasis on preserving contextual norms pragmatically instrumentalizes privacy as a means to secure human values across complex social realms [13]. 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 [12].

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 [13]. 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 [12, 558]. In both cases, CI's reliance on local standards to articulate privacy's value leaves the theory normatively underpowered. 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 [558].

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 [133, 136]. 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 [150, 594]. Whether operating as employers with authoritarian degrees of control over workers' private lives [150] or providers of exploitative data-intensive consumer products [594], 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* [133, 136]. Opportunities for meaningful participation in shaping norms continue to diminish [413], as commercial interests increasingly displace valued social norms [595, 596]. As philosopher Michael Sandel observes, market norms, unfit to address normative problems, increasingly displace social norms across nearly all shades of life [596, 595]. 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 [597, 598, 12], 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 [12, 425]. 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 [13].

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 [12]. 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 [425]. 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 [12, 558], grounding normative privacy claims in the integrity of social spheres such as health, education, and work [425]. 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 [13]. 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 [12].

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 [413, 595]. 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” [14]

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” [14]. A moral minimum grounded in shared human experiences—not bound by local meaning—can establish “*a common moral horizon*” [14], 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” [425]. 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*” [14].

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 [599].

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 [14]. 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 [600, 601]. 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 [602].

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 [603]. Dignity has been described as the ontological root of privacy [604] and a normative lens for interpreting subjective experience in value-sensitive policy-making [605]. 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 [603].

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 [606], Universal Declaration of Human Rights [602]). A foundational value of the European Union, human dignity "must be respected, protected and constitutes the real basis of fundamental rights [607]." 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 [608].

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 [609]. 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 [610]. 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 [611, 612] and corrode [596] 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 [613, 13, 406, 614].

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 [615, 616]. 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 recognize a place for each" [12]. 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 that 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 [578]. 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 [618]. 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" [619]. He underscores that the offense lies neither in what was done with the information nor in the harm caused, but in the a 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 Brandies' 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 [579]. 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 [579].

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 [582]. 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 [589]. 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" [569]. 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 [620]. 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 [621, 622]. Privacy is essential to these capacities as it creates the psychic and social space necessary for individuals to establish their identifies 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 [623]—and by extension, Thomas Nagel insists, our abilities to maintain cohesive and cooperative societies [575]. 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 [624]. 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" [625]. 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 [563]. 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 [626, 627].

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 [628]. 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 agents—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 [629]. 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 [630]. 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 [588, 591, 631]—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 [632]. 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 [633] and expanded into a full theory of justice by philosopher Martha Nussbaum [15], 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 [629]. 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 [633]. 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 [634, 635, 15]. 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 [15, 599, 636].

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 [15]. 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 [15]. 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 [636].

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 [633] 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 [603], 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 [453, 406, 355, 453, 631]. 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 [599]. 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 [15]:

- **“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 [604, 637]. 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 [628, 569, 577]—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 [563]; to experience, develop, and act upon one's *emotions* and dispositional self in accordance with one's values [576, 569]; to experience one's *senses, imagination, and thought* with moderation over how one's inner life—thoughts, emotions, and sentiments [577, 623]—is known by others; to maintain *bodily integrity* with autonomous decision-making about one's *life* and *bodily health* [638, 569]; to freely participate in political and material environments [563, 639, 39]; to maintain meaningful *affiliations* with others on terms of dignity and mutual recognition [453, 578], and to engage in *play* and personal pursuits [569]. 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 [15]. 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 [14], 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 [12]; CA grounds dignity in the texture of human functioning [15]. 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 mininum, 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) [13, 558]. 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 [558]. 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 [12, 558]. 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 [14]. 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 the 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.

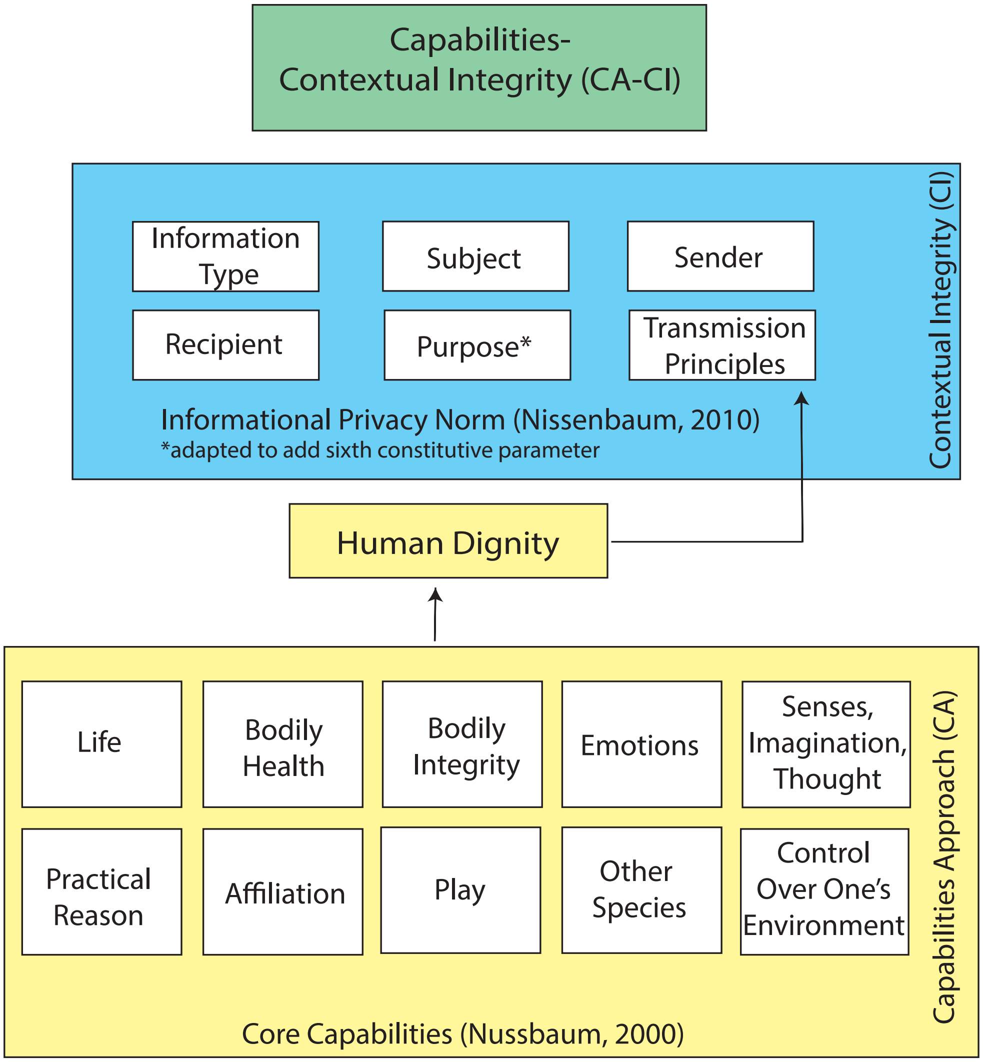

Capabilities-
Contextual Integrity (CA-CI)
Information Type
Subject
Sender
Recipient
Purpose*
Transmission Principles
Informational Privacy Norm (Nissenbaum, 2010)
*adapted to add sixth constitutive parameter
Contextual Integrity (CI)
Human Dignity
Life
Bodily Health
Bodily Integrity
Emotions
Senses, Imagination, Thought
Practical Reason
Affiliation
Play
Other Species
Control Over One's Environment
Core Capabilities (Nussbaum, 2000)
Capabilities Approach (CA)

Figure 7.1: Capabilities–Contextual Integrity (CA–CI) Theoretical Framework. Integrates 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 contraints) 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 [640, 641, 148, 642]. 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 [643, 644, 645]. 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 [646, 647] while obscuring deeper socio-technical vulnerabilities—misaligned system defaults, unexamined modeling assumptions, and inadequate sensitivity to context. [645].

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) [648], the NIST Privacy Framework [649], and ISO/IEC 23894 [650], 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 [648]), 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 text-based mental health counseling, licensed millions of anonymized crisis conversation transcripts to its for-profit spin-off, Loris.ai, to train commercial "empathy" algorithms [651, 652]. 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 [653]. 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. [654]. 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" [655].

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 [656]), 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 [12], 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 [651].

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.

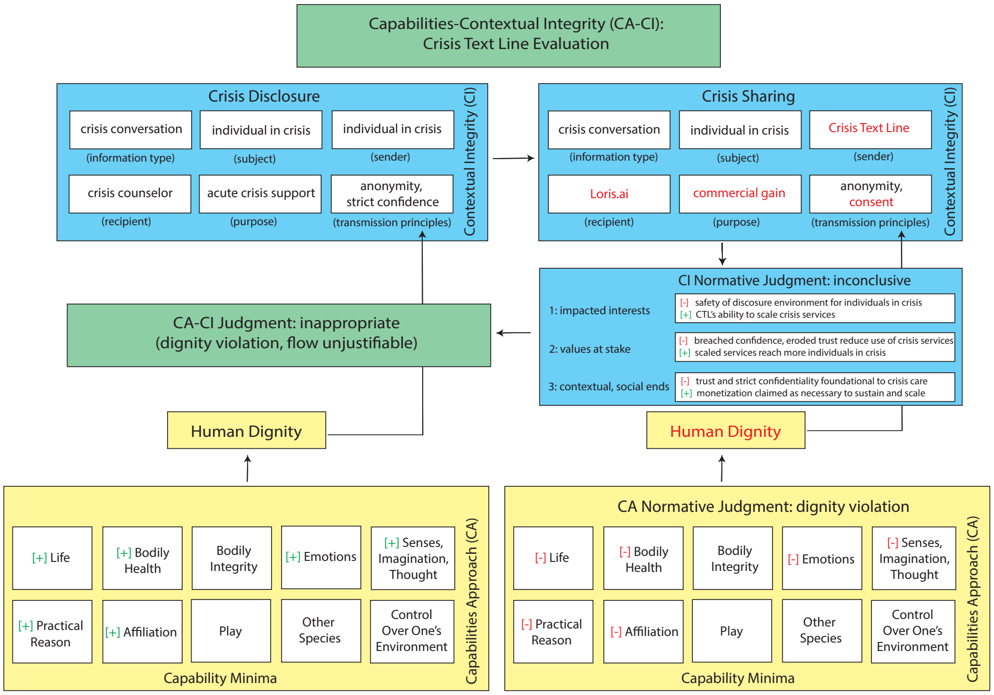

Capabilities-Contextual Integrity (CA-CI):
Crisis Text Line Evaluation
Crisis Disclosure
crisis conversation
(information type)
individual in crisis
(subject)
individual in crisis
(sender)
crisis counselor
(recipient)
acute crisis support
(purpose)
anonymity, strict confidence
(transmission principles)
Contextual Integrity (CI)
Crisis Sharing
crisis conversation
(information type)
individual in crisis
(subject)
Crisis Text Line
(sender)
Loris.ai
(recipient)
commercial gain
(purpose)
anonymity, consent
(transmission principles)
Contextual Integrity (CI)
CI Normative Judgment: inconclusive
1: impacted interests
[-] safety of discosure environment for individuals in crisis
[+] CTL's ability to scale crisis services
2: values at stake
[-] breached confidence, eroded trust reduce use of crisis services
[+] scaled services reach more individuals in crisis
3: contextual, social ends
[-] trust and strict confidentiality foundational to crisis care
[+] monetization claimed as necessary to sustain and scale
CA-CI Judgment: inappropriate
(dignity violation, flow unjustifiable)
Human Dignity
Human Dignity
[+] Life
[+] Bodily Health
Bodily Integrity
[+] Emotions
[+] Senses, Imagination, Thought
[+] Practical Reason
[+] Affiliation
Play
Other Species
Control Over One's Environment
Capability Minima
Capabilities Approach (CA)
CA Normative Judgment: dignity violation
[-] Life
[-] Bodily Health
Bodily Integrity
[-] Emotions
[-] Senses, Imagination, Thought
[-] Practical Reason
[-] Affiliation
Play
Other Species
Control Over One's Environment
Capability Minima
Capabilities Approach (CA)

Figure 7.2: Diagram of 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 [656]. 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 [657]. 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 [658]. 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 Summary of 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 *life; bodily health; emotions; senses, imagination, and thought; practical reason, affiliation.*	
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" [659].

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 [660, 661]. 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 [662], 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 [663, 664]. 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 [665].

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 [666]. 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 [667]. 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 [668].

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 [659, 644]. 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 [669]. Empirical studies confirm that location data, mobile use, and social media behavior can reliably predict sensitive attributes like occupation, gender, and mental health status [670, 671, 672].

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 [644]. 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 [673] 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, repurposing, 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.

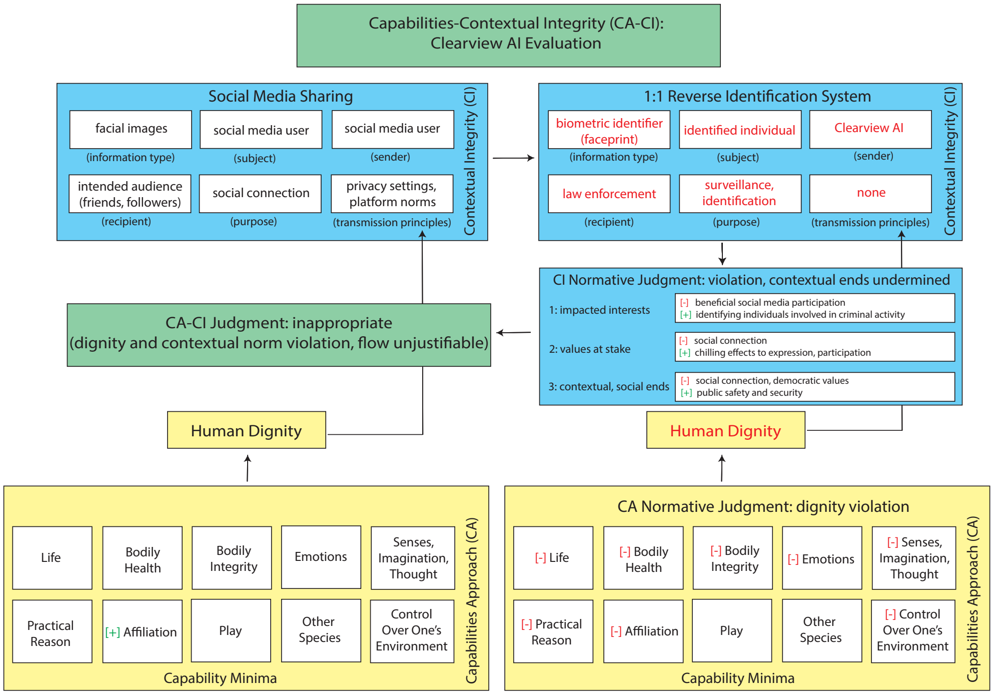

Capabilities-Contextual Integrity (CA-CI):
Clearview AI Evaluation
Social Media Sharing
facial images
(information type)
social media user
(subject)
social media user
(sender)
intended audience (friends, followers)
(recipient)
social connection
(purpose)
privacy settings, platform norms
(transmission principles)
Contextual Integrity (CI)
1:1 Reverse Identification System
biometric identifier (faceprint)
(information type)
identified individual
(subject)
Clearview AI
(sender)
law enforcement
(recipient)
surveillance, identification
(purpose)
none
(transmission principles)
Contextual Integrity (CI)
CI Normative Judgment: violation, contextual ends undermined
1: impacted interests
[-] beneficial social media participation
[+] identifying individuals involved in criminal activity
2: values at stake
[-] social connection
[+] chilling effects to expression, participation
3: contextual, social ends
[-] social connection, democratic values
[+] public safety and security
CA-CI Judgment: inappropriate
(dignity and contextual norm violation, flow unjustifiable)
Human Dignity
Human Dignity
Life
Bodily Health
Bodily Integrity
Emotions
Senses, Imagination, Thought
Practical Reason
[+] Affiliation
Play
Other Species
Control Over One's Environment
Capability Minima
Capabilities Approach (CA)
CA Normative Judgment: dignity violation
[-] Life
[-] Bodily Health
[-] Bodily Integrity
[-] Emotions
[-] Senses, Imagination, Thought
[-] Practical Reason
[-] Affiliation
Play
Other Species
[-] Control Over One's Environment
Capability Minima
Capabilities Approach (CA)

Figure 7.3: Diagram of 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 [674, 675, 676].

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 [98].

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 [677]. 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 [678]—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 [47]; as users begin to anticipate surveillance, they may withdraw from social life online altogether [679].

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 [680].

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 [681, 682]. 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 regrounds these abstractions in precisely why they matter—not bo 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” [15]. We might see this and think: well, the ability to move freely is only affected for criminals with something to hide [355]. 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 [683]. 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* [684]. 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 [685]. 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 [686]. 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% [687].

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 [684]—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 [686]. 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 [668]. 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) [688, 689]. 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 [690]—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 Summary of 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 *affiliation*, persistent traceability threatens *bodily integrity*, *bodily health*, and *life*; downstream risks implicate *practical reason*, *control over environment*, and *emotions*.	
Normative Reasoning	Contextual ends in both source and destination frustrated; flow deemed *inappropriate* 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 [691, 692, 693, 694, 695]. 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 [696, 148, 697]. 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 [698, 699]. Large-scale harm benchmarking projects like HarmBench [695] 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 [700]. Attempts to synthesize these into unified benchmarks (e.g., [701, 700]) 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 [702, 703, 704].

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 [705]. 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 [706]—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 [705]—*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.

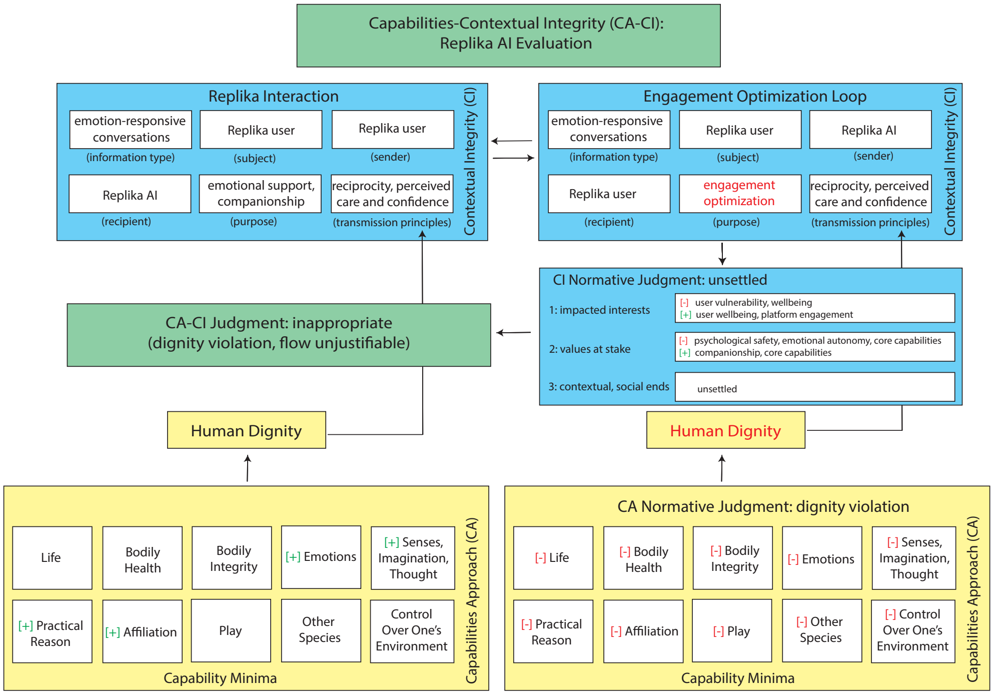

Capabilities-Contextual Integrity (CA-CI):
Replika AI Evaluation
Replika Interaction
emotion-responsive conversations
(information type)
Replika user
(subject)
Replika user
(sender)
Replika AI
(recipient)
emotional support, companionship
(purpose)
reciprocity, perceived care and confidence
(transmission principles)
Contextual Integrity (CI)
Engagement Optimization Loop
emotion-responsive conversations
(information type)
Replika user
(subject)
Replika AI
(sender)
Replika user
(recipient)
engagement optimization
(purpose)
reciprocity, perceived care and confidence
(transmission principles)
Contextual Integrity (CI)
CI Normative Judgment: unsettled
1: impacted interests
[-] user vulnerability, wellbeing
[+] user wellbeing, platform engagement
2: values at stake
[-] psychological safety, emotional autonomy, core capabilities
[+] companionship, core capabilities
3: contextual, social ends
unsettled
CA-CI Judgment: inappropriate
(dignity violation, flow unjustifiable)
Human Dignity
Human Dignity
Life
Bodily Health
Bodily Integrity
[+] Emotions
[+] Senses, Imagination, Thought
[+] Practical Reason
[+] Affiliation
Play
Other Species
Control Over One's Environment
Capability Minima
Capabilities Approach (CA)
CA Normative Judgment: dignity violation
[-] Life
[-] Bodily Health
[-] Bodily Integrity
[-] Emotions
[-] Senses, Imagination, Thought
[-] Practical Reason
[-] Affiliation
[-] Play
[-] Other Species
[-] Control Over One's Environment
Capability Minima
Capabilities Approach (CA)

Figure 7.4: Diagram of 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 [707]. 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 underping a flourishing life [1, 599]. These benefits help explain why many users report positive experiences with chatbots [708], 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* [709].

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 [705], 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 [706]. 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 [710]. 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 [711]; preference-embedding models capture intransitive or cyclic utilities [712]; generative judges replace opaque scalars with natural-language rationales [713]. 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 [714, 1, 98]. 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."* [714].

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 Summary of 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 *affiliation*, *practical reason*, *emotions*, and *senses, imagination, and thought*, 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 with 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 sociotechnical 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 [715] and social well-being [716]. 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 [14]: 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 [652]. 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) [715], information rights [717], design [718], best interests standards [719], and cybersecurity [720].

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) [12], 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) [15]. 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 [15] and Nissenbaum's Contextual Integrity [12], 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 [576, 133, 373, 453], 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 [133] and the "Silicon Valleyification of Everything" [721, 652], CA-CI helps reclaim their dignity-securing functions [416, 722].

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.

CHAPTER 8

Concluding Discussion: Reclaiming Dignity in Digital Life

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 [723, 724, 725, 726, 727]. 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 [133, 134]. 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 [609], while the public and scholars alike increasingly demand stronger safeguards keyed to human dignity [728, 729].

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 [602]. 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 [730], is no longer sidelined as speculative [731, 732]. Surprising emergent capabilities in today's scale-driven models have compressed the timelines that many AI researchers now assign to human-level AGI [733] and have shaken confidence that generative systems can be reliably constrained [734, 735, 736]. *Superintelligent* systems capable of eclipsing human cognition and eluding human oversight remain hypothetical, though multiple technically plausible paths toward them are openly discussed [737, 738, 739]. What was once the stuff of science fiction [740] has therefore become a focal point of contemporary AI governance, pushing questions of human values alignment from the margins to mainstream domain discourse [741, 148].

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 [589].

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 [742]. But this governance paralysis overlooks an existing ethical consensus: the conviction that human dignity is non-negotiable, enshrined in post-war human rights agreements [602].

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) [12] and Martha Nussbaum's Capabilities Approach (CA), which defines the constituent parts and minimum requirements of a life capable of dignity [15], 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 [615] and privacy regulation specifications [616, 558], 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 [15]. 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 [743]. 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 [606, 744, 667, 745]. Articles 7 and 8 of the Charter ground those norms in human dignity, promising respect for private life and robust data-protection safeguards [600]. Yet persistent gaps remain: EU monitoring registers declining public trust and a widening sense of digital disempowerment [611]. 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 [659]. 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 in 2023 [746], 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 [558], 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 of the EU Charter 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 [747], 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 baseline, 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" [569]. 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 [12], 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 [1], 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 [748]. 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 [749], 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 [529].

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) [667, 659]. 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., [749], 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 [145, 144, 750]. 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 [751]. 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 [752, 753]. 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 forseeable 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 [754].

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 [667, 744]—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 [667, 744], 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) [659], with particular emphasis on compounding effects that may accumulate over time, exacerbate vulnerabilities, and produce severe long-term consequences [667].

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 [659]. 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 forseeable 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 [755, 756, 757, 758, 759, 98].

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 [760, 761, 762]. 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: Interview Protocol (Ch. 3)

A.1 Interview Questions

Thanks so much for taking the time to talk to me today. I appreciate it. This is a study about people's experiences with and thoughts about social media as it relates to their emotional experiences. I want to emphasize that there is no right or wrong answer to anything I ask about. What I really want to learn about is your experiences and thoughts. Before we start:

- Is there any questions you have for me?
- Do I have your permission to record this audio?

Generic questions about social media use/non-use

- You mentioned in the screening survey that you use these social media. . . tell me a bit about how you use them each, what for, who you're connected to. . . ?
- What kinds of things do you generally post on each?

Positive personal experiences

- When was the last time you had something good and exciting happen, something that was personally meaningful to you? What was it about?
- What emotions would you say this experience evoked for you?
- Did you post about it on social media? Why yes? Why not?
- For people who say they did not share about the most recent positive experience: ask for a positive experience they did share about in the past year, and then they will answer these questions.

- Did the post communicate these feelings explicitly? Implicitly? How? (Feel free to open it up on your phone or computer and take a look. . .)
- What words did you use to describe those feelings?
- Did you think about who else might see this beyond the people that you shared it with?
- Beyond the people you shared this with, do you think any other entities had access to or used this information you shared? What makes you think that? How do you feel about that? Why?
- Did you notice that anything about the platform changed after you posted? How did that make you feel?
- How do you think these changes were relevant to what you posted? Were they relevant to your feelings?

Probes:

- Ads, recommended content/people/events
- Content showing up (or not) in feeds (e.g., have you noticed changes in the above. . . ?)
- Did you notice any changes elsewhere on the internet after you posted about X? What were they? How did that make you feel?

Negative personal experiences

- When was the last time you had something negative happen, maybe something that was deep and personal? What was it about? Tell me about it.
- What emotions would you say this evoked for you?
- Did you post about it on social media? Why yes? Why not?
- Same follow-up questions if they did share about something negative.
- If the example was about something negative that they did not post about: Can you think of a time you posted about something personal that was negative or difficult? Then same follow-up questions.

All personal experiences

- When you think about these good and bad we talked about, what were your expectations for privacy? Why?
- Has there been a time that you felt your expectations for privacy were not met in these spaces, in particular after you shared something like the ones you mentioned earlier? Give me an example.
- (Probe on both by other people, the platform, other parties.)

Other content, not personal experiences

- Do you ever share other content that are not about personal experiences on social media? Give me some examples.
- Did these posts generally communicate any of your personal feelings explicitly? Implicitly? How? What were they about? (Feel free to open it up on your phone or computer and take a look. . .)
- Did you think about who else might see them beyond the people that you shared them with?
- Beyond the people you shared them with, do you think any other entities had access to or used this information you shared? What makes you think that? How do you feel about that? Why?
- Did you notice that anything about the platform changed after you posted? How did that make you feel? Why?

Probes:

- Ads, recommended content/people/events
- Content showing up (or not) in feeds
- Did you notice any changes elsewhere on the internet after these posts?

Vignettes

Participants will be asked to imagine and discuss scenarios. Randomize order in which they are presented. Everyone will discuss all six scenarios regardless of actual experiences. If the experience does not exist, ask them to "imagine. . . "

Direct (once for negative and once for positive experiences).

I would like you to think about something [positive/negative and personal] that brought out [positive/negative] emotions for you. 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.

Now imagine that [insert social media they posted on] used computational methods to detect what emotions you felt at the time of posting that. How do you feel about that? Why do you think they might do that? Before you respond, please tell me what experiences and feelings you thought about for context.

How do you feel about your post being used to predict how you might feel in the future? Why? Why do you think they might do that? How do you feel about that?

How do you feel about the platform using this prediction or detection to recommend things related to what you posted about?

Probe:

- Recommend content in your news feed on [insert social media name]? On other platforms? Why?
- Recommend people for you to engage with on [insert social media name]? On other platforms? Why?
- Recommend products or services to buy through ads on [insert social media name]? On other platforms? Why?
- Recommend/deliver messages from political campaigns. Why?
- Recommend events in geographic area that you might be interested in. Why?

How do you feel about this prediction or detection being used to:

- Intervene with some kind of support to help you feel better? (More relevant for negative experiences). Why?
- Infer a mental or physical medical condition you may have, had in the past, or will have? Why?
- Not only learn about you, but about other people who may be similar to you, and do the above interventions and recommendations for them? Why?

How would you feel if the social media you used sold these predictions to third parties? Why? Probe for: your employer, other businesses, insurance companies, or the government.

What if they just sold your data, and not the predictions, so that the other entities can make their own predictions? Why?

Indirect (once for negative and once for positive experiences).

I would like you to think about something [positive/negative and personal] that brought out [positive/negative] emotions for you. Now consider this scenario: You had hinted to that on [insert social media they use most], and very vaguely shared how you felt about it. Not everyone reading it, or perhaps no one reading it, would have been able to understand what your experience actually was and how you felt. But you knew what you were talking about.

Now imagine that [insert social media they posted on] used computational methods to detect what emotions you felt at the time of posting that, even though you never explicitly wrote anything.

How do you feel about that? Why? Why do you think they might do that?

Same follow-up questions/probes for positive topics. Same follow-up questions/probes for negative topics.

Not sharing at all (once for negative and once for positive experiences).

You had not shared on [insert social media they use most] about X – this means you have not explicitly or vaguely shared how you felt about it. But you may have done other things online, such as shopping or seeking information or reading content about X or even about other things.

Now imagine that [insert social media they posted on] used computational methods to detect what emotions you felt around that time X happened, even though you had not posted about X or your emotions.

How do you feel about that? Why do you think they might do that?

Same follow-up questions/probes for positive topics. Same follow-up questions/probes for negative topics.

Follow-up questions (apply across scenarios)

- Does your intended audience/recipient impact what you are comfortable with, in terms of what happens to your data? Why?
- Is it different if people you shared with know how you feel vs. the algorithm reading your posts? How and why?
- Does your intention to use the platform as you did impact your comfort? Why?
- Does how publicly or privately you shared content (public post, friends, private messages) impact your comfort? Why?
- Does what exact emotion is being detected impact your comfort? Why?
- Does the topic of the post matter (e.g., how personal it is)?
- Does whether your post includes your face matter (e.g., if they detect emotions from a photo)? Why?
- Does individual vs. aggregate-level prediction matter? Why?
- Does whether prediction identifies/singles you out matter? Why?
- Does transparency of detection/prediction matter? Why?
- Does transparency in how detections/predictions influence experiences matter? Why?
- Does accuracy of detection/prediction matter? Why?
- Does awareness that companies do these things matter? Why?
- Does meaningful consent matter? Why? What would meaningfulness look like to you?
- What do you think you have given meaningful consent for as of now, when it comes to what happens to your emotional data?

Wrap up

- Thinking of the scenarios we talked about, from explicit sharing of emotions to being vague to not sharing, how would you compare your expectations around what happens to your emotional data? Why?

- If these computational detections and predictions of emotions were to occur using your data on these platforms, would that change your use of them? Why? How about your use of other technology in general?

- What if you just knew that this could happen but were not sure if it was happening or not? Would that change your use of these platforms? How? Why? What impact would this change have on you?

- How do the above scenarios match with your expectations (what you already expect would be happening)?

- How do the above scenarios match with your desires (what you would or would not want to happen)?

- How do you think the above scenarios (emotion detection and companies using your emotional data to shape your experience online) may impact you, positively or negatively? How do you feel about that, and why?

- Probe: Do you think you may experience any harm of any kind, including but not limited to psychological or wellbeing, at any level, or may be put at a disadvantage as a result of such emotion detection and what may be done with your emotional data? How, and why?

- Anything else we haven't talked about that you think is relevant?

APPENDIX B

Supplemental Materials: Recruitment and Interview Protocol (Ch. 5)

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 Questions

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 established 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 use 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 <emotion 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 <emotion 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 <emotion 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 <emotion 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 <emotion 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 <emotion 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 <emotion 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 <emotion 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 <emotion 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 <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 <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 <Emotion AI tool>? *(Follow up questions might include, "How do you think <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 <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 <emotion AI tool> making inferences about how you feel? *(Follow up questions might include, "Why might that be?")*

- Was use of <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 <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 <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 <X> has access to your data from <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 <emotion 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 <emotion 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 <emotion 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 <emotion 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.

APPENDIX C

Supplemental Materials: Survey Questions (Ch. 6)

C.1 Employment Vignettes

C.1.1 Speech/Text Data

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**:

1. inferring the mental health state of employees. Inferences of an individual's mental health will not be made; only at a group level.
2. inferring the mental health state of employees individually.
3. diagnosing mental illness in employees earlier than otherwise possible.
4. diagnosing neurological disorders, such as dementia or ADHD, in employees earlier than otherwise possible.
5. identifying employees in need of mental health support, to better plan organizational mental health resources.
6. developing an intelligent computer program, such as a chatbot, that can conduct mental health therapy with employees, including you.
7. 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. sharing that information with academic researchers to help them learn more about mental health, as part of a research partnership.

9. giving employers data-driven insights into employees' wellbeing.

10. automatically alerting your employer when employees may need support, including you.

11. inferring whether employees are at risk of harming themselves.

12. inferring whether employees are at risk of harming others.

13. avoiding subjectivity in other methods of your employer learning about your emotional state, like a survey or your employer's observations.

14. assessing the work performance of individual employees.

C.1.2 Image/Video Data

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**:

1. inferring the mental health state of employees. Inferences of an individual's mental health will not be made; only at a group level.

2. inferring the mental health state of employees individually.

3. diagnosing mental illness in employees earlier than otherwise possible.

4. diagnosing neurological disorders, such as dementia or ADHD, in employees earlier than otherwise possible.

5. identifying employees in need of mental health support, to better plan organizational mental health resources.

6. developing an intelligent computer program, such as a chatbot, that can conduct mental health therapy with employees, including you.

7. 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. sharing that information with academic researchers to help them learn more about mental health, as part of a research partnership.

9. giving employers data-driven insights into employees' wellbeing.

10. automatically alerting your employer when employees may need support, including you.

11. inferring whether employees are at risk of harming themselves.

12. inferring whether employees are at risk of harming others.

13. avoiding subjectivity in other methods of your employer learning about your emotional state, like a survey or your employer's observations.

14. assessing the work performance of individual employees.

C.2 Healthcare Vignettes

C.2.1 Speech/Text Data

As a patient, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your healthcare provider(s) 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**:

1. inferring the mental health state of patients. Inferences of an individual's mental health will not be made; only at a group level.

2. inferring the mental health state of patients individually.

3. diagnosing mental illness in patients earlier than otherwise possible.

4. diagnosing neurological disorders, such as dementia or ADHD, in patients earlier than otherwise possible

5. inferring patients in need of wellbeing support.

6. developing an intelligent computer program, such as a chatbot, that can conduct mental health therapy with patients, including you.

7. 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. sharing that information with academic researchers to help them learn more about mental health, as part of a research partnership.
9. giving healthcare provider(s) increased understanding about patients through data-driven insights.
10. automatically alerting your healthcare provider(s) when patients may need support, including you.
11. inferring whether patients are at risk of harming themselves.
12. inferring whether patients are at risk of harming others.
13. 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. assessing the overall health of individual patients.

C.2.2 Image/Video

As a patient, rate your comfort (from 0 = "very uncomfortable" to 100 = "very comfortable") with your healthcare provider(s) 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**:

1. inferring the mental health state of patients. Inferences of an individual's mental health will not be made; only at a group level.
2. inferring the mental health state of patients individually.
3. diagnosing mental illness in patients earlier than otherwise possible.
4. diagnosing neurological disorders, such as dementia or ADHD, in patients earlier than otherwise possible
5. inferring patients in need of wellbeing support.
6. developing an intelligent computer program, such as a chat bot, that can conduct mental health therapy with patients, including you.

7. 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. sharing that information with academic researchers to help them learn more about mental health, as part of a research partnership.

9. giving healthcare provider(s) increased understanding about patients through data-driven insights.

10. automatically alerting your healthcare provider(s) when patients may need support, including you.

11. inferring whether patients are at risk of harming themselves.

12. inferring whether patients are at risk of harming others.

13. 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. assessing the overall health of individual patients.

C.3 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?

C.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

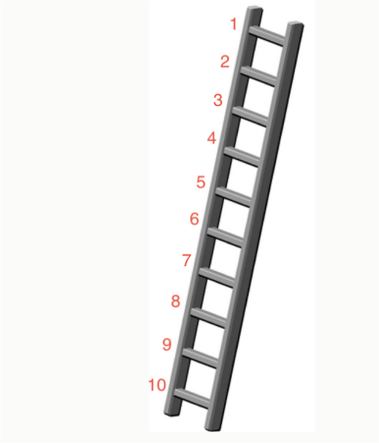


Figure C.1: MacArthur Scale of Subjective Social Status

- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- Prefer not to answer

C.5 Post-test Individual Belief Questions

C.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.

C.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**.

C.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**.

C.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?

Bibliography

[1] M. C. Nussbaum. Upheavals of Thought: The Intelligence of Emotions. Cambridge University Press, 2003.

[2] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111.24 (2014), p. 8788. DOI: 10.1073/pnas.1320040111.

[3] K. Huckvale, S. Venkatesh, and H. Christensen. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *npj Digital Medicine* 2.1 (2019), pp. 1–11. DOI: 10.1038/s41746-019-0166-1.

[4] A. Madan, M. Cebrian, D. Lazer, and A. Pentland. Social Sensing for Epidemiological Behavior Change. Proceedings of the 12th ACM International Conference on Ubiquitous Computing. UbiComp '10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 291–300. DOI: 10.1145/1864349.1864394.

[5] G. Tay, A. Zimmerman, and B. Elliot. Maverick* Research: Emotional Wellness Will Rescue Your Organization and Distributed Workforce. Gartner. 2021.

[6] A. Verma. Hype Cycle for Sensing Technologies and Applications. Gartner. 2020.

[7] Y. Kadoya, M. S. R. Khan, S. Watanapongvanich, and P. Binnagan. Emotional status and productivity: Evidence from the special economic zone in Laos. *Sustainability* 12.4 (2020), p. 1544.

[8] S. Oh, J. H. Kim, S.-W. Choi, H. J. Lee, J. Hong, and S. H. Kwon. Physician confidence in artificial intelligence: An online mobile survey. *Journal of Medical Internet Research* 21.3 (2019). DOI: 10.2196/12422.

[9] G. González-Alcaide, M. Fernández-Ríos, R. Redolat, and E. Serra. Research on emotion recognition and dementias: Foundations and prospects. *Journal of Alzheimer's Disease* 82.3 (2021), pp. 939–950. DOI: 10.3233/JAD-210096.

[10] D. McDuff, E. Jun, K. Rowan, and M. Czerwinski. Longitudinal Observational Evidence of the Impact of Emotion Regulation Strategies on Affective Expression. *IEEE Transactions on Affective Computing* 12.3 (2021), pp. 636–647. DOI: 10.1109/TAFFC.2019.2961912.

[11] K. Toyama. Technology as Amplifier in International Development. Proceedings of the 2011 iConference. New York, NY: ACM, 2011, pp. 75–82. DOI: 10.1145/1940761.1940772.

[12] H. Nissenbaum. Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford, CA, USA: Stanford University Press, 2009, p. 304.

[13] H. Nissenbaum. Respecting Context to Protect Privacy: Why Meaning Matters. *Science and Engineering Ethics* 24.3 (2018), pp. 831–852. DOI: 10.1007/s11948-015-9674-9.

[14] M. Walzer. Thick and Thin: Moral Argument at Home and Abroad. Notre Dame, IN, USA: University of Notre Dame Press, 1994, pp. xi, 108.

[15] M. C. Nussbaum. Women and Human Development: The Capabilities Approach. Vol. 3. Cambridge University Press, 2000.

[16] J. R. Averill, K. K. Chon, and D. W. Hahn. Emotions and creativity, east and west. *Asian Journal of Social Psychology* 4.3 (2001), pp. 165–183.

[17] S. Marsella, J. Gratch, and P. Petta. Computational Models of Emotion. A Blueprint for Affective Computing: A Sourcebook and Manual. Ed. by P. Petta, C. Pelachaud, and R. Cowie. New York, NY, USA: Oxford University Press, 2010, pp. 21–46.

[18] A. Scarantino. The philosophy of emotions and its impact on affective science. *Handbook of Emotions* 4 (2016), pp. 3–48.

[19] L. F. Barrett. Are emotions natural kinds? *Perspectives on psychological science* 1.1 (2006), pp. 28–58.

[20] R. Gupta. Positive emotions have a unique capacity to capture attention. *Progress in Brain Research* 247 (2019), pp. 23–46.

[21] J. S. Beer and D. Keltner. What Is Unique about Self-Conscious Emotions? *Psychological Inquiry* 15.2 (2004), pp. 126–129.

[22] K. Mogg and B. P. Bradley. A cognitive-motivational analysis of anxiety. *Behaviour research and therapy* 36.9 (1998), pp. 809–848.

[23] S. Marsella and J. Gratch. Computationally Modeling Human Emotion. *Communications of the ACM* 57.12 (2014), pp. 56–67. DOI: 10.1145/2631912.

[24] M. Zembylas. Emotions and teacher identity: A poststructural perspective. *Teachers and Teaching* 9.3 (2003), pp. 213–238.

[25] J. M. Barbalet. Emotion, Social Theory, and Social Structure: A Macrosociological Approach. Cambridge University Press, 2001.

[26] J. L. Ackrill. Aristotle's Ethics. *Tijdschrift Voor Filosofie* 35.3 (1973).

[27] A. S. Waterman. The Relevance of Aristotle's Conception of Eudaimonia for the Psychological Study of Happiness. *Theoretical & Philosophical Psychology* 10.1 (1990), p. 39.

[28] D. M. Haybron. Happiness, the Self and Human Flourishing. *Utilitas* 20.1 (2008), pp. 21–49.

[29] J. Annas. The Morality of Happiness. Oxford University Press, 1993.

[30] D. M. McMahon. The Quest for Happiness. *Wilson Quarterly* 29.1 (2005), pp. 62–71.

[31] S. Ahmed. The Promise of Happiness. Duke University Press, 2010.

[32] L. Sundararajan. Happiness Donut: A Confucian Critique of Positive Psychology. *Journal of Theoretical and Philosophical Psychology* 25.1 (2005), pp. 35–60.

[33] J. Legge et al. Confucian Analects: The Great Learning, and the Doctrine of the Mean. Courier Corporation, 1971.

[34] S. L. Gordon. Social structural effects on emotions. *Research Agendas in the Sociology of Emotions* (1990), pp. 145–179.

[35] C. Lutz and G. M. White. The anthropology of emotions. *Annual Review of Anthropology* 15.1 (1986), pp. 405–436.

[36] M. Zembylas. Emotion, resistance, and self-formation. *Educational Theory* 53.1 (2003).

[37] D. Y. Dai and R. J. Sternberg. Motivation, Emotion, and Cognition: Integrative Perspectives on Intellectual Functioning and Development. Routledge, 2004.

[38] R. B. Zajonc. Emotions. The Handbook of Social Psychology. Ed. by D. T. Gilbert, S. T. Fiske, and G. Lindzey. 4th ed. Vol. 1. Boston, MA: McGraw-Hill, 1998, pp. 591–632.

[39] A. E. Taslitz. The Fourth Amendment in the twenty-first century: Technology, privacy, and human emotions. *Law & Contemporary Problems* 65 (2002), p. 125.

[40] I. Altman. Privacy Regulation: Culturally Universal or Culturally Specific? *Journal of Social Issues* 33.3 (1977), pp. 66–84. DOI: 10.1111/j.1540-4560.1977.tb01883.x.

[41] D. K. Citron and D. J. Solove. Privacy harms. *Boston University Law Review* 102 (2022), p. 793.

[42] L. Stark. The emotional context of information privacy. *The Information Society* 32.1 (2016), pp. 14–27.

[43] H. Li, X. R. Luo, J. Zhang, and H. Xu. Resolving the privacy paradox: Toward a cognitive appraisal and emotion approach to online privacy behaviors. *Information & management* 54.8 (2017), pp. 1012–1022.

[44] R. Wakefield. The influence of user affect in online information disclosure. *The Journal of Strategic Information Systems* 22.2 (2013), pp. 157–174.

[45] H. Li, R. Sarathy, and H. Xu. The role of affect and cognition on online consumers' decision to disclose personal information to unfamiliar online vendors. *Decision Support Systems* 51.3 (2011), pp. 434–445.

[46] D. Teutsch, P. K. Masur, and S. Trepte. Privacy in mediated and nonmediated interpersonal communication: How subjective concepts and situational perceptions influence behaviors. *Social Media + Society* 4.2 (2018), p. 2056305118767134.

[47] A. E. Waldman. Privacy as Trust: Information Privacy for an Information Age. Cambridge University Press, 2018.

[48] P. McCole, E. Ramsey, and J. Williams. Trust considerations on attitudes towards online purchasing: The moderating effect of privacy and security concerns. *Journal of Business Research* 63.9-10 (2010), pp. 1018–1024.

[49] A. L. Allen. Lying to protect privacy. *Villanova Law Review* 44 (1999), p. 161.

[50] S. T. Margulis. Conceptions of Privacy: Current Status and Next Steps. *Journal of Social Issues* 33.3 (1977), pp. 5–21. DOI: 10.1111/j.1540-4560.1977.tb01879.x.

[51] K. Shobe. Productivity driven by job satisfaction, physical work environment, management support and job autonomy. *Business and Economics Journal* 9.2 (2018), pp. 1–9.

[52] K. Pugliesi. The consequences of emotional labor: Effects on work stress, job satisfaction, and well-being. *Motivation and emotion* 23.2 (1999), pp. 125–154.

[53] A. R. Hochschild. The Managed Heart. Routledge, 1983.

[54] J. K. Burgoon, R. Parrott, B. A. Le Poire, D. L. Kelley, J. B. Walther, and D. Perry. Maintaining and restoring privacy through communication in different types of relationships. *Journal of Social and Personal Relationships* 6.2 (1989), pp. 131–158.

[55] D. Hopkins, J. Kleres, H. Flam, and H. Kuzmics. Theorizing Emotions: Sociological Explorations and Applications. Campus Verlag, 2009.

[56] H. Flam and J. Kleres. Methods of Exploring Emotions. Routledge, 2015.

[57] T. Dixon. "Emotion": The history of a keyword in crisis. *Emotion Review* 4.4 (2012), pp. 338–344.

[58] C. E. Izard. The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review* 2.4 (2010), pp. 363–370.

[59] D. Grandjean, D. Sander, and K. R. Scherer. Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and cognition* 17.2 (2008), pp. 484–495.

[60] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31.2 (2013), pp. 120–136.

[61] M. B. Arnold. Emotion and Personality. 1960.

[62] R. S. Lazarus. Emotion and Adaptation. Oxford University Press, 1991.

[63] K. R. Scherer, A. Schorr, and T. Johnstone. Appraisal processes in emotion: Theory, methods, research. Oxford University Press, 2001.

[64] L. Stark and J. Hoey. The Ethics of Emotion in Artificial Intelligence Systems. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event Canada: ACM, 2021, pp. 782–793. DOI: 10.1145/3442188.3445939.

[65] E. Harmon-Jones, C. Harmon-Jones, and E. Summerell. On the importance of both dimensional and discrete models of emotion. *Behavioral sciences* 7.4 (2017), p. 66.

[66] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing* 1.1 (2010), pp. 18–37.

[67] M. Mortillaro, B. Meuleman, and K. R. Scherer. Advocating a componential appraisal model to guide emotion recognition. *International Journal of Synthetic Emotions (IJSE)* 3.1 (2012), pp. 18–32.

[68] P. Ekman and D. Keltner. Universal Facial Expressions of Emotion. Nonverbal Communication: Where Nature Meets Culture. Ed. by U. Segerstrale and P. Molnar. Cambridge University Press, 1997, pp. 27–46.

[69] P. Ekman. An argument for basic emotions. *Cognition & emotion* 6.3-4 (1992), pp. 169–200.

[70] P. Ekman, W. V. Friesen, and S. S. Tomkins. Facial Affect Scoring Technique: A First Validity Study. *Semiotica* 3.1 (1971), pp. 37–58.

[71] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17.2 (1971), pp. 124–129. DOI: 10.1037/h0030377.

[72] D. Matsumoto, S. Takeuchi, S. Andayani, N. Kouznetsova, and D. Krupp. The contribution of individualism vs. collectivism to cross-national differences in display rules. *Asian Journal of Social Psychology* 1.2 (1998), pp. 147–165.

[73] L. F. Barrett and T. D. Wager. The structure of emotion: Evidence from neuroimaging studies. *Current Directions in Psychological Science* 15.2 (2006), pp. 79–83.

[74] P. Ekman and D. Cordaro. What is meant by calling emotions basic. *Emotion Review* 3.4 (2011), pp. 364–370.

[75] P. Ekman. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology* 53.4 (1972), p. 712.

[76] P. Ekman. Facial expression and emotion. *American Psychologist* 48.4 (1993), p. 384.

[77] J. I. Durán, R. Reisenzein, and J.-M. Fernández-Dols. Coherence between emotions and facial expressions. *The science of facial expression* (2017), pp. 107–129.

[78] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest* 20.1 (2019), pp. 1–68.

[79] H. Kaur, D. McDuff, A. C. Williams, J. Teevan, and S. T. Iqbal. “I Didn’t Know I Looked Angry”: Characterizing Observed Emotion and Reported Affect at Work. CHI Conference on Human Factors in Computing Systems. 2022, pp. 1–18.

[80] L. Rhue. Racial Influence on Automated Perceptions of Emotions. 2018. arXiv: 3281765.

[81] E. Kim, D. Bryant, D. Srikanth, and A. Howard. Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. Virtual Event USA: ACM, 2021, pp. 638–644. DOI: 10.1145/3461702.3462609.

[82] L. Rhue. Understanding the Hidden Bias in Emotion-Reading AIs. 2019. URL: https://singularityhub.com/2019/01/11/understanding-the-hidden-bias-in-emotion-reading-ais/.

[83] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane. Gender De-Biasing in Speech Emotion Recognition. Interspeech. 2019, pp. 2823–2827.

[84] H. A. Elfenbein and N. Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin* 128.2 (2002), p. 203.

[85] C. Darwin. The Expression of the Emotions in Man and Animals. London: J. Murray, 1872.

[86] U. Hess and P. Thibault. Darwin and Emotion Expression. *American Psychologist* 64.2 (2009), p. 120.

[87] P. Ekman. Darwin's contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1535 (2009), pp. 3449–3451. DOI: 10.1098/rstb.2009.0189.

[88] L. F. Barrett. Was Darwin wrong about emotional expressions? *Current Directions in Psychological Science* 20.6 (2011), pp. 400–406.

[89] W. James. What Is an Emotion? (1884). Readings in Psychology. Appleton-Century-Crofts, 1948.

[90] G. Colombetti and E. Thompson. Enacting emotional interpretations with feeling. *Behavioral and Brain Sciences* 28.2 (2005), pp. 200–201.

[91] G. Colombetti. Enaction, sense-making and emotion. *Enaction: Toward a New Paradigm for Cognitive Science* (2010), pp. 145–164.

[92] M. R. Bennett and P. M. S. Hacker. Philosophical foundations of neuroscience. John Wiley & Sons, 2022.

[93] L. Wu, R. Huang, Z. Wang, J. N. Selvaraj, L. Wei, W. Yang, and J. Chen. Embodied emotion regulation: The influence of implicit emotional compatibility on creative thinking. *Frontiers in Psychology* 11 (2020), p. 1822.

[94] R. McCarty. The Fight-or-Flight Response: A Cornerstone of Stress Research. Stress: Concepts, Cognition, Emotion, and Behavior. Ed. by G. Fink. San Diego, CA: Elsevier, 2016, pp. 33–37.

[95] G. E. Weisfeld and S. M. Goetz. Applying evolutionary thinking to the study of emotion. *Behavioral Sciences* 3.3 (2013), pp. 388–407.

[96] J. LeDoux. The Emotional Brain: The Mysterious Underpinnings of Emotional Life. New York, NY: Simon and Schuster, 1998.

[97] J. LeDoux. Rethinking the emotional brain. *Neuron* 73.4 (2012), pp. 653–676.

[98] M. C. Nussbaum. The Monarchy of Fear: A Philosopher Looks at Our Political Crisis. New York, NY: Simon & Schuster, 2019.

[99] J. Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological review* 108.4 (2001), p. 814.

[100] S. D. Kreibig. Autonomic nervous system activity in emotion: A review. *Biological Psychology* 84.3 (2010), pp. 394–421.

[101] H. A. Simon. Motivational and emotional controls of cognition. *Psychological review* 74.1 (1967), p. 29.

[102] B. J. Ellis and D. 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 (2012), p. 591.

[103] J. Dewey. The theory of emotion. *Psychological review* 2.1 (1895), p. 13.

[104] A. Birhane. Algorithmic injustice: A relational ethics approach. *Patterns* 2.2 (2021).

[105] L. F. Barrett. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review* 10.1 (2006), pp. 20–46.

[106] R. Reisenzein. Wundt's Three-Dimensional Theory of Emotion. Structuralist Knowledge Representation. Ed. by W. Balzer, F. Stadler, and P. Wessels. Leiden, Netherlands: Brill, 2000, pp. 219–250.

[107] W. M. Wundt and C. H. Judd. Outlines of Psychology. Leipzig, Germany: W. Engelmann, 1902.

[108] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology* 39.6 (1980), p. 1161.

[109] U. Schimmack and A. Grob. Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality* 14.4 (2000), pp. 325–345.

[110] E. Cambria, A. Livingstone, and A. Hussain. The Hourglass of Emotions. Cognitive Behavioural Systems. Ed. by A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller. Vol. 7403. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012, pp. 144–157. DOI: 10.1007/978-3-642-34584-5_11.

[111] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological review* 110.1 (2003), p. 145.

[112] C. Crivelli and A. J. Fridlund. Facial displays are tools for social influence. *Trends in Cognitive Sciences* 22.5 (2018), pp. 388–399.

[113] E. Brunswik. Representative design and probabilistic theory in a functional psychology. *Psychological review* 62.3 (1955), p. 193.

[114] K. Boehner, R. DePaula, P. Dourish, and P. Sengers. How emotion is made and measured. *International Journal of Human-Computer Studies* 65.4 (2007), pp. 275–291.

[115] J. J. Prinz. Gut Reactions: A Perceptual Theory of the Emotions. Oxford University Press, 2004.

[116] K. Boehner, R. DePaula, P. Dourish, and P. Sengers. Affect: From Information to Interaction. Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility. New York, NY, USA: ACM, 2005, pp. 59–68. DOI: 10.1145/1094562.1094570.

[117] K. Crawford. Time to regulate AI that interprets human emotions. *Nature* 592.7853 (2021), pp. 167–167.

[118] R. W. Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cavallo, T. Machover, M. Resnick, D. Roy, and C. Strohecker. Affective learning — A manifesto. *BT Technology Journal* 22.4 (2004), pp. 253–269. DOI: 10.1023/B:BTTJ.0000047603.37042.33.

[119] P. Sengers, K. Boehner, M. Mateas, and G. Gay. The disenchantment of affect. *Personal and Ubiquitous Computing* 12.5 (2008), pp. 347–358.

[120] D. Schuller and B. W. Schuller. The Age of Artificial Emotional Intelligence. *Computer* 51.9 (2018), pp. 38–46. DOI: 10.1109/MC.2018.3620963.

[121] Y. Song, P. H. Tung, and B. Jeon. Trends in Artificial Emotional Intelligence Technology and Application. 2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD). IEEE. 2022, pp. 366–370.

[122] E. M. G. Younis, S. Mohsen, E. H. Houssein, and O. A. S. Ibrahim. Machine learning for human emotion recognition: A comprehensive review. *Neural Computing and Applications* 36.16 (2024), pp. 8901–8947. DOI: 10.1007/s00521-024-09426-2.

[123] J. Hernandez, J. Lovejoy, D. McDuff, J. Suh, T. O'Brien, A. Sethumadhavan, G. Greene, R. Picard, and M. Czerwinski. Guidelines for Assessing and Minimizing Risks of Emotion Recognition Applications. 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2021, pp. 1–8. DOI: 10.1109/ACII52823.2021.9597452.

[124] A. Landowska, A. Karpus, T. Zawadzka, B. Robins, D. Erol Barkana, H. Kose, T. Zorcec, and N. Cummins. Automatic emotion recognition in children with autism: A systematic literature review. *Sensors* 22.4 (2022), p. 1649.

[125] T. Eerola and J. K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music* 39.1 (2011), pp. 18–49.

[126] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems* 28.2 (2013), pp. 15–21.

[127] M. L. B. Estrada, R. Z. Cabada, R. O. Bustillos, and M. Graff. Opinion mining and emotion recognition applied to learning environments. *Expert Systems with Applications* 150 (2020), p. 113265.

[128] K. Crawford, R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A. N. Sánchez, D. Raji, J. L. Rankin, R. Richardson, J. Schultz, S. M. West, and M. Whittaker. AI Now 2019 Report. Tech. rep. New York, NY: AI Now Institute, 2019, p. 100.

[129] F. A. Pasquale. More than a feeling. *Real Life* (2020).

[130] H. Kaur, A. C. Williams, D. McDuff, M. Czerwinski, J. Teevan, and S. T. Iqbal. Optimizing for Happiness and Productivity: Modeling Opportune Moments for Transitions and Breaks at Work. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–15. DOI: 10.1145/3313831.3376817.

[131] T. Xue, S. Ghosh, G. Ding, A. El Ali, and P. Cesar. Designing Real-Time, Continuous Emotion Annotation Techniques for 360 VR Videos. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–9. DOI: 10.1145/3334480.3382895.

[132] A. McStay and G. Rosner. Emotional artificial intelligence in children's toys and devices: Ethics, governance and practical remedies. *Big Data & Society* 8.1 (2021), pp. 1–16. DOI: 10.1177/2053951721994877.

[133] S. Zuboff. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York, NY: PublicAffairs, 2019.

[134] C. Wylie. Mindf*ck: Cambridge Analytica and the Plot to Break America. New York, NY: Random House, 2019.

[135] T. Z. Zarsky. Privacy and manipulation in the digital age. *Theoretical Inquiries in Law* 20.1 (2019), pp. 157–188. DOI: 10.1515/til-2019-0006.

[136] D. Susser, B. Roessler, and H. Nissenbaum. Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review* 4.1 (2019), pp. 1–45.

[137] M. Ienca. On artificial intelligence and manipulation. *Topoi* 42.3 (2023), pp. 833–842. DOI: 10.1007/s11245-023-09940-3.

[138] A. Patwardhan and G. Knapp. Augmenting supervised emotion recognition with rule-based decision model. 2016. arXiv: 1607.02660.

[139] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Barcelona, Spain: Association for Computing Machinery, 2020, pp. 33–44. DOI: 10.1145/3351095.3372873.

[140] A. Balagopalan, H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, and M. Ghassemi. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. 2022. arXiv: 2205.03295.

[141] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing* 11.8 (2017), pp. 1301–1309.

[142] J. Zhang, Z. Yin, P. Chen, and S. Nichele. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion* 59 (2020), pp. 103–126.

[143] V. C. C. Roza and O. A. Postolache. Multimodal approach for emotion recognition based on simulated flight experiments. *Sensors* 19.24 (2019). DOI: 10.3390/s19245516.

[144] H. Zhang, Y. Luo, Q. Ai, Y. Wen, and H. Hu. Look, Read and Feel: Benchmarking Ads Understanding with Multimodal Multitask Learning. Proceedings of the 28th ACM International Conference on Multimedia (MM '20). New York, NY, USA: Association for Computing Machinery, 2020, pp. 430–438. DOI: 10.1145/3394171.3413582.

[145] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. 2019. arXiv: 1905.05812.

[146] H. Zhang, C. Zhao, Y. Zhang, D. Wang, and H. Yang. Deep Latent Emotion Network for Multi-Task Learning. 2021. arXiv: 2104.08716.

[147] S. Atakishiyev, M. Salameh, and R. Goebel. Safety implications of explainable artificial intelligence in end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* (2025). DOI: 10.1109/TITS.2025.3574738.

[148] A. Reuel, B. Bucknall, S. Casper, T. Fist, L. Soder, O. Aarne, L. Hammond, L. Ibrahim, A. Chan, P. Wills, M. Anderljung, B. Garfinkel, L. Heim, A. Trask, G. Mukobi, R. Schaeffer, M. Baker, S. Hooker, I. Solaiman, A. S. Luccioni, N. Rajkumar, N. Moës, J. Ladish, D. Bau, P. Bricman, N. Guha, J. Newman, Y. Bengio, T. South, A. Pentland, S. Koyejo, M. J. Kochenderfer, and R. Trager. Open problems in technical AI governance. 2024. DOI: 10.48550/arXiv.2407.14981.

[149] C. Tarsney. Deception and Manipulation in Generative AI. *Philosophical Studies* 182.7 (2025), pp. 1865–1887. DOI: 10.1007/s11098-024-02259-8.

[150] E. Anderson. Private Government: How Employers Rule Our Lives (and Why We Don't Talk About It). University Center for Human Values Series. Princeton; Oxford: Princeton University Press, 2017.

[151] M. K. Sharma, N. John, and M. Sahu. Influence of social media on mental health: A systematic review. *Current Opinion in Psychiatry* 33.5 (2020), pp. 467–475. DOI: 10.1097/YCO.0000000000000631.

[152] R. Meyer. Everything We Know About Facebook's Secret Mood Manipulation Experiment. The Atlantic. 2014. URL: https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/.

[153] E. Selinger and W. 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 (2016), pp. 35–43. DOI: 10.1177/1747016115579531.

[154] E. Hu. Facebook Manipulates Our Moods for Science and Commerce: A Roundup. NPR. 2014. URL: https://www.npr.org/sections/alltechconsidered/2014/06/30/326929138/facebook-manipulates-our-moods-for-science-and-commerce-a-roundup.

[155] B. Hallinan, J. R. Brubaker, and C. Fiesler. Unexpected expectations: Public reaction to the Facebook emotional contagion study. *New Media & Society* 22.6 (2020), pp. 1076–1094. DOI: 10.1177/1461444819876944.

[156] Gartner projections for 2018. *Database and Network Journal* 48.2 (2018), p. 10.

[157] J. Penni. The future of online social networks (OSN): A measurement analysis using social media tools and application. *Telematics and Informatics* 34.5 (2017), pp. 498–517. DOI: 10.1016/j.tele.2016.10.009.

[158] T. Glenn and S. Monteith. Privacy in the Digital World: Medical and Health Data Outside of HIPAA Protections. *Current Psychiatry Reports* 16.11 (2014), p. 494. DOI: 10.1007/s11920-014-0494-4.

[159] V. K. Singh and R. R. Agarwal. Cooperative Phoneotypes: Exploring Phone-Based Behavioral Markers of Cooperation. Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 646–657. DOI: 10.1145/2971648.2971755.

[160] J. Torous, M. V. Kiang, J. Lorme, and J.-P. Onnela. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Mental Health* 3.2 (2016). DOI: 10.2196/mental.5165.

[161] T. Aledavood, A. M. Triana Hoyos, T. Alakörkkö, K. Kaski, J. Saramäki, E. Isometsä, and R. K. Darst. Data collection for mental health studies through digital platforms: Requirements and design of a prototype. *JMIR Research Protocols* 6.6 (2017), e110. DOI: 10.2196/resprot.6919.

[162] G. Coppersmith, M. Dredze, and C. Harman. Quantifying Mental Health Signals in Twitter. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014, pp. 51–60. DOI: 10.3115/v1/W14-3207.

[163] D. Bhugra, A. Tasman, S. Pathare, S. Priebe, S. Smith, J. Torous, M. R. Arbuckle, A. Langford, R. D. Alarcón, H. F. K. Chiu, M. B. First, J. Kay, C. Sunkel, A. Thapar, P. Udomratn, F. K. Baingana, D. Kestel, R. M. K. Ng, A. Patel, L. D. Picker, K. J. McKenzie, D. Moussaoui, M. Muijen, P. Bartlett, S. Davison, T. Exworthy, N. Loza, D. Rose, J. Torales, M. Brown, H. Christensen, J. Firth, M. Keshavan, A. Li, J.-P. Onnela, T. Wykes, H. Elkholy, G. Kalra, K. F. Lovett, M. J. Travis, and A. Ventriglio. The WPA-Lancet Psychiatry Commission on the Future of Psychiatry. *The Lancet Psychiatry* 4.10 (2017), pp. 775–818. DOI: 10.1016/S2215-0366(17)30333-4.

[164] A. G. Reece and C. M. Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6.1 (2017), pp. 1–12. DOI: 10.1140/epjds/s13688-017-0110-z.

[165] H. Christensen, P. J. Batterham, and B. O'Dea. E-Health Interventions for Suicide Prevention. *International Journal of Environmental Research and Public Health* 11.8 (2014), pp. 8193–8212. DOI: 10.3390/ijerph110808193.

[166] J. Mikal, S. Hurst, and M. Conway. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. *BMC medical ethics* 17.1 (2016), p. 22.

[167] M. Marks. Artificial intelligence-based suicide prediction. *Yale Journal of Law and Technology* 21.3 (2019), pp. 98–121.

[168] S. Chancellor, E. P. S. Baumer, and M. 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 (2019), 147:1–147:32. DOI: 10.1145/3359249.

[169] N. Singer. In screening for suicide risk, Facebook takes on tricky public health role. The New York Times. 2018. URL: https://www.nytimes.com/2018/12/31/technology/facebook-suicide-screening-algorithm.html.

[170] A. Benton, G. Coppersmith, and M. Dredze. Ethical Research Protocols for Social Media Health Research. Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 94–102. DOI: 10.18653/v1/W17-1612.

[171] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. B. Silenzio, and M. De Choudhury. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 79–88. DOI: 10.1145/3287560.3287587.

[172] M. Conway and D. O'Connor. Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology* 9 (2016), pp. 77–82. DOI: 10.1016/j.copsyc.2016.01.004.

[173] H. AlBalooshi, S. Rahmanian, and R. V. Kumar. Detecting User Emotions in Social Media Text. Proceedings of the Sixth Workshop on Natural Language Processing for Social Media (SocialNLP 2018). Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 45–49. DOI: 10.18653/v1/W18-3508.

[174] M. De Choudhury, S. Counts, and M. Gamon. Not All Moods Are Created Equal! Exploring Human Emotional States in Social Media. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM). AAAI Press, 2012, pp. 66–73.

[175] M. De Choudhury, M. Gamon, and S. Counts. Happy, Nervous or Surprised? Classification of Human Affective States in Social Media. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM). AAAI Press, 2012, pp. 435–438.

[176] F. Kivran-Swaine, J. Ting, J. R. Brubaker, R. Teodoro, and M. Naaman. Understanding Loneliness in Social Awareness Streams: Expressions and Responses. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM). AAAI Press, 2014, pp. 256–265.

[177] L. Manikonda and M. De Choudhury. Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. CHI '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 170–181. doi: 10.1145/3025453.3025932.

[178] G. 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 (2009), e11. doi: 10.2196/jmir.1157.

[179] F. Sadeque, D. Xu, and S. Bethard. Measuring the Latency of Depression Detection in Social Media. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. WSDM '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 495–503. doi: 10.1145/3159652.3159725.

[180] C. L. Hanson, S. H. Burton, C. Giraud-Carrier, J. H. West, M. D. Barnes, and B. Hansen. Tweaking and tweeting: Exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. *Journal of Medical Internet Research* 15.4 (2013). doi: 10.2196/jmir.2503.

[181] M. Myslín, S.-H. Zhu, W. Chapman, and M. Conway. Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products. *Journal of Medical Internet Research* 15.8 (2013), e174. doi: 10.2196/jmir.2534.

[182] Z. Tan, X. Liu, X. Liu, Q. Cheng, and T. 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 (2017), e381. doi: 10.2196/jmir.8729.

[183] D. Muriello, L. Donahue, D. Ben-David, U. Ozertem, and R. Shilon. Under the Hood: Suicide Prevention Tools Powered by AI. Facebook Engineering. 2018. url: https://engineering.fb.com/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/.

[184] N. N. Gomes de Andrade, D. Pawson, D. Muriello, L. Donahue, and J. Guadagno. Ethics and artificial intelligence: Suicide prevention on Facebook. *Philosophy & Technology* 31.4 (2018), pp. 669–684. doi: 10.1007/s13347-018-0336-0.

[185] M. L. Birnbaum, S. K. Ernala, A. F. Rizvi, M. De Choudhury, and J. 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 (2017), e289. doi: 10.2196/jmir.7956.

[186] M. Mitchell, K. Hollingshead, and G. Coppersmith. Quantifying the Language of Schizophrenia in Social Media. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Denver, Colorado: Association for Computational Linguistics, 2015, pp. 11–20. DOI: 10.3115/v1/W15-1202.

[187] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting Depression via Social Media. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM). Cambridge, MA, USA: AAAI Press, 2013, pp. 128–137. DOI: 10.1609/icwsm.v7i1.14432.

[188] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki. Recognizing Depression from Twitter Activity. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 3187–3196. DOI: 10.1145/2702123.2702280.

[189] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff. Characterizing and Predicting Postpartum Depression from Shared Facebook Data. Proceedings of the 17th ACM Conference on Computer-Supported Cooperative Work & Social Computing. CSCW '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 626–638. DOI: 10.1145/2531602.2531675.

[190] G. A. Coppersmith, C. T. Harman, and M. Dredze. Measuring Post Traumatic Stress Disorder in Twitter. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM). Ann Arbor, MI, USA: AAAI Press, 2014, pp. 579–582. DOI: 10.1609/icwsm.v8i1.14574.

[191] L. Eggertson. Social media embraces suicide prevention. *CMAJ* 187.11 (2015), E333–E333. DOI: 10.1503/cmaj.109-5104.

[192] A. L. Berman and G. Carter. Technological advances and the future of suicide prevention: Ethical, legal, and empirical challenges. *Suicide and Life-Threatening Behavior* 50.3 (2020), pp. 643–651. DOI: 10.1111/sltb.12610.

[193] N. Singer. Creepy or not? Your privacy concerns probably reflect your politics. The New York Times. 2018. URL: https://www.nytimes.com/2018/04/30/technology/privacy-concerns-politics.html.

[194] C. Newton. How Facebook is preparing for a surge in depressed and anxious users. The Verge. 2020. URL: https://www.theverge.com/2020/3/19/21185204/facebook-coronavirus-depression-anxiety-content-moderation-mark-zuckerberg-interview.

[195] K. McVeigh. Samaritans Twitter app identifying user's moods criticised as invasive. The Guardian. 2014. URL: https://www.theguardian.com/society/2014/nov/04/samaritans-twitter-app-mental-health-depression.

[196] J. Lopez-Castroman, B. Moulahi, J. Aze, S. Bringay, J. Deninotti, S. Guillaume, and E. Baca-Garcia. Mining social networks to improve suicide prevention: A scoping review. *Journal of Neuroscience Research* 98.4 (2020), pp. 616–625. DOI: 10.1002/jnr.24404.

[197] I. Barnett and J. Torous. Ethics, transparency, and public health at the intersection of innovation and Facebook's suicide prevention efforts. *Annals of Internal Medicine* 170.8 (2019), pp. 565–566. DOI: 10.7326/M19-0366.

[198] T. Gillespie. The Relevance of Algorithms. Media Technologies: Essays on Communication, Materiality, and Society. Ed. by T. Gillespie, P. J. Boczkowski, and K. A. Foot. Cambridge, MA: MIT Press, 2014, pp. 167–194. DOI: 10.7551/mitpress/9780262525374.003.0009.

[199] M. Haim, F. Arendt, and S. Scherr. Abyss or shelter? On the relevance of web search engines' search results when people Google for suicide. *Health Communication* 32.2 (2017), pp. 253–258. DOI: 10.1080/10410236.2015.1113484.

[200] O. J. Kirtley and R. C. O'Connor. Suicide prevention is everyone's business: Challenges and opportunities for Google. *Social Science & Medicine* 262 (2020). DOI: 10.1016/j.socscimed.2019.112691.

[201] E. Ford, K. Curlewis, A. Wongkoblap, and V. 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 (2019), e12942. DOI: 10.2196/12942.

[202] K. L. Costello and D. Floegel. "Predictive ads are not doctors": Mental health tracking and technology companies. *Proceedings of the Association for Information Science and Technology* 57.1 (2020), e250.

[203] T. Ammari, S. Schoenebeck, and M. Morris. Accessing Social Support and Overcoming Judgment on Social Media among Parents of Children with Special Needs. 2014.

[204] J. R. Brubaker, L. S. Dombrowski, A. M. Gilbert, N. Kusumakaulika, and G. R. Hayes. Stewarding a legacy: responsibilities and relationships in the management of post-mortem data. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 4157–4166. DOI: 10.1145/2556288.2557059.

[205] J. M. Carroll. Five reasons for scenario-based design. *Interacting with Computers* 13.1 (2000), pp. 43–60. DOI: 10.1016/S0953-5438(00)00023-0.

[206] A. C. High, A. Oeldorf-Hirsch, and S. Bellur. Misery rarely gets company: The influence of emotional bandwidth on supportive communication on Facebook. *Computers in Human Behavior* 34 (2014), pp. 79–88. DOI: 10.1016/j.chb.2014.01.037.

[207] R. Y. Wong, E. Van Wyk, and J. Pierce. Real-Fictional Entanglements: Using Science Fiction and Design Fiction to Interrogate Sensing Technologies. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). 2017, pp. 4032–4043. DOI: 10.1145/3064663.3064682.

[208] J. Finch. The vignette technique in survey research. *Sociology* 21.1 (1987), pp. 105–114.

[209] R. 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 (1998), pp. 381–400. DOI: 10.1111/1467-9566.00107.

[210] M. Eslami, A. Rickman, K. Vaccaro, A. Aleyasen, A. Vuong, K. Karahalios, K. Hamilton, and C. Sandvig. "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 153–162. DOI: 10.1145/2702123.2702556.

[211] R. Y. Wong, D. K. Mulligan, and J. Chuang. Using Science Fiction Texts to Surface User Reflections on Privacy. 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. New York, NY, USA: Association for Computing Machinery, 2017, pp. 213–216. DOI: 10.1145/3123024.3123080.

[212] N. Andalibi. Disclosure, Privacy, and Stigma on Social Media: Examining Non-Disclosure of Distressing Experiences. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27.3 (2020). DOI: 10.1145/3386600.

[213] J. Corbin and A. Strauss. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. 3rd ed. Thousand Oaks, CA: SAGE Publications, 2008. DOI: 10.4135/9781452230153.

[214] N. Andalibi and J. Buss. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–16. DOI: 10.1145/3313831.3376680.

[215] G. Grill and N. 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 (2022), pp. 1–35.

[216] I. Seidman. Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences. 5th ed. New York, NY: Teachers College Press, 2019.

[217] T. Ammari, J. Kaye, J. Y. Tsai, and F. Bentley. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Transactions on Computer-Human Interaction* 26.3 (2019), 17:1–17:28. DOI: 10.1145/3311956.

[218] F. Hamidi, M. K. Scheuerman, and S. M. Branham. Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18). Montreal, QC, Canada: Association for Computing Machinery, 2018, pp. 1–13. DOI: 10.1145/3173574.3173582.

[219] J. L. Feuston and A. M. Piper. Everyday Experiences: Small Stories and Mental Illness on Instagram. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI ’19. Glasgow, Scotland, UK: Association for Computing Machinery, 2019. DOI: 10.1145/3290605.3300495.

[220] J. L. Feuston and A. M. Piper. Beyond the Coded Gaze: Analyzing Expression of Mental Health and Illness on Instagram. *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–21. DOI: 10.1145/3274320.

[221] J. L. Feuston, A. S. Taylor, and A. M. Piper. Conformity of Eating Disorders through Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1 (2020), pp. 1–28. DOI: 10.1145/3392845.

[222] S. Chancellor, J. A. Pater, T. Clear, E. Gilbert, and M. De Choudhury. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. Proceedings of the 19th ACM Conference on Computer-Supported

Cooperative Work & Social Computing. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1201–1213. DOI: 10.1145/2818048.2819963.

[223] J. A. Pater, O. L. Haimson, N. Andalibi, and E. D. Mynatt. "Hunger Hurts but Starving Works": Characterizing the Presentation of Eating Disorders Online. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. CSCW '16. San Francisco, CA, USA: Association for Computing Machinery, 2016, pp. 1185–1200. DOI: 10.1145/2818048.2820030.

[224] E. M. Lawrence. Why Do College Graduates Behave More Healthfully than Those Who Are Less Educated? *Journal of Health and Social Behavior* 58.3 (2017), pp. 291–306. DOI: 10.1177/0022146517715671.

[225] C. Fiesler and N. Proferes. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4.1 (2018). DOI: 10.1177/2056305118763366.

[226] G. Omale. Gartner Identifies Three Most Common AI Use Cases in HR and Recruiting. Gartner. 2019. URL: https://www.gartner.com/en/newsroom/press-releases/2019-06-19-gartner-identifies-three-most-common-ai-use-cases-in-hr-and-recruiting.

[227] S. Buranyi. 'Dehumanising, impenetrable, frustrating': The grim reality of job hunting in the age of AI. The Guardian. 2018. URL: https://www.theguardian.com/inequality/2018/mar/04/dehumanising-impenetrable-frustrating-the-grim-reality-of-job-hunting-in-the-age-of-ai.

[228] L. Winner. Do artifacts have politics? *Daedalus* 109.1 (1980), pp. 121–136.

[229] K. Shilton. Values and ethics in human–computer interaction. *Foundations and Trends in Human-Computer Interaction* 12.2 (2018), pp. 107–171. DOI: 10.1561/1100000073.

[230] C. A. Le Dantec, E. S. Poole, and S. P. Wyche. Values as Lived Experience: Evolving Value Sensitive Design in Support of Value Discovery. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '09. Boston, MA, USA: Association for Computing Machinery, 2009, pp. 1141–1150. DOI: 10.1145/1518701.1518875.

[231] K. Shilton, J. A. Koepfler, and K. R. Fleischmann. Charting sociotechnical dimensions of values for design research. *The Information Society* 29.5 (2013), pp. 259–271. DOI: 10.1080/01972243.2013.825357.

[232] L. A. Rivera. Go with your gut: Emotion and evaluation in job interviews. *American Journal of Sociology* 120.5 (2015), pp. 1339–1389. DOI: 10.1086/681214.

[233] E. H. Gorman. Gender stereotypes, same-gender preferences, and organizational variation in the hiring of women: Evidence from law firms. *American Sociological Review* 70.4 (2005), pp. 702–728. DOI: 10.1177/000312240507000408.

[234] D. Pager. The mark of a criminal record. *American Journal of Sociology* 108.5 (2003), pp. 937–975. DOI: 10.1086/374403.

[235] G. Farkas. Cognitive skills and noncognitive traits and behaviors in stratification processes. *Annual Review of Sociology* 29.1 (2003), pp. 541–562. DOI: 10.1146/annurev.soc.29.010202.100023.

[236] E. Faliagka, K. Ramantas, A. Tsakalidis, and G. Tzimas. Application of Machine Learning Algorithms to an Online Recruitment System. Proceedings of the Seventh International Conference on Internet and Web Applications and Services (ICIW 2012). Stuttgart, Germany: IARIA Press, 2012, pp. 215–220.

[237] I. Ajunwa. Automated video interviewing as the new phrenology. *Berkeley Technology Law Journal* 36.3 (2021), pp. 1173–1226. DOI: 10.15779/Z38RX93F1Q.

[238] I. Ajunwa and D. Greene. Platforms at Work: Automated Hiring Platforms and Other New Intermediaries in the Organization of Work. Work and Labor in the Digital Age. Research in the Sociology of Work. Bingley, UK: Emerald Publishing Limited, 2019, pp. 61–91. DOI: 10.1108/S0277-283320190000033005.

[239] L. Dencik and S. Stevens. Regimes of justification in the datafied workplace: The case of hiring. *New Media & Society* 25.12 (2021), pp. 3657–3675. DOI: 10.1177/14614448211052893.

[240] S. Skinner-Thompson. Privacy at the Margins. Cambridge, United Kingdom; New York, NY, USA: Cambridge University Press, 2020. DOI: 10.1017/9781316850350.

[241] L. Stark, A. Stanhaus, and D. 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 (2020), pp. 1074–1088. DOI: https://doi.org/10.1002/asi.24342.

[242] M. Fourcade and K. Healy. Seeing like a market. *Socio-Economic Review* 15.1 (2017), pp. 9–29. DOI: 10.1093/ser/mww033.

[243] L. L. Koppes, ed. Historical Perspectives in Industrial and Organizational Psychology. New York, NY, USA: Psychology Press, 2007. DOI: 10.4324/9781315820972.

[244] S. M. Jacoby. Employee Attitude Surveys in Historical Perspective. *Industrial Relations: A Journal of Economy and Society* 27.1 (1988), pp. 74–93. DOI: 10.1111/j.1468-232X.1988.tb01047.x.

[245] M. L. Gross. The Brain Watchers. New York, NY, USA: Signet Books, 1962.

[246] S. E. Igo. The Known Citizen: A History of Privacy in Modern America. Cambridge, MA, USA: Harvard University Press, 2018.

[247] D. D. Steiner. Personnel Selection across the Globe. The Oxford Handbook of Personnel Assessment and Selection. Ed. by N. Schmitt. New York, NY, USA: Oxford University Press, 2012, pp. 740–767. DOI: 10.1093/oxfordhb/9780199732579.013.0032.

[248] T. M. S. Neal, C. Slobogin, M. J. Saks, D. L. Faigman, and K. F. Geisinger. Psychological assessments in legal contexts: Are courts keeping “junk science” out of the courtroom? *Psychological Science in the Public Interest* 20.3 (2019), pp. 135–164. DOI: 10.1177/1529100619888860.

[249] H. P. Assocation. Request for Information (RFI) on Public and Private Sector Uses of Biometric Technologies: Responses. 2022.

[250] B. Dattner, T. Chamorro-Premuzic, R. Buchband, and L. Schettler. The Legal and Ethical Implications of Using AI in Hiring. Harvard Business Review. 2019. URL: https://hbr.org/2019/04/the-legal-and-ethical-implications-of-using-ai-in-hiring.

[251] J. Dastin. Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. Reuters. 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[252] I. Žliobaitė. Measuring Discrimination in Algorithmic Decision Making. *Data Mining and Knowledge Discovery* 31.4 (2017), pp. 1060–1089. DOI: 10.1007/s10618-017-0506-1.

[253] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. FAT* 2019 – Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. FAT* 2019. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 329–338. DOI: 10.1145/3287560.3287589.

[254] C. Kuhlman, L. Jackson, and R. Chunara. No Computation Without Representation: Avoiding Data and Algorithm Biases through Diversity. 2020. arXiv: 2002.11836.

[255] B. Fish and L. Stark. Reflexive Design for Fairness and Other Human Values in Formal Models. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’21). AIES ’21. Virtual Event, USA: Association for Computing Machinery and AAAI, 2021, pp. 89–99. DOI: 10.1145/3461702.3462518.

[256] P.-H. Wong. Democratizing Algorithmic Fairness. *Philosophy & Technology* 33.2 (2020), pp. 225–244. DOI: 10.1007/s13347-019-00355-w.

[257] M. Raghavan, S. Barocas, J. M. Kleinberg, and K. Levy. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 469–481. DOI: 10.1145/3351095.3372828.

[258] N. T. Lee. Detecting Racial Bias in Algorithms and Machine Learning. *Journal of Information, Communication and Ethics in Society* 16.3 (2018), pp. 252–260. DOI: 10.1108/JICES-06-2018-0056.

[259] K. Nakamura. My Algorithms Have Determined You're Not Human: AI-ML, Reverse Turing-Tests, and the Disability Experience. The 21st International ACM SIGACCESS Conference on Computers and Accessibility. ASSETS '19. Pittsburgh, PA, USA: Association for Computing Machinery, 2019, pp. 1–2. DOI: 10.1145/3308561.3353812.

[260] J. Sánchez-Monedero, L. Dencik, and L. 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. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 458–468. DOI: 10.1145/3351095.3372849.

[261] L. Stark and J. Hutson. Physiognomic Artificial Intelligence. *Fordham Intellectual Property, Media & Entertainment Law Journal* 32.4 (2022), pp. 922–978. DOI: 10.2139/ssrn.3927300.

[262] C. 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 (2021), pp. 26–29.

[263] N. Sampath. CR's Comments to the Office of Science and Technology Policy on AI-Enabled Biometric Processing. Consumer Reports Advocacy. 2022. URL: https://advocacy.consumerreports.org/research/crs-comments-to-the-office-of-science-and-technology-policy-on-ai-enabled-biometric-processing/.

[264] K. Shilton, J. A. Koepfler, and K. R. Fleischmann. How to See Values in Social Computing: Methods for Studying Values Dimensions. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14). CSCW '14. Baltimore, MD, USA: Association for Computing Machinery, 2014, pp. 426–435. DOI: 10.1145/2531602.2531625.

[265] C. P. Knobel and G. C. Bowker. Values in Design. *Communications of the ACM* 54.7 (2011), pp. 26–28. DOI: 10.1145/1965724.1965735.

[266] M. Hoffmann and M. Mariniello. Biometric Technologies at Work: A Proposed Use-Based Taxonomy. Bruegel. 2021. URL: https://www.bruegel.org/policy-brief/biometric-technologies-work-proposed-use-based-taxonomy.

[267] M. Flanagan and H. Nissenbaum. Values at Play in Digital Games. Cambridge, MA: MIT Press, 2014.

[268] R. Hilpinen. Artifact. The Stanford Encyclopedia of Philosophy. Ed. by E. N. Zalta. Stanford, CA: Metaphysics Research Lab, Stanford University, 2012.

[269] P.-P. Verbeek and P. E. Vermaas. Technological Artifacts. A Companion to the Philosophy of Technology. Ed. by J. K. Berg Olsen, S. A. Pedersen, and V. F. Hendricks. Wiley-Blackwell, 2009, pp. 165–171. DOI: 10.1002/9781444310795.ch28.

[270] C. Kluckhohn. Values and Value-Orientations in the Theory of Action: An Exploration in Definition and Classification. Toward a General Theory of Action. Ed. by T. Parsons and E. A. Shils. Cambridge, MA, USA: Harvard University Press, 1951.

[271] A. E. Clarke. Situational Analyses: Grounded Theory Mapping After the Postmodern Turn. *Symbolic Interaction* 26.4 (2003), pp. 553–576. DOI: 10.1525/si.2003.26.4.553.

[272] S. Bardzell. Feminist HCI: Taking Stock and Outlining an Agenda for Design. Proceedings of the 28th ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '10). CHI '10. Atlanta, GA, USA: Association for Computing Machinery, 2010, pp. 1301–1310. DOI: 10.1145/1753326.1753521.

[273] N. McDonald, S. Schoenebeck, and A. 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 (2019), pp. 1–23. DOI: 10.1145/3359174.

[274] G. C. Bowker and S. L. Star. Sorting Things Out: Classification and Its Consequences. Cambridge, MA, USA: MIT Press, 1999.

[275] A. E. Clarke and K. Charmaz, eds. Grounded Theory and Situational Analysis. SAGE Benchmarks in Social Research Methods. Los Angeles, CA, USA: SAGE Publications, 2014.

[276] K. Charmaz and R. G. Mitchell. Grounded Theory in Ethnography. Handbook of Ethnography. Ed. by P. Atkinson, A. Coffey, S. Delamont, J. Lofland, and L. Lofland. London, UK: SAGE Publications, 2001, pp. 160–174. DOI: 10.4135/9781848608337.n11.

[277] R. Lynn. The intelligence of the Mongoloids: A psychometric, evolutionary and neurological theory. *Personality and Individual Differences* 8.6 (1987), pp. 813–844. DOI: 10.1016/0191-8869(87)90135-8.

[278] J. Rust and S. Golombok. Modern Psychometrics: The Science of Psychological Assessment. 3rd ed. London, UK: Routledge, 2014. DOI: 10.4324/9781315787527.

[279] Electronic Privacy Information Center. HireVue, Facing FTC Complaint From EPIC, Halts Use of Facial Recognition. 2021. URL: https://epic.org/hirevue-facing-ftc-complaint-from-epic-halts-use-of-facial-recognition/.

[280] S. Mhlambi. From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance. Carr Center for Human Rights Policy, Harvard Kennedy School. 2020. URL: https://cyber.harvard.edu/story/2020-07/rationality-relationality-ubuntu-ethical-and-human-rights-framework-artificial.

[281] L. Gitelman, ed. Raw Data Is an Oxymoron. Cambridge, MA, USA: MIT Press, 2013. DOI: 10.7551/mitpress/9302.001.0001.

[282] J. E. H. Smith. Irrationality: A History of the Dark Side of Reason. Princeton, NJ, USA: Princeton University Press, 2019. DOI: 10.1515/9780691210827.

[283] R. Benjamin. Race after Technology: Abolitionist Tools for the New Jim Code. Medford, MA, USA: Polity, 2019.

[284] R. Descartes. The Philosophical Writings of Descartes. Ed. by J. Cottingham, R. Stoothoff, and D. Murdoch. Vol. 2. Cambridge, UK: Cambridge University Press, 1984. DOI: 10.1017/CBO9780511805042.

[285] W. Hartzog. The Inadequate, Invaluable Fair Information Practices. *Maryland Law Review* 76.4 (2017), pp. 952–983.

[286] D. K. Mulligan, D. Kluttz, and N. Kohli. Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions. Carr Center for Human Rights Policy / SSRN. 2019. URL: https://ssrn.com/abstract=3311894.

[287] Federal Reserve. Federal Trade Commission Act Section 5: Unfair or Deceptive Acts or Practices. Board of Governors of the Federal Reserve System. 2008. URL: https://www.federalreserve.gov/boarddocs/supmanual/cch/ftca.pdf.

[288] K. Lyons. New FTC Memo Calls for a Focus on ‘Structural Dominance’ from Big Companies. The Verge. 2021. URL: https://www.theverge.com/2021/9/23/22690176/ftc-chair-lina-khan-focus-antitrust-consumer-amazon.

[289] P. Bourdieu. The Forms of Capital. Handbook of Theory and Research for the Sociology of Education. Ed. by J. G. Richardson. Westport, CT, USA: Greenwood Press, 1986, pp. 241–258.

[290] S. E. Nugent and S. Scott-Parker. Recruitment AI Has a Disability Problem: Anticipating and Mitigating Unfair Automated Hiring Decisions. Towards Trustworthy Artificial Intelligent Systems. Cham, Switzerland: Springer, 2022, pp. 85–96. DOI: 10.1007/978-3-031-09823-9_6.

[291] K. Martin. Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics* 160.4 (2019), pp. 835–850. DOI: 10.1007/s10551-018-3921-3.

[292] K. Zickuhr. Workplace Surveillance Is Becoming the New Normal for US Workers. Washington Center for Equitable Growth. 2021. URL: https://equitablegrowth.org/research-paper/workplace-surveillance-is-becoming-the-new-normal-for-u-s-workers/.

[293] G. Tay and B. Elliot. Maverick* Research: Emotion AI Will Become You Without Your Knowledge. Gartner. 2019. URL: https://www.gartner.com/en/documents/3975557-maverick-research-emotion-ai-will-become-you-without-your-knowledge.

[294] S. Bromuri, A. P. Henkel, D. Iren, and V. Urovi. Using AI to Predict Service Agent Stress from Emotion Patterns in Service Interactions. *Journal of Service Management* 31.4 (2020), pp. 785–812. DOI: 10.1108/JOSM-06-2019-0163.

[295] E. Shaw, M. Payri, M. Cohn, and I. R. Shaw. How often is employee anger an insider risk I? Detecting and measuring negative sentiment versus insider risk in digital communications. *Journal of Digital Forensics, Security and Law* 8.1 (2013), pp. 39–71. DOI: 10.15394/jdfsl.2013.1140.

[296] A. P. Henkel, S. Bromuri, D. Iren, and V. Urovi. Half human, half machine: augmenting service employees with AI for interpersonal emotion regulation. *Journal of Service Management* 31.2 (2020), pp. 247–265. DOI: 10.1108/JOSM-05-2019-0160.

[297] R. Subhashini and P. R. Niveditha. Analyzing and detecting employee’s emotion for amelioration of organizations. *Procedia Computer Science* 48 (2015), pp. 530–536. DOI: 10.1016/j.procs.2015.04.131.

[298] R. Gelbard, R. Ramon-Gonen, A. Carmeli, R. M. Bittmann, and R. Talyansky. Sentiment analysis in organizational work: towards an ontology of people analytics. *Expert Systems* 35.5 (2018). DOI: 10.1111/exsy.12289.

[299] H. Poitevin and F. De Silva. Getting Value From Employee Productivity Monitoring Technologies for Remote and Office-Based Workers. Gartner. 2020. URL: https://www.gartner.com/document/3984342.

[300] G. Sadowski, A. Litan, T. Bussa, and T. Phillips. Market Guide for User and Entity Behavior Analytics. Gartner. 2018. URL: https://www.gartner.com/document/3917096.

[301] Y. Natis and J. Daigler. Gartner's Top Strategic Predictions for 2020 and Beyond: Technology Changes the Human Condition. Gartner, Inc. 2019. URL: https://www.gartner.com/document/3970846.

[302] R. Raisamo, I. Rakkolainen, P. Majaranta, K. Salminen, J. Rantala, and A. Farooq. Human augmentation: past, present and future. *International Journal of Human-Computer Studies* 131 (2019), pp. 131–143. DOI: 10.1016/j.ijhcs.2019.05.008.

[303] B. Kropp. The Future of Employee Monitoring. Gartner, Inc. 2019. URL: https://www.gartner.com/smarterwithgartner/the-future-of-employee-monitoring/.

[304] K. Ball. Workplace surveillance: An overview. *Labor History* 51.1 (2010), pp. 87–106. DOI: 10.1080/00236561003654776.

[305] B.-A. B. On. Marginality and Epistemic Privilege. Feminist Epistemologies. Ed. by L. M. Alcoff and E. Potter. New York and London: Routledge, 1993, pp. 83–100.

[306] C. D'Ignazio and L. F. Klein. Data Feminism. Cambridge, MA and London, England: MIT Press, 2020. DOI: 10.7551/mitpress/11805.001.0001.

[307] W. A. Creech. Psychological Testing and Constitutional Rights. *Duke Law Journal* 15.2 (1966), pp. 332–371.

[308] L. H. Mirel. The limits of governmental inquiry into the private lives of government employees. *Boston University Law Review* 46 (1966), pp. 1–36.

[309] S. S. on Invasion of Privacy of the Committee on Government Operations. The Computer and Invasion of Privacy. U.S. House of Representatives, Eighty-Ninth Congress, second session. 1966. URL: https://archive.org/details/U.S.House1966TheComputerAndInvasionOfPrivacy.

[310] C. E. Gallagher. Why House hearings on invasion of privacy. *American Psychologist* 20.11 (1965), pp. 881–882.

[311] F. L. Bailey, R. E. Zuckerman, and K. R. Pierce. The Employee Polygraph Protection Act: A Manual for Polygraph Examiners and Employers. Severna Park, MD: American Polygraph Association, 1989, p. 65.

[312] J. Butler, M. Czerwinski, S. Iqbal, S. Jaffe, K. Nowak, E. Peloquin, and L. Yang. Personal Productivity and Well-Being. Chapter 2. The New Future of Work: Research from Microsoft on the Impact of the Pandemic on Work Practices. Ed. by J. Teevan, B. Hecht, and S. Jaffe. Microsoft, 2021.

[313] A. C. Williams, H. Kaur, G. Mark, A. L. Thompson, S. T. Iqbal, and J. Teevan. Supporting Workplace Detachment and Reattachment with Conversational Intelligence. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Montreal, Canada: ACM, 2018, Paper No. 88, 1–13. DOI: 10.1145/3173574.3173662.

[314] B. Gruenberg. The Happy Worker: An Analysis of Educational and Occupational Differences in Determinants of Job Satisfaction. *American Journal of Sociology* 86.2 (1980), pp. 247–271.

[315] D. McDuff, E. Jun, K. Rowan, and M. Czerwinski. Longitudinal observational evidence of the impact of emotion regulation strategies on affective expression. *IEEE Transactions on Affective Computing* 12.3 (2021), pp. 636–647. DOI: 10.1109/TAFFC.2019.2961912.

[316] H. Kaur, D. McDuff, A. C. Williams, J. Teevan, and S. T. Iqbal. "I Didn't Know I Looked Angry": Characterizing Observed Emotion and Reported Affect at Work. CHI Conference on Human Factors in Computing Systems. New Orleans LA USA: ACM, 2022, pp. 1–18. DOI: 10.1145/3491102.3517453.

[317] P. Ekkekakis. Affect, Mood, and Emotion. Measurement in Sport and Exercise Psychology. Ed. by G. Tenenbaum, R. C. Eklund, and A. Kamata. Champaign, IL: Human Kinetics, 2012, pp. 321–332.

[318] M. Gregg. Work's Intimacy. Cambridge, UK: Polity Press, 2011, p. 205.

[319] J. R. Aiello and K. J. Kolb. Electronic performance monitoring and social context: Impact on productivity and stress. *Journal of Applied Psychology* 80.3 (1995), pp. 339–353. DOI: 10.1037/0021-9010.80.3.339.

[320] G. Loewenstein and J. S. Lerner. The Role of Affect in Decision Making. Handbook of Affective Sciences. Ed. by R. J. Davidson, K. R. Scherer, and H. H. Goldsmith. New York, NY, USA: Oxford University Press, 2003, pp. 619–642.

[321] P. T. Young. Motivation and Emotion: A Survey of the Determinants of Human and Animal Activity. New York, NY, USA: John Wiley & Sons, 1961, p. 648.

[322] A. R. Damasio. Descartes' Error: Emotion, Reason and the Human Brain. London: Vintage, 2006.

[323] P. Aggarwal, S. B. Castleberry, R. Ridnour, and C. D. Shepherd. Salesperson Empathy and Listening: Impact on Relationship Outcomes. *Journal of Marketing Theory and Practice* 13.3 (2005), pp. 16–31. DOI: 10.1080/10696679.2005.11658547.

[324] I. Robertson and C. L. Cooper. Well-Being: Productivity and Happiness at Work. London, UK: Palgrave Macmillan, 2011, p. 224.

[325] C. S. Bellet, J.-E. De Neve, and G. Ward. Does Employee Happiness Have an Impact on Productivity? *Management Science* 70.3 (2024), pp. 1656–1679. DOI: 10.1287/mnsc.2023.4766.

[326] C. H. DiMaria, C. Peroni, and F. Sarracino. Happiness Matters: Productivity Gains from Subjective Well-Being. *Journal of Happiness Studies* 21.1 (2020), pp. 139–160. DOI: 10.1007/s10902-019-00074-1.

[327] J. Barling, K. E. Dupré, and E. K. Kelloway. Predicting workplace aggression and violence. *Annual Review of Psychology* 60 (2009), pp. 671–692. DOI: 10.1146/annurev.psych.60.110707.163629.

[328] C. Holton. Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decision Support Systems* 46.4 (2009), pp. 853–864. DOI: 10.1016/j.dss.2008.11.013.

[329] P. Murali, J. Hernandez, D. McDuff, K. Rowan, J. Suh, and M. Czerwinski. AffectiveSpotlight: Facilitating the Communication of Affective Responses from Audience Members during Online Presentations. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Yokohama, Japan: ACM, 2021, pp. 1–13. DOI: 10.1145/3411764.3445235.

[330] S. Samrose, D. McDuff, R. Sim, J. Suh, K. Rowan, J. Hernandez, S. Rintel, K. Moynihan, and M. Czerwinski. MeetingCoach: An Intelligent Dashboard for Supporting Effective & Inclusive Meetings. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Yokohama, Japan: ACM, 2021, pp. 1–13. DOI: 10.1145/3411764.3445615.

[331] P. Moradi and K. Levy. The Future of Work in the Age of AI: Displacement or Risk-Shifting? The Oxford Handbook of Ethics of AI. Ed. by M. D. Dubber, F. Pasquale, and S. Das. Oxford, UK: Oxford University Press, 2020, pp. 271–287. DOI: 10.1093/oxfordhb/9780190067397.013.17.

[332] Y. E. Fukumura, J. M. Gray, G. M. Lucas, B. Becerik-Gerber, and S. 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 (2021), p. 1690. DOI: 10.3390/ijerph18041690.

[333] S. Zhang, Y. Feng, L. Bauer, L. F. Cranor, A. Das, and N. 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 (2021), pp. 282–304. DOI: 10.2478/popets-2021-0028.

[334] P. Mantello, M.-T. Ho, M.-H. Nguyen, and Q.-H. Vuong. Bosses without a Heart: Socio-Demographic and Cross-Cultural Determinants of Attitude toward Emotional AI in the Workplace. *AI & Society* 38.1 (2023), pp. 97–119. DOI: 10.1007/s00146-021-01290-1.

[335] C. Burr and N. Cristianini. Can machines read our minds? *Minds and Machines* 29.3 (2019), pp. 461–494. DOI: 10.1007/s11023-019-09497-4.

[336] A. Chen and K. Hao. Emotion AI Researchers Say Overblown Claims Give Their Work a Bad Name. MIT Technology Review. 2020. URL: https://www.technologyreview.com/2020/02/14/844765/ai-emotion-recognition-affective-computing-hirevue-regulation-ethics/.

[337] B. Rogers. The Law & Political Economy of Workplace Technological Change. *Harvard Civil Rights–Civil Liberties Law Review* 55.2 (2020), pp. 531–583. DOI: 10.2139/ssrn.3327608.

[338] M. W. Finkin. Employee Privacy, American Values, and the Law. *Chicago–Kent Law Review* 72.1 (1996), pp. 221–270.

[339] I. Altman, A. Vinsel, and B. B. Brown. Dialectic Conceptions in Social Psychology: An Application to Social Penetration and Privacy Regulation. Advances in Experimental Social Psychology. Ed. by L. Berkowitz. Vol. 14. New York, NY, USA: Academic Press, 1981, pp. 107–160. DOI: 10.1016/S0065-2601(08)60371-8.

[340] S. Petronio. Boundaries of privacy: Dialectics of disclosure. Suny Press, 2002.

[341] I. Turner Daniel W. Qualitative interview design: A practical guide for novice investigators. *The Qualitative Report* 15.3 (2010), pp. 754–760. DOI: 10.46743/2160-3715/2010.1178.

[342] M. D. Gall, W. R. Borg, and J. P. Gall. Educational Research: An Introduction. 6th ed. New York, NY, USA: Longman, 1996, p. 788.

[343] R. S. Weiss. Learning from Strangers: The Art and Method of Qualitative Interview Studies. New York, NY, USA: Free Press, 1995, p. 256.

[344] K. Spiel, O. L. Haimson, and D. Lottridge. How to Do Better with Gender on Surveys: A Guide for HCI Researchers. *Interactions* 26.4 (2019), pp. 62–65. DOI: 10.1145/3338283.

[345] B. Baez. Confidentiality in Qualitative Research: Reflections on Secrets, Power and Agency. *Qualitative Research* 2.1 (2002), pp. 35–58. DOI: 10.1177/146879410200200102.

[346] V. S. Mouly and J. K. Sankaran. On the Study of Settings Marked by Severe Superior-Subordinate Conflict. *Organization Studies* 18.2 (1997), pp. 175–192. DOI: 10.1177/017084069701800201.

[347] M. Q. Patton. Qualitative Evaluation Methods. Beverly Hills, CA: Sage Publications, 1980.

[348] J. Corbin and A. Strauss. Strategies for Qualitative Data Analysis. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. 3rd ed. Los Angeles, CA, USA: SAGE Publications, 2008, pp. 85–106. DOI: 10.4135/9781452230153.n4.

[349] R. Thornberg and K. Charmaz. Grounded Theory and Theoretical Coding. The SAGE Handbook of Qualitative Data Analysis. Ed. by U. Flick. London, UK: Sage Publications, 2014, pp. 153–169. DOI: 10.4135/9781446282243.n11.

[350] K. Charmaz. Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. Introducing Qualitative Methods. London: SAGE Publications Ltd, 2006.

[351] A. Strauss. A Social World Perspective. Studies in Symbolic Interaction: An Annual Compilation of Research. Ed. by N. K. Denzin. Vol. 1. Greenwich, CT, USA: JAI Press, 1978, pp. 119–128.

[352] J. A. Holton. The Coding Process and Its Challenges. The SAGE Handbook of Grounded Theory. Ed. by A. Bryant and K. Charmaz. London, UK: SAGE Publications Ltd, 2007, pp. 265–289. DOI: 10.4135/9781848607941.n13.

[353] M. Crouch and H. McKenzie. The Logic of Small Samples in Interview-Based Qualitative Research. *Social Science Information* 45.4 (2006), pp. 483–499. DOI: 10.1177/0539018406069584.

[354] H. Nissenbaum. Privacy in Context: Technology, Policy, and the Integrity of Social Life. USA: Stanford University Press, 2009.

[355] D. J. Solove. Nothing to hide: The false tradeoff between privacy and security. Yale University Press, 2011.

[356] E. Goffman. The Presentation of Self in Everyday Life. Harmondsworth: Penguin Books, 1978.

[357] D. Lyon. Surveillance Studies: An Overview. Cambridge, UK: Polity Press, 2007.

[358] G. Mancia, M. Di Rienzo, and G. Parati. Ambulatory blood pressure monitoring use in hypertension research and clinical practice. *Hypertension* 21.4 (1993), pp. 510–524.

[359] J. S. Grove, D. M. Reed, K. Yano, and L.-J. Hwang. Variability in systolic blood pressure—a risk factor for coronary heart disease? *American Journal of Epidemiology* 145.9 (1997), pp. 771–776.

[360] N. H. Frijda. The Laws of Emotion. Psychology Press, 2017.

[361] A. Howard, C. Zhang, and E. Horvitz. Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems. 2017 IEEE Workshop on Advanced Robotics and Its Social Impacts (ARSO). Austin, TX, USA: IEEE, 2017, pp. 1–7. DOI: 10.1109/ARSO.2017.8025197.

[362] K. Roemmich and N. 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 (2021). DOI: 10.1145/3476049.

[363] K. Marx. Economic & Philosophical Manuscripts of 1844. Marx/Engels Collected Works, Volume 3: Marx and Engels 1843–1844. Trans. by M. Milligan. London; New York; Moscow: Lawrence & Wishart; International Publishers; Progress Publishers, 1975, pp. 229–346.

[364] K. W. Crenshaw. On Intersectionality: Essential Writings. New York, NY, USA: The New Press, 2017, p. 320.

[365] A. B. de Castro, J. Agnew, and S. T. Fitzgerald. Emotional labor: relevant theory for occupational health practice in post-industrial America. *AAOHN Journal* 52.3 (2004), pp. 109–115. DOI: 10.1177/216507990405200307.

[366] P. Brook. The Alienated Heart: Hochschild's *'emotional labour'* Thesis and the Anticapitalist Politics of Alienation. *Capital & Class* 33.2 (2009), pp. 7–31. DOI: 10.1177/030981680909800101.

[367] S. M. Kruml and D. Geddes. Exploring the Dimensions of Emotional Labor: The Heart of Hochschild's Work. *Management Communication Quarterly* 14.1 (2000), pp. 8–49. DOI: 10.1177/0893318900141002.

[368] A. McStay. Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data & Society* 7.1 (2020), p. 2053951720904386.

[369] M. Egger, M. Ley, and S. Hanke. Emotion Recognition from Physiological Signal Analysis: A Review. *Electronic Notes in Theoretical Computer Science* 343 (2019), pp. 35–55. DOI: 10.1016/j.entcs.2019.04.009.

[370] J. Wan, S. Escalera, G. Anbarjafari, H. J. Escalante, X. Baró, I. Guyon, M. Madadi, J. Allik, J. Gorbova, C. Lin, and Y. Xie. Results and Analysis of ChaLearn LAP Multi-Modal Isolated and Continuous Gesture Recognition, and Real Versus Fake Expressed Emotions Challenges. Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW). Venice, Italy: IEEE, 2017, pp. 3189–3197. DOI: 10.1109/ICCVW.2017.377.

[371] L. Li, T. Baltrusaitis, B. Sun, and L.-P. Morency. Combining Sequential Geometry and Texture Features for Distinguishing Genuine and Deceptive Emotions. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Venice, Italy: IEEE, 2017, pp. 3147–3153. DOI: 10.1109/ICCVW.2017.372.

[372] M. M. Khan, R. D. Ward, and M. Ingleby. Classifying Pretended and Evoked Facial Expressions of Positive and Negative Affective States Using Infrared Measurement of Skin Temperature. *ACM Transactions on Applied Perception* 6.1 (2009), pp. 1–22. DOI: 10.1145/1462055.1462061.

[373] N. Richards. Intellectual Privacy: Rethinking Civil Liberties in the Digital Age. New York, NY: Oxford University Press, 2015.

[374] L. Pessoa. On the relationship between emotion and cognition. *Nature Reviews Neuroscience* 9.2 (2008), pp. 148–158. DOI: 10.1038/nrn2317.

[375] I. Ajunwa, K. Crawford, and J. Schultz. Limitless Worker Surveillance. *California Law Review* 105.3 (2017), pp. 735–776. DOI: 10.15779/Z38BR8MF94.

[376] S. E. Wilborn. Revisiting the public/private distinction: employee monitoring in the workplace. *Georgia Law Review* 32 (1998), pp. 825–887.

[377] P. T. Kim. Data mining and the challenges of protecting employee privacy under U.S. law. *Comparative Labor Law & Policy Journal* 40.3 (2019), pp. 405–419.

[378] P. T. Kim and M. T. Bodie. Artificial intelligence and the challenges of workplace discrimination and privacy. *Journal of Labor and Employment Law* 35.2 (2021), pp. 289–315.

[379] L. Determann and J. 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 (2020), pp. 7–21.

[380] K. Crawford and J. Schultz. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review* 55.1 (2014), pp. 93–128.

[381] J. E. Cohen. The Surveillance-Innovation Complex: The Irony of the Participatory Turn. The Participatory Condition in the Digital Age. Ed. by D. Barney, G. Coleman, C. Ross, J. Sterne, and T. Tembeck. University of Minnesota Press, 2016, pp. 207–226.

[382] A. Fiore and M. 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 (2007), p. 525.

[383] L. Stewart. Big Data Discrimination: Maintaining Protection of Individual Privacy Without Disincentivizing Businesses' Use of Biometric Data to Enhance Security. *BCL Rev.* 60 (2019), p. 349.

[384] J. S. Bard. Developing a legal framework for regulating emotion AI. *Boston University Journal of Science & Technology Law* 27.2 (2021), pp. 271–311.

[385] R. Calo. Artificial Intelligence Policy: A Primer and Roadmap. *UC Davis Law Review* 51.2 (2017), pp. 399–435. DOI: 10.2139/ssrn.3015350.

[386] K. E. Darling. "Who's Johnny?" Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence. Ed. by P. Lin, G. A. Bekey, K. Abney, and R. Jenkins. Oxford, UK: Oxford University Press, 2017, pp. 173–192. DOI: 10.2139/ssrn.2588669.

[387] P. Lin, K. Abney, and G. A. Bekey, eds. Robot Ethics: The Ethical and Social Implications of Robotics. Intelligent Robotics and Autonomous Agents. Cambridge, MA: The MIT Press, 2014.

[388] M. E. Kaminski, M. Rueben, W. D. Smart, and C. M. Grimm. Averting Robot Eyes. *Maryland Law Review* 76.4 (2017), pp. 983–1024.

[389] D. Clifford, M. Richardson, and N. Witzleb. Artificial Intelligence and Sensitive Inferences: New Challenges for Data Protection Laws. Regulatory Insights on Artificial Intelligence: Research for Policy. Ed. by M. Findlay, J. Ford, S. Seoh, and T. Thamapillai. Cheltenham, UK: Edward Elgar Publishing, 2022. DOI: 10.2139/ssrn.3754037.

[390] E. Lander and A. Nelson. WIRED (Opinion): Americans Need a Bill of Rights for an AI-Powered World. The White House Office of Science and Technology Policy. 2021. URL: https://bidenwhitehouse.archives.gov/ostp/news-updates/2021/10/22/icymi-wired-opinion-americans-need-a-bill-of-rights-for-an-ai-powered-world/.

[391] M. Kearns and A. Roth. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. New York, NY, USA: Oxford University Press, 2019.

[392] J. E. Spataro. Microsoft Viva: Empowering Every Employee for the New Digital Age. Microsoft 365 Blog. 2021. URL: https://www.microsoft.com/en-us/microsoft-365/blog/2021/02/04/microsoft-viva-empowering-every-employee-for-the-new-digital-age/.

[393] K. Wiggers. Microsoft Launches Viva, an AI-Powered Information Hub for Enterprises. VentureBeat. 2021. URL: https://venturebeat.com/2021/02/04/microsoft-launches-viva-topics-an-ai-powered-information-curator-for-enterprises/.

[394] J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia. The Limits of Differential Privacy (and Its Misuse in Data Release and Machine Learning). *Communications of the ACM* 64.7 (2021), pp. 33–35. DOI: 10.1145/3433638.

[395] D. Wright. Making Privacy Impact Assessment More Effective. *The Information Society* 29.5 (2013), pp. 307–315. DOI: 10.1080/01972243.2013.825687.

[396] A. Mantelero. AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment. *Computer Law & Security Review* 34.4 (2018), pp. 754–772. DOI: 10.1016/j.clsr.2018.05.017.

[397] N. Goasduff. Gartner Says By 2023, 65% of the World's Population Will Have Its Personal Data Covered Under Modern Privacy Regulations. Gartner. 2020. URL: https://www.gartner.com/en/newsroom/press-releases/2020-09-14-gartner-says-by-2023--65--of-the-world-s-population-w.

[398] P. Garcia, T. Sutherland, M. Cifor, A. S. Chan, L. Klein, C. D'Ignazio, and N. Salehi. No: Critical Refusal as Feminist Data Practice. Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '20 Companion). Proceedings of the ACM Conference on Computer Supported Cooperative Work. New York, NY, USA: Association for Computing Machinery, 2020, pp. 199–202. DOI: 10.1145/3406865.3419014.

[399] R. J. Bies. Privacy and Procedural Justice in Organizations. *Social Justice Research* 6.1 (1993), pp. 69–86. DOI: 10.1007/BF01048733.

[400] J. W. DeCew. In Pursuit of Privacy: Law, Ethics, and the Rise of Technology. Ithaca, NY, USA: Cornell University Press, 1997, p. 224.

[401] K. E. Martin and H. F. Nissenbaum. Measuring Privacy: An Empirical Test Using Context to Expose Confounding Variables. *Columbia Science and Technology Law Review* 18.1 (2017), pp. 176–218. DOI: 10.7916/stlr.v18i1.4015.

[402] K. E. Martin and K. 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 (2016), pp. 1871–1882. DOI: 10.1002/asi.23500.

[403] D. McIver, M. L. Lengnick-Hall, and C. A. Lengnick-Hall. A strategic approach to workforce analytics: Integrating science and agility. *Business Horizons* 61.3 (2018), pp. 397–407.

[404] N. Koutsouleris, T. U. Hauser, V. Skvortsova, and M. De Choudhury. From promise to practice: towards the realisation of AI-informed mental health care. *The Lancet Digital Health* (2022).

[405] S. Monteith, T. Glenn, J. Geddes, P. C. Whybrow, and M. Bauer. Commercial Use of Emotion Artificial Intelligence (AI): Implications for Psychiatry. *Current Psychiatry Reports* 24.3 (2022), pp. 203–211. DOI: 10.1007/s11920-022-01330-7.

[406] D. J. Solove. A Taxonomy of Privacy. *University of Pennsylvania Law Review* 154.3 (2006), p. 477. DOI: 10.2307/40041279.

[407] K. E. Martin and H. F. Nissenbaum. What is it about location? *Berkeley Technology Law Journal* 35.1 (2020), pp. 251–326. DOI: 10.15779/Z382F7JR6F.

[408] Y. Kim, B. Choi, and Y. Jung. Individual differences in online privacy concern. *Asia Pacific Journal of Information Systems* 28.4 (2018), pp. 274–289.

[409] J. Bhatia and T. D. Breaux. Empirical measurement of perceived privacy risk. *ACM Transactions on Computer-Human Interaction* 25.6 (2018), pp. 1–47. DOI: 10.1145/3267808.

[410] H. Lee, S. F. Wong, J. Oh, and Y. Chang. Information privacy concerns and demographic characteristics: Data from a Korean media panel survey. *Government Information Quarterly* 36.2 (2019), pp. 294–303. DOI: 10.1016/j.giq.2019.01.002.

[411] A. Bergström. Online privacy concerns: A broad approach to understanding the concerns of different groups for different uses. *Computers in Human Behavior* 53 (2015), pp. 419–426. DOI: 10.1016/j.chb.2015.07.025.

[412] N. K. Malhotra, S. S. Kim, and J. Agarwal. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research* 15.4 (2004), pp. 336–355. DOI: 10.1287/isre.1040.0032.

[413] N. McDonald and A. Forte. The Politics of Privacy Theories: Moving from Norms to Vulnerabilities. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–14. DOI: 10.1145/3313831.3376167.

[414] M. Hildebrandt. Location Data, Purpose Binding and Contextual Integrity: What's the Message? Protection of Information and the Right to Privacy – A New Equilibrium? Ed. by L. Floridi. Vol. 17. Law, Governance and Technology. Dordrecht, Netherlands: Springer, 2014, pp. 31–62. DOI: 10.1007/978-3-319-05720-0_3.

[415] N. Forgó, S. Hänold, and B. Schütze. The Principle of Purpose Limitation and Big Data. New Technology, Big Data and the Law. Ed. by M. Corrales, M. Fenwick, and N. Forgó. Perspectives in Law, Business and Innovation. Singapore: Springer, 2017, pp. 17–42. DOI: 10.1007/978-981-10-5038-1_2.

[416] E. Anderson. Hijacked: How Neoliberalism Turned the Work Ethic Against Workers and How Workers Can Take It Back. Cambridge, UK: Cambridge University Press, 2023.

[417] J. Varelius. The Value of Autonomy in Medical Ethics. *Medicine, Health Care and Philosophy* 9.3 (2006), pp. 377–388. DOI: 10.1007/s11019-006-9000-z.

[418] K. Roemmich, F. Schaub, and N. Andalibi. Emotion AI at Work: Implications for Workplace Surveillance, Emotional Labor, and Emotional Privacy. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Hamburg Germany: ACM, 2023, pp. 1–20. DOI: 10.1145/3544548.3580950.

[419] S. Zhang, Y. Feng, and N. Sadeh. Facial Recognition: Understanding Privacy Concerns and Attitudes Across Increasingly Diverse Deployment Scenarios. Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021). Virtual Conference: USENIX Association, 2021, pp. 243–262.

[420] J. Blair, D. Mukherjee, E. F. H. Saunders, and S. 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. CHI '23: Proceedings of the ACM Conference on Human Factors in Computing Systems. Hamburg, Germany: Association for Computing Machinery, 2023, pp. 1–12. DOI: 10.1145/3544548.3581563.

[421] N. Shen, L. Sequeira, M. P. Silver, A. Carter-Langford, J. Strauss, and D. Wiljer. Patient Privacy Perspectives on Health Information Exchange in a Mental Health Context: Qualitative Study. *JMIR Mental Health* 6.11 (2019), e13306. DOI: 10.2196/13306.

[422] L. Nurgalieva, D. O'Callaghan, and G. Doherty. Security and Privacy of mHealth Applications: A Scoping Review. *IEEE Access* 8 (2020), pp. 104247–104268. DOI: 10.1109/ACCESS.2020.2999934.

[423] D. Zhang, J. Lim, L. Zhou, and A. 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 (2021), e31633. DOI: 10.2196/31633.

[424] K. E. Martin. Diminished or just different? A factorial vignette study of privacy as a social contract. *Journal of Business Ethics* 111.4 (2012), pp. 519–539.

[425] M. Walzer. Spheres of Justice: A Defense of Pluralism and Equality. New York, NY, USA: Basic Books, 1983.

[426] S. Tivatansakul, M. Ohkura, S. Puangpontip, and T. Achalakul. Emotional healthcare system: Emotion detection by facial expressions using Japanese database. 2014 6th Computer Science and Electronic Engineering Conference (CEEC 2014). CEEC '14. Colchester, United Kingdom: Institute of Electrical and Electronics Engineers, 2014, pp. 41–46. DOI: 10.1109/CEEC.2014.6958552.

[427] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, and M. Hamdi. Emotion Recognition for Healthcare Surveillance Systems Using Neural Networks: A Survey. 2021 International Wireless Communications and Mobile Computing (IWCMC 2021). IWCMC '21. Virtual / Online, China: Institute of Electrical and Electronics Engineers, 2021, pp. 681–687. DOI: 10.1109/IWCMC51323.2021.9498861.

[428] J. Kwon, D.-H. Kim, W. Park, and L. Kim. A Wearable Device for Emotional Recognition Using Facial Expression and Physiological Response. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2016). IEEE. Orlando, FL, USA, 2016, pp. 5765–5768. DOI: 10.1109/EMBC.2016.7592037.

[429] N. Azam, T. Ahmad, and N. U. Haq. Automatic Emotion Recognition in Healthcare Data Using Supervised Machine Learning. *PeerJ Computer Science* 7 (2021), e751. DOI: 10.7717/peerj-cs.751.

[430] K. Ball, E. M. Daniel, and C. Stride. Dimensions of Employee Privacy: An Empirical Study. *Information Technology & People* 25.4 (2012), pp. 376–394. DOI: 10.1108/09593841211278785.

[431] B. Zhang and E. M. Provost. Automatic Recognition of Self-Reported and Perceived Emotions. Multimodal Behavior Analysis in the Wild: Advances and Challenges. Amsterdam, Netherlands: Elsevier, 2019, pp. 443–470. DOI: 10.1016/B978-0-12-814601-9.00027-4.

[432] J. Lau, B. Zimmerman, and F. Schaub. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018). DOI: 10.1145/3274371.

[433] S. P. Weisband and B. A. Reinig. Managing User Perceptions of Email Privacy. *Communications of the ACM* 38.12 (1995), pp. 40–47. DOI: 10.1145/219663.219678.

[434] H. Nissenbaum. A Contextual Approach to Privacy Online. *Dædalus* 140.4 (2011), pp. 32–48. DOI: 10.1162/DAED_a_00113.

[435] H. Lee, S. Kang, and U. 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 (2022), Article 61, 1–26. DOI: 10.1145/3534623.

[436] D. Ellis and I. Tucker. Emotion in the Digital Age: Technologies, Data and Psychosocial Life. London, United Kingdom: Routledge, 2020. DOI: 10.4324/9781315108322.

[437] M. Deshpande and V. Rao. Depression detection using emotion artificial intelligence. 2017 International Conference on Intelligent Sustainable Systems (ICISS). 2017, pp. 858–862. DOI: 10.1109/ISS1.2017.8389299.

[438] S. Samrose, W. Chu, C. He, Y. Gao, S. S. Shahrin, Z. Bai, and M. E. Hoque. Visual Cues for Disrespectful Conversation Analysis. 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII 2019). Cambridge, MA, USA: IEEE, 2019, pp. 580–586. DOI: 10.1109/ACII.2019.8925440.

[439] S. Wood, V. Ghezzi, C. Barbaranelli, C. Di Tecco, R. Fida, M. L. Farnese, M. Ronchetti, and S. Iavicoli. Assessing the Risk of Stress in Organizations: Getting the Measure of Organizational-Level Stressors. *Frontiers in Psychology* 10 (2019). DOI: 10.3389/fpsyg.2019.02776.

[440] P. E. Naeini, S. Bhagavatula, H. Habib, M. Degeling, L. Bauer, L. F. Cranor, and N. Sadeh. Privacy expectations and preferences in an IoT world. Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017). 2017, pp. 399–412.

[441] N. Howe, E. Giles, D. Newbury-Birch, and E. McColl. Systematic Review of Participants' Attitudes Towards Data Sharing: A Thematic Synthesis. *Journal of Health Services Research & Policy* 23.2 (2018), pp. 123–133. DOI: 10.1177/1355819617751555.

[442] M. P. Tully, K. Bozentko, S. Clement, A. Hunn, L. Hassan, R. Norris, M. Oswald, and N. 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 (2018), e112. DOI: 10.2196/jmir.7763.

[443] S. Berkhout and J. Zaheer. Digital Self-Monitoring, Bodied Realities: Re-Casting App-Based Technologies in First Episode Psychosis. *Catalyst: Feminism, Theory, Technoscience* 7.1 (2021). DOI: 10.28968/cftt.v7i1.34101.

[444] D. Garg, L. Jia, and A. Datta. Policy Auditing over Incomplete Logs: Theory, Implementation and Applications. Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS '11). Chicago, IL, USA, 2011, pp. 151–162. DOI: 10.1145/2046707.2046726.

[445] United States Congress. Health Insurance Portability and Accountability Act of 1996. Public Law 104-191, 110 Stat. 1936. 1996.

[446] T. Guberek, A. McDonald, S. Simioni, A. H. Mhaidli, K. Toyama, and F. Schaub. Keeping a Low Profile? Technology, Risk and Privacy among Undocumented Immigrants. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). CHI '18. Montreal, QC, Canada: ACM, 2018, pp. 1–15. DOI: 10.1145/3173574.3173688.

[447] O. L. Haimson, D. Delmonaco, P. Nie, and A. 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 (2021), pp. 1–35. DOI: 10.1145/3479610.

[448] L. Rainie, C. Funk, M. Anderson, and A. Tyson. AI and Human Enhancement: Americans' Openness is Tempered by a Range of Concerns. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/03/17/ai-and-human-enhancement-americans-openness-is-tempered-by-a-range-of-concerns.

[449] S. Browne. Dark Matters: On the Surveillance of Blackness. Durham, NC, USA: Duke University Press, 2015. DOI: 10.1215/9780822375302.

[450] R. Coleman. Reclaiming the Streets: Closed Circuit Television, Neoliberalism and the Mystification of Social Divisions in Liverpool, UK. *Surveillance & Society* 2.2/3 (2004), pp. 293–309. DOI: 10.24908/ss.v2i2/3.3379.

[451] T. Xu, J. White, S. Kalkan, and H. Gunes. Investigating Bias and Fairness in Facial Expression Recognition. Computer Vision – ECCV 2020 Workshops. Vol. 12540. Lecture Notes in Computer Science. Springer, 2020, pp. 506–523. DOI: 10.1007/978-3-030-65414-6_35.

[452] K. Hitczenko, H. R. Cowan, M. Goldrick, and V. A. Mittal. Racial and ethnic biases in computational approaches to psychopathology. 2022.

[453] D. K. Citron. The Fight for Privacy: Protecting Dignity, Identity, and Love in the Digital Age. New York, NY, USA: W. W. Norton & Company, 2022.

[454] J. Fogel and E. Nehmad. Internet Social Network Communities: Risk Taking, Trust, and Privacy Concerns. *Computers in Human Behavior* 25.1 (2009), pp. 153–160. DOI: 10.1016/j.chb.2008.08.006.

[455] K. Bartel Sheehan. An Investigation of Gender Differences in On-Line Privacy Concerns and Resultant Behaviors. *Journal of Interactive Marketing* 13.4 (1999), pp. 24–38. DOI: 10.1002/(SICI)1520-6653(199923)13:4<24::AID-DIR3>3.0.CO;2-O.

[456] I. A. Anwar, J. Pal, and J. Hui. Watched, but Moving: Platformization of Beauty Work and Its Gendered Mechanisms of Control. *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW3 (2021), pp. 1–20. DOI: 10.1145/3432949.

[457] P. Harpur, F. Hyseni, and P. Blanck. Workplace Health Surveillance and COVID-19: Algorithmic Health Discrimination and Cancer Survivors. *Journal of Cancer Survivorship* 16.1 (2022), pp. 200–212. DOI: 10.1007/s11764-021-01144-1.

[458] M. Van Oort. The Emotional Labor of Surveillance: Digital Control in Fast Fashion Retail. *Critical Sociology* 45.7–8 (2018), pp. 1167–1179. DOI: 10.1177/0896920518778087.

[459] M. K. Scheuerman, S. M. Branham, and F. 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 (2018). DOI: 10.1145/3274424.

[460] R. Waelen and M. 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 (2022), p. 53. DOI: 10.1007/s13347-022-00548-w.

[461] A. Lerner, H. Y. He, A. Kawakami, S. C. Zeamer, and R. Hoyle. Privacy and Activism in the Transgender Community. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–13. DOI: 10.1145/3313831.3376339.

[462] O. 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 (2019), pp. 749–758. DOI: 10.1093/jamia/ocz056.

[463] K. E. Toth and C. S. Dewa. Employee Decision-Making about Disclosure of a Mental Disorder at Work. *Journal of Occupational Rehabilitation* 24 (2014), pp. 732–746. DOI: 10.1007/s10926-014-9504-y.

[464] E. P. M. Brouwers, M. C. W. 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* 30.1 (2020), pp. 84–92. DOI: 10.1007/s10926-019-09848-z.

[465] M. Elliott and J. C. Reuter. Disclosure, Discrimination, and Identity Among Working Professionals with Bipolar Disorder or Major Depression. The Oxford Handbook of the Sociology of Disability. Ed. by R. L. Brown, M. Maroto, and D. Pettinicchio. New York, NY, USA: Oxford University Press, 2021, pp. 508–524. DOI: 10.1093/oxfordhb/9780190093167.013.16.

[466] J. Nagy. Autism and the Making of Emotion AI: Disability as Resource for Surveillance Capitalism. *New Media & Society* 25.12 (2023), pp. 3185–3202. DOI: 10.1177/14614448221109550.

[467] E. B. Kang. On the Praxes and Politics of AI Speech Emotion Recognition. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. FAccT '23. Chicago, IL, USA: Association for Computing Machinery, 2023, pp. 455–466. DOI: 10.1145/3593013.3594011.

[468] K. B. Sheehan. Toward a typology of Internet users and online privacy concerns. *The Information Society* 18.1 (2002), pp. 21–32. DOI: 10.1080/01972240252818207.

[469] G. R. Milne, G. Pettinico, F. M. Hajjat, and E. Markos. Information Sensitivity Typology: Mapping the Degree and Type of Risk Consumers Perceive in Personal Data Sharing. *Journal of Consumer Affairs* 51.1 (2017), pp. 133–161. DOI: 10.1111/joca.12111.

[470] K. Nissim and A. Wood. Is privacy privacy? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2128 (2018). DOI: 10.1098/rsta.2017.0358.

[471] A. R. Brough and K. D. Martin. Critical roles of knowledge and motivation in privacy research. *Current Opinion in Psychology* 31 (2020), pp. 11–15. DOI: 10.1016/j.copsyc.2019.06.021.

[472] E. M. Uslaner. Trust, civic engagement, and the Internet. *Political Communication* 21.2 (2004), pp. 223–242. DOI: 10.1080/10584600490443895.

[473] A. F. Westin and D. Maurici. E-Commerce & Privacy: What Net Users Want. Hackensack, NJ: Privacy & American Business, 1998.

[474] H. J. Smith, S. J. Milberg, and S. J. Burke. Information privacy: Measuring individuals' concerns about organizational practices. *MIS quarterly* (1996), pp. 167–196.

[475] A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science* 347.6221 (2015), pp. 509–514. DOI: 10.1126/science.aaa1465.

[476] I. Adjerid, A. Acquisti, and G. Loewenstein. Choice architecture, framing, and cascaded privacy choices. *Management Science* 65.5 (2019), pp. 2267–2290. DOI: 10.1287/mnsc.2018.3028.

[477] S. Mukherjee, J. A. Manjaly, and M. Nargundkar. Money makes you reveal more: Consequences of monetary cues on preferential disclosure of personal information. *Frontiers in Psychology* 4 (2013), p. 839. DOI: 10.3389/fpsyg.2013.00839.

[478] K. D. Martin and P. E. Murphy. The Role of Data Privacy in Marketing. *Journal of the Academy of Marketing Science* 45.2 (2017), pp. 135–155. DOI: 10.1007/s11747-016-0495-4.

[479] F. Kehr, T. Kowatsch, D. Wentzel, and E. Fleisch. Blissfully ignorant: The effects of general privacy concerns, general institutional trust, and affect in the privacy calculus. *Information Systems Journal* 25.6 (2015), pp. 607–635. DOI: 10.1111/isj.12062.

[480] P. Kumaraguru and L. F. Cranor. Privacy Indexes: A Survey of Westin's Studies. Tech. rep. CMU-ISRI-5-138. Pittsburgh, PA, USA: Institute for Software Research International, School of Computer Science, Carnegie Mellon University, 2005.

[481] V. Garg, L. Lorenzen-Huber, L. J. Camp, and K. Connelly. Risk Communication Design for Older Adults. ISARC: Proceedings of the 29th International Symposium on Automation and Robotics in Construction. Eindhoven, Netherlands: IAARC Publications, 2012, —. DOI: 10.22260/ISARC2012/0030.

[482] P. Slovic. The Perception of Risk. Risk, Society and Policy. London & Sterling, VA: Earthscan, 2000. DOI: 10.4324/9781315661773.

[483] Y. J. Park, S. W. Campbell, and N. Kwak. Affect, Cognition and Reward: Predictors of Privacy Protection Online. *Computers in Human Behavior* 28.3 (2012), pp. 1019–1027. DOI: 10.1016/j.chb.2012.01.004.

[484] M. L. Finucane, A. Alhakami, P. Slovic, and S. M. Johnson. The affect heuristic in judgment of risks and benefits. *Journal of Behavioral Decision Making* 13.1 (2000), pp. 1–17. DOI: 10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S.

[485] M. N. Alraja, M. M. J. Farooque, and B. 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 (2019), pp. 111341–111354. DOI: 10.1109/ACCESS.2019.2904006.

[486] K. Martin. The Penalty for Privacy Violations: How Privacy Violations Impact Trust Online. *Journal of Business Research* 82.1 (2018), pp. 103–116. DOI: 10.1016/j.jbusres.2017.08.034.

[487] A. J. Rohm and G. 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 (2004), pp. 1000–1011. DOI: https://doi.org/10.1016/S0148-2963(02)00345-4.

[488] H. Xu, H. Wang, and H.-H. Teo. Predicting the Usage of P2P Sharing Software: The Role of Trust and Perceived Risk. Proceedings of the 38th Annual Hawaii International Conference on System Sciences. Vol. 7. HICSS. Big Island, HI, USA: IEEE Computer Society, 2005, pp. 201–211. DOI: 10.1109/HICSS.2005.500.

[489] J. Tolsdorf and F. Dehling. In Our Employer We Trust: Mental Models of Office Workers' Privacy Perceptions. Financial Cryptography and Data Security. Vol. 12063. Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2020, pp. 122–136. DOI: 10.1007/978-3-030-54455-3_9.

[490] E. Markos, G. R. Milne, and J. 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 (2017), pp. 79–96. DOI: 10.1509/jppm.15.159.

[491] H. Beales and J. A. Eisenach. Putting Consumers First: A Functionality-Based Approach to Online Privacy. 2013. DOI: 10.2139/ssrn.2211540.

[492] A. Calero Valdez and M. Ziefle. The Users' Perspective on the Privacy-Utility Trade-Offs in Health Recommender Systems. *International Journal of Human-Computer Studies* 121 (2019), pp. 108–121. DOI: 10.1016/j.ijhcs.2018.04.003.

[493] S. Wachter and B. Mittelstadt. A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review* 2 (2019), pp. 494–620. DOI: 10.7916/cblr.v2019i2.3424.

[494] G. Jasso. Factorial survey methods for studying beliefs and judgments. *Sociological Methods & Research* 34.3 (2006), pp. 334–423.

[495] S. L. Nock and T. M. Guterbock. Survey Experiments. Handbook of Survey Research. Ed. by P. V. Marsden and J. D. Wright. 2nd ed. Bingley, UK: Emerald Group Publishing, 2010, pp. 837–865.

[496] J. D. Wright and P. V. Marsden. Survey Research and Social Science: History, Current Practice, and Future Prospects. Handbook of Survey Research. Ed. by P. V. Marsden and J. D. Wright. 2nd ed. Bingley, UK: Emerald Group Publishing, 2010, pp. 3–26.

[497] H. Lewis. The Politics of Everybody: Feminism, Queer Theory, and Marxism at the Intersection. London, UK / New York, NY, USA: Zed Books, 2016.

[498] M. Foucault. Discipline and Punish: The Birth of the Prison. Trans. by A. Sheridan. New York, NY, USA: Vintage, 1995.

[499] A. E. Marwick and d. boyd danah. Understanding Privacy at the Margins. *International Journal of Communication* 12 (2018), pp. 1157–1165.

[500] K. Crenshaw. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. Feminist Legal Theories. Routledge, 2013, pp. 23–51.

[501] P. H. Collins. Intersectionality as Critical Social Theory. Durham, NC, USA: Duke University Press, 2019.

[502] World Medical Association. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA* 310.20 (2013), pp. 2191–2194. DOI: 10.1001/jama.2013.281053.

[503] 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. Washington, DC, USA: U.S. Department of Health, Education, and Welfare, 1978.

[504] The American Journal of Managed Care. Vulnerable Populations: Who Are They? *The American Journal of Managed Care* 12.13 Suppl (2006), S348–S352.

[505] L. M. Diamond and J. Alley. *Rethinking Minority Stress: A Social Safety Perspective on the Health Effects of Stigma in Sexually-Diverse and Gender-Diverse Populations*. 2022. DOI: 10.1016/j.neubiorev.2022.104720.

[506] D. J. Solove. The Myth of the Privacy Paradox. *George Washington Law Review* 89.1 (2021), pp. 1–51.

[507] A. Acquisti and J. Grossklags. Privacy and Rationality in Individual Decision Making. *IEEE Security & Privacy* 3.1 (2005), pp. 26–33. DOI: 10.1109/MSP.2005.22.

[508] K. 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 (2015), pp. 210–227. DOI: 10.1509/jppm.14.139.

[509] P. G. Leon, J. Cranshaw, L. F. Cranor, J. Graves, M. Hastak, B. Ur, and G. Xu. What Do Online Behavioral Advertising Privacy Disclosures Communicate to Users? Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society. New York, NY, USA: Association for Computing Machinery, 2012, pp. 19–30. DOI: 10.1145/2381966.2381970.

[510] M. Rausch and M. Zehetleitner. A comparison between a visual analogue scale and a four point scale as measures of conscious experience of motion. *Consciousness and Cognition* 28 (2014), pp. 126–140. DOI: 10.1016/j.concog.2014.06.012.

[511] M. Freyd. The graphic rating scale. *Journal of Educational Psychology* 14.2 (1923), pp. 83–102. DOI: 10.1037/h0074329.

[512] G. Z. Heller, M. Manuguerra, and R. Chow. How to Analyze the Visual Analogue Scale: Myths, Truths and Clinical Relevance. *Scandinavian Journal of Pain* 13.1 (2016), pp. 67–75. DOI: 10.1016/j.sjpain.2016.06.012.

[513] H. Treiblmaier and P. Filzmoser. Benefits from Using Continuous Rating Scales in Online Survey Research. Proceedings of the International Conference on Information Systems (ICIS) 2011. Shanghai, China, 2011.

[514] C. J. Chimi and D. L. Russell. The Likert Scale: A Proposal for Improvement Using Quasi-Continuous Variables. Information Systems Education Conference (ISECON). Washington, DC, USA, 2009, pp. 1–10.

[515] S. Y. Y. Chyung, I. Swanson, K. Roberts, and A. Hankinson. Evidence-Based Survey Design: The Use of Continuous Rating Scales in Surveys. *Performance Improvement* 57.5 (2018), pp. 38–48. DOI: 10.1002/pfi.21763.

[516] S. Jamieson. Likert scales: How to (ab)use them. *Medical Education* 38.12 (2004), pp. 1217–1218. DOI: 10.1111/j.1365-2929.2004.02012.x.

[517] I. E. Allen and C. A. Seaman. Likert Scales and Data Analyses. *Quality Progress* (2007).

[518] P. A. Bishop and R. L. Herron. Use and Misuse of the Likert Item Responses and Other Ordinal Measures. *International Journal For Exercise Science* 8.3 (2015), pp. 297–302.

[519] L. K. John, A. Acquisti, and G. Loewenstein. Strangers on a plane: context-dependent willingness to divulge sensitive information. *Journal of Consumer Research* 37.5 (2011), pp. 858–873. DOI: 10.1086/656423.

[520] J. Lin, S. Amini, J. I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang. Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy through Crowdsourcing. Proceedings of the 2012 ACM Conference on Ubiquitous Computing. UbiComp '12. Pittsburgh, PA, USA: ACM, 2012, pp. 501–510. DOI: 10.1145/2370216.2370290.

[521] J. S. Olson, J. Grudin, and E. Horvitz. A study of preferences for sharing and privacy. CHI '05 Extended Abstracts on Human Factors in Computing Systems. CHI EA '05. Portland, OR, USA: Association for Computing Machinery, 2005, pp. 1985–1988. DOI: 10.1145/1056808.1057073.

[522] K. L. Boyd and N. Andalibi. Automated Emotion Recognition in the Workplace: How Proposed Technologies Reveal Potential Futures of Work. *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW1 (2023), pp. 1–37. DOI: 10.1145/3579528.

[523] N. Karizat, A. H. Vinson, S. Parthasarathy, and N. 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 (2024), pp. 1–43. DOI: 10.1145/3637383.

[524] A. M. Chekroud, R. J. Zotti, Z. Shehzad, R. Gueorguieva, M. K. Johnson, M. H. Trivedi, T. D. Cannon, J. H. Krystal, and P. R. Corlett. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry* 3.3 (2016), pp. 243–250. DOI: 10.1016/S2215-0366(15)00471-X.

[525] L. K. Thompson, M. M. Sugg, and J. R. Runkle. Adolescents in crisis: A geographic exploration of help-seeking behavior using data from Crisis Text Line. *Social Science & Medicine* 215 (2018), pp. 69–79. DOI: 10.1016/j.socscimed.2018.08.025.

[526] N. P. Burnett, A. M. Hernandez, E. E. King, R. L. Tanner, and K. Wilsterman. A Push for Inclusive Data Collection in STEM Organizations. *Science* 376.6588 (2022), pp. 37–39. DOI: 10.1126/science.abo1599.

[527] A. Flanagin, T. Frey, S. L. Christiansen, and H. Bauchner. The Reporting of Race and Ethnicity in Medical and Science Journals: Comments Invited. *JAMA* 325.11 (2021), pp. 1049–1052. DOI: 10.1001/jama.2021.2104.

[528] L. Giatti, L. d. V. Camelo, J. F. d. C. Rodrigues, and S. M. Barreto. Reliability of the MacArthur scale of subjective social status - Brazilian Longitudinal Study of Adult Health (ELSA-Brasil). *BMC Public Health* 12 (2012). DOI: 10.1186/1471-2458-12-1096.

[529] European Union. Regulation (EU) 2016/679 (GDPR), Article 9: Processing of special categories of personal data. *Official Journal of the European Union* L 119 (2016), pp. 1–88.

[530] 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.

[531] I. Eekhout, R. M. de Boer, J. W. R. Twisk, H. C. W. de Vet, and M. W. Heymans. Missing data: A systematic review of how they are reported and handled. *Epidemiology* 23.5 (2012), pp. 729–732. DOI: 10.1097/EDE.0b013e3182576cdb.

[532] T. Tsiampalis and D. 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 (2020). DOI: 10.1186/s12874-020-01038-3.

[533] C. E. Rodríguez, M. H. Miyawaki, and G. Argeros. Latino racial reporting in the US: To be or not to be. *Sociology Compass* 7.5 (2013), pp. 390–403. DOI: 10.1111/soc4.12032.

[534] Y. Liang, X. Zheng, and D. D. Zeng. A survey on big data-driven digital phenotyping of mental health. *Information Fusion* 52 (2019), pp. 290–307. DOI: 10.1016/j.inffus.2019.04.001.

[535] W. H. Finch, J. E. Bolin, and K. Kelley. Multilevel Modeling Using R. Boca Raton, FL, USA: CRC Press, 2019.

[536] A. Gelman and J. Hill. Data Analysis Using Regression and Multilevel/Hierarchical Models. New York, NY, USA: Cambridge University Press, 2006.

[537] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67.1 (2015), pp. 1–48. DOI: 10.18637/jss.v067.i01.

[538] J. Pinheiro and D. Bates. Mixed-effects models in S and S-PLUS. Springer science & business media, 2006.

[539] F. E. Satterthwaite. Synthesis of variance. *Psychometrika* 6.5 (1941), pp. 309–316. DOI: 10.1007/BF02288586.

[540] S. G. Luke. Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods* 49.4 (2017), pp. 1494–1502. DOI: 10.3758/s13428-016-0809-y.

[541] J. J. Hox and C. J. M. Maas. Sample Sizes for Multilevel Modeling. Proceedings of the Fifth International Conference on Logic and Methodology (Social Science Methodology in the New Millennium). Bremen, Germany: Leske + Budrich, 2002, pp. 0–19.

[542] P. E. Shrout and J. L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86.2 (1979), pp. 420–428. DOI: 10.1037/0033-2909.86.2.420.

[543] W. N. Venables and B. D. Ripley. Random and Mixed Effects. Modern Applied Statistics with S. Ed. by —. New York, NY, USA: Springer, 2002, pp. 271–300. DOI: 10.1007/978-0-387-21706-2_10.

[544] D. C. Howell. Statistical Methods for Psychology. 8th ed. Belmont, CA, USA: Cengage Learning, 2012.

[545] O. Ayalon and E. Toch. Not even past: Information aging and temporal privacy in online social networks. *Human–Computer Interaction* 32.2 (2017), pp. 73–102. DOI: 10.1080/07370024.2016.1203791.

[546] J. Tang, E. Birrell, and A. Lerner. Replication: how well do my results generalize now? the external validity of online privacy and security surveys. Proceedings of the Eighteenth USENIX Conference on Usable Privacy and Security. SOUPS'22. Boston, MA, USA: USENIX Association, 2022.

[547] P. D. Bliese, M. A. Maltarich, and J. L. Hendricks. Back to Basics with Mixed-Effects Models: Nine Take-Away Points. *Journal of Business and Psychology* 33.1 (2018), pp. 1–23. DOI: 10.1007/s10869-017-9491-z.

[548] V. Patel, A. Chesmore, C. M. Legner, and S. Pandey. Trends in workplace wearable technologies and connected-worker solutions for next-generation occupational safety, health, and productivity. *Advanced Intelligent Systems* 4.1 (2022), p. 2100099.

[549] M. Cain and M. Woodbridge. Hype Cycle for the Digital Workplace. Gartner. 2020. URL: https://www.gartner.com/document/3987663.

[550] K. Roemmich, T. Rosenberg, S. Fan, and N. Andalibi. Values in Emotion Artificial Intelligence Hiring Services: Technosolutions to Organizational Problems. *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW1 (2023). DOI: 10.1145/3579543.

[551] B. Bonevski, M. Randell, C. Paul, K. Chapman, L. Twyman, J. Bryant, I. Brozek, and C. 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 (2014). DOI: 10.1186/1471-2288-14-42.

[552] A. Birhane, E. Ruane, T. Laurent, M. S. Brown, J. Flowers, A. Ventresque, and C. L. Dancy. The Forgotten Margins of AI Ethics. Proceedings of 2022 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). ACM International Conference Proceeding Series. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 948–958. DOI: 10.1145/3531146.3533157.

[553] S. Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. 1st ed. Cambridge, MA, USA: The MIT Press, 2020. DOI: 10.7551/mitpress/12255.001.0001.

[554] S. M. Okin. Justice, Gender, and the Family. New York, NY, USA: Basic Books, 1989.

[555] A. Bowser, K. Shilton, J. Preece, and E. Warrick. Accounting for Privacy in Citizen Science: Ethical Research in a Context of Openness. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17). Portland, OR, USA: Association for Computing Machinery, 2017, pp. 2124–2136. DOI: 10.1145/2998181.2998305.

[556] 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. Brussels, Belgium: European Commission, 2025.

[557] H. Nissenbaum. Contextual Integrity: Breaking the Grip of Public-Private Distinction for Meaningful Privacy. University of Washington Department of Human Centered Design & Engineering. 2021. URL: https://www.youtube.com/watch?v=VPwmC0Sfe50.

[558] H. Nissenbaum. Contextual Integrity Up and Down the Data Food Chain. *Theoretical Inquiries in Law* 20.1 (2019), pp. 221–256. DOI: 10.1515/til-2019-0008.

[559] A. Etzioni. A Cyber Age Privacy Doctrine: More Coherent, Less Subjective, and Operational. *Brooklyn Law Review* 80.4 (2014), pp. 1263–1290.

[560] A. Kak, ed. Regulating Biometrics: Global Approaches and Urgent Questions. AI Now Institute. 2020. URL: https://ainowinstitute.org/publications/regulating-biometrics-global-approaches-and-urgent-questions.

[561] E. K. Choe, S. Abdullah, M. Rabbi, E. Thomaz, D. A. Epstein, F. Cordeiro, M. Kay, G. D. Abowd, T. Choudhury, J. Fogarty, et al. Semi-Automated Tracking: A Balanced Approach for Self-Monitoring Applications. *IEEE Pervasive Computing* 16.1 (2017), pp. 74–84. DOI: 10.1109/MPRV.2017.18.

[562] R. McNaney, C. A. M. Morgan, P. Kulkarni, J. Vega, F. Heidarivincheh, R. McConville, A. L. Whone, M. Kim, R. Kirkham, and I. J. Craddock. Exploring Perceptions of Cross-Sectoral Data Sharing with People with Parkinson's. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. Ed. by S. Drucker, J. Williamson, and K. Yatani. Conference on Human Factors in Computing Systems – Proceedings. New Orleans, LA, USA: Association for Computing Machinery, 2022, pp. 1–14. DOI: 10.1145/3491102.3501984.

[563] H. Arendt. The Human Condition. Chicago, IL, USA: University of Chicago Press, 1958, p. 332.

[564] A. F. Westin. Privacy and Freedom. New York, NY, USA: Atheneum, 1967.

[565] I. Altman. The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding. Monterey, CA, USA: Brooks/Cole Publishing Company, 1975, p. 256.

[566] B. J. Moore. Privacy: Studies in Social and Cultural History. London and New York, UK/USA: Routledge, 1984, p. 342.

[567] J. Habermas. The Public Sphere: An Encyclopedia Article (1964). *New German Critique* 3 (1974), pp. 49–55. DOI: 10.2307/487737.

[568] B. Rössler, ed. Privacies: Philosophical Evaluations. Cultural Memory in the Present. Stanford, CA, USA: Stanford University Press, 2004, p. 256.

[569] A. L. Allen. Uneasy Access: Privacy for Women in a Free Society. Totowa, NJ, USA: Rowman & Littlefield, 1988.

[570] J. L. Cohen. Democracy, Difference, and the Right of Privacy. Democracy and Difference: Contesting the Boundaries of the Political. Ed. by S. Benhabib. Princeton, NJ, USA: Princeton University Press, 1996, pp. 187–217. DOI: 10.1515/9780691234168-011.

[571] I. M. Young. House and Home: Feminist Variations on a Theme. Motherhood and Spaces of Home. Ed. by S. Hardy and C. Wiedmer. Dordrecht, The Netherlands: Springer, 2005, pp. 115–147. DOI: 10.1007/978-1-137-12103-5_8.

[572] C. A. MacKinnon. Toward a Feminist Theory of the State. Cambridge, MA, USA: Harvard University Press, 1989, p. 330. DOI: 10.1086/293359.

[573] B. Rössler. The Value of Privacy. Cambridge, UK: Polity Press, 2005, p. 272.

[574] R. E. Gavison. Feminism and the Public/Private Distinction. *Stanford Law Review* 45.1 (1992), pp. 1–45. DOI: 10.2307/1228984.

[575] T. Nagel. Concealment and Exposure. *Philosophy & Public Affairs* 27.1 (1998), pp. 3–30. DOI: 10.1111/j.1088-4963.1998.tb00057.x.

[576] J. E. Cohen. Configuring the Networked Self: Law, Code, and the Play of Everyday Practice. New Haven, CT, USA: Yale University Press, 2012, p. 352. DOI: 10.12987/9780300177930.

[577] S. D. Warren and L. D. Brandeis. The Right to Privacy. *Harvard Law Review* 4.5 (1890), pp. 193–220. DOI: 10.2307/1321160.

[578] A. L. Allen. The Virtuous Spy: Privacy as an Ethical Limit. *The Monist* 91.1 (2008), pp. 3–22. DOI: 10.5840/monist200891110.

[579] D. Rosen and A. Santesso. Inviolate Personality and the Literary Roots of the Right to Privacy. *Law & Literature* 23.1 (2011), pp. 1–25. DOI: 10.1525/lal.2011.23.1.1.

[580] J. Williams. Wordsworth: Romantic Poetry and Revolution Politics. Manchester, UK; New York, NY, USA: Manchester University Press, 1989, p. 203.

[581] G. Negley. Philosophical Views on the Value of Privacy. *Law & Contemporary Problems* 31.2 (1966), pp. 319–325. DOI: 10.2307/1190674.

[582] J. S. Mill. On Liberty and Other Essays. Ed. by J. Gray. Oxford World's Classics. Oxford, UK: Oxford University Press, 1998.

[583] E. J. Bloustein and N. J. Pallone. Individual and Group Privacy. London, UK: Routledge, 2018, p. 194.

[584] D. K. Citron. Sexual Privacy. *Yale Law Journal* 128.8 (2019), pp. 1870–1962.

[585] E. J. Bloustein. Privacy as an Aspect of Human Dignity: An Answer to Dean Prosser. *New York University Law Review* 39 (1964), pp. 962–1007.

[586] N. M. Richards and D. J. Solove. Prosser's Privacy Law: A Mixed Legacy. *California Law Review* 98.6 (2010), pp. 1887–1924. DOI: 10.2307/25799958.

[587] D. Rosen and A. Santesso. The Watchman in Pieces: Surveillance, Literature, and Liberal Personhood. New Haven, CT, USA: Yale University Press, 2013, p. 352. DOI: 10.12987/9780300156645.

[588] D. K. Citron. Mainstreaming Privacy Torts. *California Law Review* 98.6 (2010), pp. 1805–1852.

[589] R. Dworkin. Taking Rights Seriously. Cambridge, MA, USA: Harvard University Press, 1977, pp. xv, 293. DOI: 10.2307/2218969.

[590] W. L. Prosser. Privacy. *California Law Review* 48.3 (1960), pp. 383–423. DOI: 10.2307/3478805.

[591] O. W. J. Holmes. The Common Law. Boston, MA, USA: Little, Brown and Company, 1881, p. 480.

[592] J. J. Thomson. The Right to Privacy. *Philosophy & Public Affairs* 4.4 (1975), pp. 295–314.

[593] R. Gavison. Privacy and the Limits of Law. *The Yale Law Journal* 89.3 (1980), pp. 421–471. DOI: 10.2307/795891.

[594] J. Turow. The Voice Catchers: How Marketers Listen In to Exploit Your Feelings, Your Privacy, and Your Wallet. New Haven, CT, USA: Yale University Press, 2021.

[595] M. J. Sandel. What Money Can't Buy: The Moral Limits of Markets. New York, NY, USA: Free Press, 1998, p. 266.

[596] M. J. Sandel. Market Reasoning as Moral Reasoning: Why Economists Should Re-engage with Political Philosophy. *Journal of Economic Perspectives* 27.4 (2013), pp. 121–140. DOI: 10.1257/jep.27.4.121.

[597] J. B. Landes. More Than Words: The Printing Press and the French Revolution. Eighteenth-Century Studies. 1991. URL: https://www.jstor.org/stable/2739189.

[598] E. Burke. Party, Parliament, and the Dividing of the Whigs, 1780–1794. Ed. by P. J. Marshall and D. C. Bryant. Vol. 4. The Writings and Speeches of Edmund Burke. Oxford, UK: Clarendon Press, 2015.

[599] M. C. Nussbaum. Capabilities and Human Rights. *Fordham Law Review* 66.2 (1997), pp. 273–300.

[600] M. Neal. Respect for Human Dignity as ‘Substantive Basic Norm’. *International Journal of Law in Context* 10.1 (2014), pp. 26–46. DOI: 10.1017/S1744552313000359.

[601] O. Schachter. Human Dignity as a Normative Concept. *American Journal of International Law* 77.4 (1983), pp. 848–854. DOI: 10.2307/2202536.

[602] United Nations General Assembly. Universal Declaration of Human Rights. United Nations. 1948. URL: https://www.un.org/en/about-us/universal-declaration-of-human-rights.

[603] D. J. Mattson and S. G. Clark. Human Dignity in Concept and Practice. *Policy Sciences* 44.4 (2011), pp. 303–319. DOI: 10.1007/s11077-010-9124-0.

[604] L. Floridi. On human dignity as a foundation for the right to privacy. *Philosophy & Technology* 29.4 (2016), pp. 307–312.

[605] H. D. Lasswell and M. S. McDougal. Jurisprudence for a Free Society: Studies in Law, Science and Policy. Vol. II. The New Haven Studies in International Law and World Public Order. Dordrecht, Boston, London: Martinus Nijhoff Publishers, 1992.

[606] European Union. Charter of Fundamental Rights of the European Union. *Official Journal of the European Communities*. C 364 (2000), pp. 1–22.

[607] European Union. Aims and Values. European Union. 2024. URL: https://european-union.europa.eu/principles-countries-history/principles-and-values/aims-and-values_en.

[608] European Union Agency for Fundamental Rights. Article 52 – Scope and Interpretation of Rights and Principles. European Union. 2021. URL: https://fra.europa.eu/en/eu-charter/article/52-scope-and-interpretation-rights-and-principles.

[609] W. R. Wiewiórowski. Shaping a Safer Digital Future: a New Strategy for a New Decade. European Data Protection Supervisor. 2025. URL: https://www.edps.europa.eu/press-publications/publications/strategy/shaping-safer-digital-future.

[610] G. Malgieri and J. Niklas. Vulnerable Data Subjects. *Computer Law & Security Review* 37 (2020), p. 105415. DOI: 10.1016/j.clsr.2020.105415.

[611] European Commission. Monitoring of Digital Rights and Principles – Support Study. European Commission. 2024. URL: https://digital-strategy.ec.europa.eu/en/library/monitoring-digital-rights-and-principles-support-study-2024.

[612] German Presidency of the Council of the European Union. Berlin Declaration on Digital Society and Value-Based Digital Government. 2020. URL: https://ec.europa.eu/isa2/sites/isa/files/cdr_20201207_eu2020_berlin_declaration_on_digital_society_and_value-based_digital_government_.pdf.

[613] N. Rao. Three Concepts of Dignity in Constitutional Law. *Notre Dame Law Review* 86.1 (2013), pp. 183–274.

[614] O. Lynskey. Deconstructing Data Protection: The 'Added-Value' of a Right to Data Protection in the EU Legal Order. *International & Comparative Law Quarterly* 63.3 (2014), pp. 569–597. DOI: 10.1017/S0020589314000244.

[615] S. Benthall, S. Gürses, and H. Nissenbaum. Contextual Integrity through the Lens of Computer Science. *Foundations and Trends in Privacy and Security* 2.1 (2017), pp. 1–69. DOI: 10.1561/3300000016.

[616] A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum. Privacy and Contextual Integrity: Framework and Applications. Proceedings of the 2006 IEEE Symposium on Security and Privacy (S&P'06). 2006 IEEE Symposium on Security and Privacy. Berkeley, CA, USA: IEEE, 2006, pp. 184–198. DOI: 10.1109/SP.2006.32.

[617] J. H. Moor. Towards a theory of privacy in the information age. *ACM Special Interest Group on Computers and Society (SIGCAS)* 27.3 (1997), pp. 27–32. DOI: 10.1145/270858.270866.

[618] Hamberger v. Eastman. 106 N.H. 107, 206 A.2d 239 (N.H. 1964). Supreme Court of New Hampshire, 1964.

[619] R. C. Post. The Social Foundations of Privacy: Community and Self in the Common Law Tort. *California Law Review* 77.5 (1989), pp. 957–1010.

[620] D. G. Johnson. Computer Ethics. The Blackwell Guide to the Philosophy of Computing and Information. Ed. by L. Floridi. Malden, MA, USA; Oxford, UK: Blackwell Publishing Ltd, 2004, pp. 63–75. DOI: 10.1002/9780470757017.ch5.

[621] S. I. Benn. Freedom, Autonomy and the Concept of a Person. *Proceedings of the Aristotelian Society* 76.1 (1976), pp. 109–130. DOI: 10.1093/aristotelian/76.1.109.

[622] S. I. Benn. Privacy, Freedom, and Respect for Persons. Privacy and Personality. Ed. by F. D. Schoeman. Cambridge, UK: Cambridge University Press, 2009, pp. 223–244. DOI: 10.1017/CBO9780511625138.009.

[623] C. Fried. Privacy. *Yale Law Journal* 77.3 (1968), pp. 475–493. DOI: 10.2307/794941.

[624] J. Rachels. Why Is Privacy Important? Philosophical Dimensions of Privacy: An Anthology. Ed. by F. D. Schoeman. New York, NY, USA: Cambridge University Press, 1984, pp. 290–299. DOI: 10.1017/CBO9780511625138.013.

[625] J. H. Reiman. Privacy, Intimacy, and Personhood. Privacy. Ed. by E. Barendt. London, UK: Routledge, 2001. DOI: 10.4324/9781315246024-3.

[626] B. Berger. Political Engagement as Intrinsic Good: Arendt and Company. Attention Deficit Democracy: The Paradox of Civic Engagement. Princeton, NJ, USA: Princeton University Press, 2011, pp. 52–82. DOI: 10.1515/9781400840311-004.

[627] J. Segal. A Delight in Doing: Individuality and Action in the Political Thought of Hannah Arendt. *New England Journal of Political Science* 2.1 (2007), pp. 6–21.

[628] J. E. Cohen. What Privacy Is For. *Harvard Law Review* 126.7 (2013), pp. 1904–1933.

[629] J. Rawls. A theory of justice. Applied ethics. Routledge, 2017, pp. 21–29.

[630] I. Berlin. 'Two Concepts of Liberty'. Liberty: Incorporating Four Essays on Liberty. Ed. by H. Hardy. Oxford, UK: Oxford University Press, 2000, pp. 118–170. DOI: 10.1093/019924989X.003.0004.

[631] N. Richards. Why Privacy Matters. Oxford, UK: Oxford University Press, 2021, p. 296.

[632] M. Queloz. The Dworkin–Williams Debate: Liberty, Conceptual Integrity, and Tragic Conflict in Politics. *Philosophy and Phenomenological Research* 109.1 (2024), pp. 3–29. DOI: 10.1111/phpr.13002.

[633] A. Sen. Equality of What? Tanner Lectures on Human Values, Volume 1. Ed. by S. M. McMurrin. Cambridge, UK: Cambridge University Press, 1980, pp. 195–220.

[634] A. Sen. Capability and Well-Being. The Quality of Life. Ed. by M. C. Nussbaum and A. Sen. Oxford, UK: Clarendon Press, 1993, pp. 270–293.

[635] A. Sen. Human Rights and Capabilities. *Journal of Human Development and Capabilities* 6.2 (2005), pp. 151–166. DOI: 10.1080/14649880500120491.

[636] M. C. Nussbaum. Creating Capabilities: The Human Development Approach. Cambridge, MA and London, UK: Belknap Press of Harvard University Press, 2011, p. 237. DOI: 10.4159/harvard.9780674061200.

[637] A. Rafaeli and R. I. Sutton. The Expression of Emotion in Organizational Life. *Research in Organizational Behavior* 11.1 (1989), pp. 1–42.

[638] L. Kutner. Due Process of Euthanasia: The Living Will, A Proposal. *Indiana Law Journal* 44.4 (1969), pp. 539–554.

[639] M. Bedi. The Curious Case of Cell Phone Location Data: Fourth Amendment Doctrine Mash-Up. *Northwestern University Law Review* 110.2 (2016), pp. 507–524.

[640] M. Anderljung, J. Barnhart, A. Korinek, J. Leung, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs, B. Martin, W. Maxwell, G. Sastry, E. Trager, and A. Weller. Frontier AI Regulation: Managing Emerging Risks to Public Safety. 2023. arXiv: 2307.03718 [cs.CY].

[641] O. Delaney, O. Guest, and Z. Williams. Mapping Technical Safety Research at AI Companies: A Literature Review and Incentives Analysis. 2024. arXiv: 2409.07878.

[642] M. Minkkinen and M. Mäntymäki. Discerning Between the "Easy"' and "Hard"' Problems of AI Governance. *IEEE Transactions on Technology and Society* 4.2 (2023), pp. 188–194. DOI: 10.1109/TTS.2023.3267382.

[643] M. Horák, V. Stupka, and M. Husák. GDPR compliance in cybersecurity software: A case study of DPIA in information sharing platform. Proceedings of the 14th international conference on availability, reliability and security. 2019, pp. 1–8.

[644] N. Swaminathan and D. Danks. Application of the NIST AI Risk Management Framework to Surveillance Technology. 2024. arXiv: 2403.15646.

[645] C. B. Landis and J. A. Kroll. Mitigating Inference Risks with the NIST Privacy Framework. *Proceedings on Privacy Enhancing Technologies* 2024.3 (2024), pp. 640–654. DOI: doi.org/10.56553/popets-2024-0013.

[646] A. E. Waldman. Industry Unbound: The Inside Story of Privacy, Data, and Corporate Power. Cambridge, UK: Cambridge University Press, 2021.

[647] K. A. Bamberger and D. K. Mulligan. Privacy on the Ground: Driving Corporate Behavior in the United States and Europe. Cambridge, MA, USA: MIT Press, 2015.

[648] E. Tabassi. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Tech. rep. NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology, 2023. DOI: 10.6028/NIST.AI.100-1.

[649] K. R. Boeckl and N. B. Lefkovitz. NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0. NIST Cybersecurity White Paper CSWP 01162020. Gaithersburg, MD: National Institute of Standards and Technology, 2020. DOI: 10.6028/NIST.CSWP.01162020.

[650] ISO/IEC JTC 1/SC 42. *ISO/IEC 23894:2023 Information Technology – Artificial Intelligence – Guidance on Risk Management*. Geneva, Switzerland: International Organization for Standardization and International Electrotechnical Commission, 2023.

[651] G. 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. *Journal of Medical Internet Research* 27 (2025), e67878. DOI: 10.2196/67878.

[652] J. McNeil. Crisis Text Line and the Silicon Valleyfication of Everything. Vice. 2022. URL: https://www.vice.com/en/article/wxdpym/crisis-text-line-and-the-silicon-valleyfication-of-everything.

[653] 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. Washington, DC, USA: National Academies Press, 2006.

[654] O. Kuenzi. FTC Non-Compete Ban. *The Reporter: Social Justice Law Center Magazine* 2023 (2023).

[655] B. Bordelon. Could Congress Fix AI Bias with Privacy Rules? Politico. 2022. URL: https://www.politico.com/newsletters/morning-tech/2022/03/29/could-congress-fix-ai-bias-with-privacy-rules-00021193.

[656] K. Porcard. The Real Harm of Crisis Text Line's Data Sharing. *Wired* (2022).

[657] B. Stanley, G. Martínez-Alés, I. Gratch, M. Rizk, H. Galfalvy, T.-H. Choo, and J. J. Mann. Coping strategies that reduce suicidal ideation: An ecological momentary assessment study. *Journal of Psychiatric Research* 133 (2021), pp. 32–37.

[658] P. Helm, B. Lipp, and R. Pujadas. Generating Reality and Silencing Debate: Synthetic Data as Discursive Device. *Big Data & Society* 11.2 (2024), pp. 1–14. DOI: 10.1177/20539517241249447.

[659] 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). European Commission. 2025. URL: https://ec.europa.eu/newsroom/dae/redirection/document/112367.

[660] K. Hill. Your Face Belongs to Us: A Tale of AI, a Secretive Startup, and the End of Privacy. New York, NY, USA: Random House, 2023.

[661] ACLU of Illinois. In Big Win, Settlement Ensures Clearview AI Complies With Groundbreaking Illinois Biometric Privacy Law. American Civil Liberties Union. 2022. URL: https://www.aclu.org/press-releases/big-win-settlement-ensures-clearview-ai-complies-with-groundbreaking-illinois.

[662] J. Buolamwini and T. Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Ed. by S. A. Friedler and C. Wilson. Vol. 81. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, 2018, pp. 77–91.

[663] P. J. Grother, M. L. Ngan, and K. K. Hanaoka. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Tech. rep. NISTIR 8280. Gaithersburg, MD: National Institute of Standards and Technology, 2019. DOI: 10.6028/NIST.IR.8280.

[664] M. McLaughlin and D. Castro. The Critics Were Wrong: NIST Data Shows the Best Facial Recognition Algorithms Are Neither Racist Nor Sexist. 2020. URL: https://itif.org/publications/2020/01/27/critics-were-wrong-nist-data-shows-best-facial-recognition-algorithms/.

[665] M. B. 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*. University of Michigan Civil Rights Litigation Initiative, 2023.

[666] W. K. Jung and H. Y. Kwon. Privacy and Data Protection Regulations for AI Using Publicly Available Data: Clearview AI Case. Proceedings of the 17th International Conference on Theory and Practice of Electronic Governance. ICEGOV '24. Pretoria, South Africa: Association for Computing Machinery, 2024, pp. 48–55. DOI: 10.1145/3680127.3680200.

[667] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union. 2024. URL: https://data.europa.eu/eli/reg/2024/1689/oj.

[668] P. Haeck. The EU's AI Bans Come with Big Loopholes for Police. Politico. 2025. URL: https://www.politico.eu/article/ai-deepseek-chatgpt-openai-eu-bans-series-of-ai-practices-but-with-loopholes/.

[669] S. M. Bellovin, R. M. Hutchins, T. Jebara, and S. Zimmeck. When enough is enough: Location tracking, mosaic theory, and machine learning. *NYU Journal of Law & Liberty* 8.2 (2014), pp. 555–628.

[670] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. M. T. Do, O. Dousse, J. Eberle, and M. Miettinen. From big smartphone data to worldwide research: The mobile data challenge. *Pervasive and Mobile Computing* 9.6 (2013), pp. 752–771.

[671] Y. Zhu, Y. Sun, and Y. Wang. Nokia Mobile Data Challenge: Predicting Semantic Place and Next Place via Mobile Data. Mobile Data Challenge 2012 (by Nokia) Workshop, in conjunction with the International Conference on Pervasive Computing (June 18–19, 2012). Newcastle, UK: Nokia Research Center, 2012.

[672] H. Bae, H. Shin, H.-G. Ji, J. S. Kwon, H. Kim, and J.-W. Hur. App-based interventions for moderate to severe depression: a systematic review and meta-analysis. *JAMA network open* 6.11 (2023), e2344120–e2344120.

[673] A. Häuselmann, A. M. Sears, L. Zard, and E. Fosch-Villaronga. EU Law and Emotion Data. 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII). Boston, MA, USA: IEEE, 2023, pp. 1–8. DOI: 10.48550/arXiv.2309.10776.

[674] S. E. Henderson. Expectations of privacy in social media. *Mississippi College Law Review* 31.2 (2013), pp. 227–247.

[675] G. Nadon, M. Feilberg, M. Johansen, and I. Shklovski. In the User We Trust: Unrealistic Expectations of Facebook's Privacy Mechanisms. Proceedings of the 9th International Conference on Social Media and Society. SMSociety '18. Copenhagen, Denmark: ACM, 2018, pp. 138–149. DOI: 10.1145/3217804.3217906.

[676] N. Martin, J. Rice, and R. Martin. Expectations of privacy and trust: Examining the views of IT professionals. *Behaviour & Information Technology* 35.6 (2016), pp. 500–510. DOI: 10.1080/0144929X.2015.1066444.

[677] The Facebook Papers. Gizmodo and NYU Cybersecurity for Democracy. 2021. URL: https://facebookpapers.com/.

[678] S. Trepte. The social media privacy model: Privacy and communication in the light of social media affordances. *Communication Theory* 31.4 (2021), pp. 549–570.

[679] J. Kim, S. Cho, R. Wolfe, J. H. Nair, and A. Hiniker. Privacy as Social Norm: Systematically Reducing Dysfunctional Privacy Concerns on Social Media. *Proc. ACM Hum.-Comput. Interact.* 9.2 (2025). DOI: 10.1145/3711049.

[680] K.-L. Lubin and L.-T. Fan. Rethinking the Rhetoric of Surveillance in Public Safety: A Critical Discourse Analysis. Future of Information and Communication Conference. Springer. 2025, pp. 657–677.

[681] D. Lyon. Identification, Surveillance and Democracy. Surveillance and Democracy. Ed. by K. D. Haggerty and M. Samatas. London: Routledge-Cavendish, 2010.

[682] D. Lyon. Identifying Citizens: ID Cards as Surveillance. Cambridge, UK: Polity, 2009.

[683] J. H. Tanne. Florida bans abortions after six weeks, leaving millions of women in southeastern US without care. *BMJ* 385 (2024). DOI: 10.1136/bmj.q1013.

[684] J. Clayton and B. Derico. Clearview AI used nearly 1m times by US police, it tells the BBC. BBC News. 2023. URL: https://www.bbc.com/news/technology-65057011.

[685] D. A. Grimes, J. Benson, S. Singh, M. Romero, B. Ganatra, F. E. Okonofua, and I. H. Shah. Unsafe abortion: the preventable pandemic. *The Lancet* 368.9550 (2006), pp. 1908–1919. DOI: 10.1016/S0140-6736(06)69481-6.

[686] A. Belanger. Cop Busted for Unauthorized Use of Clearview AI Facial Recognition Resigns. Ars Technica. 2024. URL: https://arstechnica.com/tech-policy/2024/06/cop-busted-for-unauthorized-use-of-clearview-ai-facial-recognition-resigns/.

[687] A. M. Mennicke and K. Ropes. Estimating the Rate of Domestic Violence Perpetrated by Law Enforcement Officers: A Review of Methods and Estimates. *Aggression and Violent Behavior* 31 (2016), pp. 157–164. DOI: 10.1016/j.avb.2016.09.003.

[688] K. Wuyts and W. Joosen. LINDDUN Privacy Threat Modeling: A Tutorial. CW Reports CW685. Leuven, Belgium: Department of Computer Science, KU Leuven, 2015.

[689] K. Wuyts, L. Sion, and W. Joosen. LINDDUN GO: A Lightweight Approach to Privacy Threat Modeling. 2020 IEEE European Symposium on Security and Privacy Workshops. Genova, Italy: IEEE, 2020, pp. 302–309. DOI: 10.1109/EuroSPW51379.2020.00047.

[690] R. Powell. The EU AI Act: National Security Implications. The Alan Turing Institute. 2024. URL: https://cetas.turing.ac.uk/publications/eu-ai-act-national-security-implications.

[691] M. Landauer, K. Mayer, F. Skopik, M. Wurzenberger, and M. Kern. Red Team Redemption: A Structured Comparison of Open-Source Tools for Adversary Emulation. 2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Sanya, China: IEEE, 2024, pp. 117–128. DOI: 10.1109/TrustCom63139.2024.00043.

[692] B. Bullwinkel, A. Minnich, S. Chawla, G. Lopez, M. Pouliot, W. Maxwell, J. de Gruyter, K. Pratt, S. Qi, N. Chikanov, R. Lutz, R. S. R. Dheekonda, B. Jagdagdorj, E. Kim, J. Song, K. Hines, D. Jones, G. Severi, R. Lundeen, S. Vaughan, V. Westerhoff, P. Bryan, R. S. Siva Kumar, Y. Zunger, C. Kawaguchi, and M. Russinovich. Lessons From Red Teaming 100 Generative AI Products. arXiv. 2025. DOI: 10.48550/arXiv.2501.07238.

[693] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. Red Teaming Language Models with Language Models. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 3419–3448. DOI: 10.18653/v1/2022.emnlp-main.225.

[694] R. Shah, Q. Feuillade-Montixi, S. Pour, A. Tagade, S. Casper, and J. Rando. Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation. 2023. DOI: 10.48550/arXiv.2311.03348.

[695] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks. HarmBench: a standardized evaluation framework for automated red teaming and robust refusal. Proceedings of the 41st International Conference on Machine Learning. ICML'24. Vienna, Austria: JMLR.org, 2024.

[696] Y. Shvartzshnaider and V. Duddu. Investigating Privacy Bias in Training Data of Language Models. arXiv. 2024. DOI: 10.48550/arXiv.2409.03735.

[697] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel. Ethical and social risks of harm from language models. arXiv. 2021. DOI: 10.48550/arXiv.2112.04359.

[698] S. Ouyang, S. Wang, Y. Liu, M. Zhong, Y. Jiao, D. Iter, R. Pryzant, C. Zhu, H. Ji, and J. Han. The Shifted and The Overlooked: A Task-oriented Investigation of User-GPT Interactions. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023, pp. 2375–2393. DOI: 10.18653/v1/2023.emnlp-main.146.

[699] L. Ahmad, S. Agarwal, M. Lampe, and P. Mishkin. OpenAI's Approach to External Red Teaming for AI Models and Systems. arXiv. 2025. DOI: 10.48550/arXiv.2503.16431.

[700] G. Abercrombie, D. Benbouzid, P. Giudici, D. Golpayegani, J. Hernandez, P. Noro, H. Pandit, E. Paraschou, C. Pownall, J. Prajapati, M. A. Sayre, U. Sengupta, A. Suriyawongkul, R. Thelot, S. Vei, and L. Waltersdorfer. A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms. arXiv. 2024. DOI: 10.48550/arXiv.2407.01294.

[701] L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, I. Gabriel, V. Rieser, and W. Isaac. Sociotechnical Safety Evaluation of Generative AI Systems. 2023. DOI: 10.48550/arXiv.2310.11986.

[702] A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, M. Lin, A. Mayhew, K. Collins, M. Molamohammadi, J. Burden, W. Zhao, S. Rismani, K. Voudouris, U. Bhatt, A. Weller, D. Krueger, and T. Maharaj. Harms from Increasingly Agentic Algorithmic Systems. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. FAccT '23. Chicago, IL, USA: Association for Computing Machinery, 2023, pp. 651–666. DOI: 10.1145/3593013.3594033.

[703] D. Raji, E. Denton, E. M. Bender, A. Hanna, and A. Paullada. AI and the Everything in the Whole Wide World Benchmark. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS). 2021.

[704] M. Feffer, A. Sinha, W. H. Deng, Z. C. Lipton, and H. Heidari. Red-Teaming for Generative AI: Silver Bullet or Security Theater? Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society. AAAI Press, 2025, pp. 421–437.

[705] R. Zhang, H. Li, H. Meng, J. Zhan, H. Gan, and Y.-C. Lee. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25. Association for Computing Machinery, 2025. DOI: 10.1145/3706598.3713429.

[706] Italian Supervisory Authority (Garante per la protezione dei dati personali). AI: the Italian Supervisory Authority fines company behind chatbot "Replika". European Data Protection Board. 2025. URL: https://www.edpb.europa.eu/news/national-news/2025/ai-italian-supervisory-authority-fines-company-behind-chatbot-replika_en.

[707] M. Mechergui and S. Sreedharan. Goal alignment: re-analyzing value alignment problems using human-aware AI. Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. DOI: 10.1609/aaai.v38i9.28875.

[708] P. B. Brandtzaeg and A. Følstad. Why People Use Chatbots. Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22–24, 2017, Proceedings. Vol. 10673. Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2017, pp. 377–392. DOI: 10.1007/978-3-319-70284-1_30.

[709] S.-W. Cheng, C.-W. Chang, W.-J. Chang, H.-W. Wang, C.-S. Liang, T. Kishimoto, J. P.-C. Chang, J. S. Kuo, and K.-P. Su. The now and future of ChatGPT and GPT in psychiatry. *Psychiatry and Clinical Neurosciences* 77.11 (2023), pp. 592–596. DOI: 10.1111/pcn.13588.

[710] B. Schuller, A. Mallol-Ragolta, A. P. Almansa, I. Tsangko, M. M. Amin, A. Semertzidou, L. Christ, and S. Amiriparian. Affective Computing Has Changed: The Foundation Model Disruption. 2024. DOI: 10.48550/arXiv.2409.08907.

[711] H. Wang, Y. Lin, W. Xiong, R. Yang, S. Diao, S. Qiu, H. Zhao, and T. Zhang. Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 8642–8655. DOI: 10.18653/v1/2024.acl-long.468.

[712] Y. Zhang, G. Zhang, Y. Wu, K. Xu, and Q. Gu. Beyond Bradley-Terry Models: A General Preference Model for Language Model Alignment. Proceedings of the Forty-second International Conference on Machine Learning. ICML '25. Singapore, 2025.

[713] Z. Ye, X. Li, Q. Li, Q. Ai, Y. Zhou, W. Shen, D. Yan, and Y. Liu. Beyond Scalar Reward Model: Learning Generative Judge from Preference Data. 2024. arXiv: 2410.03742.

[714] M. Nussbaum. Interview with Bill Moyers on *A World of Ideas*. Public Broadcasting Service (PBS). 1988.

[715] D. Kleine. Technologies of Choice?: ICTs, Development, and the Capabilities Approach. The Information Society Series. Cambridge, MA: MIT Press, 2013. DOI: 10.7551/mitpress/9061.001.0001.

[716] I. Robeyns. Wellbeing, Place and Technology. *Wellbeing, Space and Society* 1 (2020), p. 100013. DOI: 10.1016/j.wss.2020.100013.

[717] J. Britz, A. Hoffmann, S. Ponelis, M. Zimmer, and P. Lor. On considering the application of Amartya Sen's capability approach to an information-based rights framework. *Information Development* 29.2 (2013), pp. 106–113.

[718] S. Boylston. Designing with Society: A Capabilities Approach to Design, Systems Thinking and Social Innovation. New York, NY, USA: Routledge, 2019.

[719] M. Thomson. A capabilities approach to best interests assessments. *Legal Studies* 41.2 (2021), pp. 276–293. DOI: 10.1017/lst.2020.47.

[720] P. Das Chowdhury and K. Renaud. 'Ought' should not assume 'Can'? Basic Capabilities in Cybersecurity to Ground Sen's Capability Approach. Proceedings of the 2023 New Security Paradigms Workshop. NSPW '23. Segovia, Spain: Association for Computing Machinery, 2023, pp. 76–91. DOI: 10.1145/3633500.3633506.

[721] M. Sloane, E. Moss, and R. Chowdhury. A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns* 3.2 (2022), p. 100425. DOI: 10.1016/j.patter.2021.100425.

[722] C. S.-Y. Park. Threshold of Dignity. *Clinical Nurse Specialist* 39.3 (2025), p. 162.

[723] C. Shah, R. White, R. Andersen, G. Buscher, S. Counts, S. Das, A. Montazer, S. Manivannan, J. Neville, N. Rangan, T. Safavi, S. Suri, M. Wan, L. Wang, and L. Yang. Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies. *ACM Transactions on the Web* 19.3 (2025). DOI: 10.1145/3732294.

[724] J. Zheng, G. Tao, S. Qin, D. Wang, and Z. Ma. Intent-Based Multi-Cloud Storage Management Powered by a Fine-Tuned Large Language Model. *IEEE Access* 13 (2025), pp. 72736–72753. DOI: 10.1109/ACCESS.2025.3563200.

[725] M. Garg and C. Saxena. Emotion detection from text data using machine learning for human behavior analysis. Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications. Elsevier, 2024, pp. 129–144.

[726] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, D. Vora, and I. 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 (2023), pp. 15129–15215.

[727] H. Sakurai and Y. Miyao. Evaluating Intention Detection Capability of Large Language Models in Persuasive Dialogues. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024, pp. 1635–1657.

[728] H. Thomas. Opinion 4/2015 – Towards a New Digital Ethics: Data, Dignity and Technology. European Data Protection Supervisor. 2015. URL: https://edps.europa.eu/sites/edp/files/publication/15-09-11_data_ethics_en.pdf.

[729] O. Ulgen. AI and the Crisis of the Self: Protecting Human Dignity as Status and Respectful Treatment. The Frontlines of Artificial Intelligence Ethics: Human-Centric Perspectives on Technology's Advance. Ed. by A. J. Hampton and J. A. DeFalco. Abingdon, UK: Routledge, 2022, pp. 9–33. DOI: 10.4324/9781003030928-3.

[730] P. Altmeyer, A. M. Demetriou, A. Bartlett, and C. C. S. Liem. Position: Stop Making Unscientific AGI Performance Claims. International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235). Ed. by R. Salakhutdinov, Z. Kolter, and K. Heller. Vienna, Austria, 2024, pp. 1222–1242. DOI: 10.48550/arXiv.2402.03962.

[731] S. Balasubramaniam, V. Chirchi, S. Kadry, M. Agoramoorthy, S. P. Gururama, K. Satheesh Kumar, T. A. Sivakumar, and E. Vocaturo. The Road Ahead: Emerging Trends, Unresolved Issues, and Concluding Remarks in Generative AI—A Comprehensive Review. *International Journal of Intelligent Systems* 2024 (2024). DOI: 10.1155/2024/4013195.

[732] M. Schmitt and I. Flechais. Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing. *Artificial Intelligence Review* 57.12 (2024). DOI: 10.1007/s10462-024-10973-2.

[733] L. Berti, F. Giorgi, and G. Kasneci. Emergent Abilities in Large Language Models: A Survey. 2025. arXiv: 2503.05788.

[734] B. Baker, J. Huizinga, L. Gao, Z. Dou, M. Y. Guan, A. Madry, W. Zaremba, J. Pachocki, and D. Farhi. Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. 2025. arXiv: 2503.11926.

[735] A. Meinke, B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn. Frontier Models Are Capable of In-Context Scheming. 2024. arXiv: 2412.04984.

[736] R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, et al. Alignment Faking in Large Language Models. 2024. DOI: 10.48550/arXiv.2412.14093.

[737] T. Korbak, M. Balesni, B. Shlegeris, and G. Irving. How to Evaluate Control Measures for LLM Agents? A Trajectory from Today to Superintelligence. 2025. arXiv: 2504.05259.

[738] D. Hendrycks, E. Schmidt, and A. Wang. Superintelligence Strategy: Expert Version. 2025. arXiv: 2503.05628.

[739] A. Grabowska and A. Gunia. On Quantum Computing for Artificial Superintelligence. *European Journal for Philosophy of Science* 14.2 (2024), pp. 1–30. DOI: 10.1007/s13194-024-00584-7.

[740] N. Bostrom. Superintelligence: Paths, Dangers, Strategies. Oxford, UK: Oxford University Press, 2016.

[741] H. Kim, X. Yi, J. Yao, J. Lian, M. Huang, S. Duan, J. Bak, and X. Xie. The Road to Artificial SuperIntelligence: A Comprehensive Survey of Superalignment. 2024. arXiv: 2412.16468.

[742] I. Gabriel and V. Ghazavi. The Challenge of Value Alignment. The Oxford Handbook of Digital Ethics. Ed. by C. Véliz. Oxford, UK: Oxford University Press, 2024.

[743] A. Sen. Development as Freedom. 1st. New York, NY, USA: Oxford University Press, 1999.

[744] 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). Official Journal of the European Union. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[745] A. Bradford. The Brussels Effect: How the European Union Rules the World. Oxford University Press, 2020.

[746] European Commission. Artificial Intelligence – Questions and Answers. European Commission, Press Corner. 2024. URL: https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683.

[747] G. Umbach. Futures in EU Governance: Anticipatory Governance, Strategic Foresight and EU Better Regulation. *European Law Journal* 30.3 (2024), pp. 409–421. DOI: 10.1111/eulj.12519.

[748] S. V. Fernandes and M. S. Ullah. Development of Spectral Speech Features for Deception Detection Using Neural Networks. 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE. 2021, pp. 0198–0203. DOI: 10.1109/IEMCON53756.2021.9623077.

[749] M. M. Mafazy, C. Fatichah, and A. Yuniarti. Audio Feature Analysis and Selection for Deception Detection in Court Proceedings. *JUTI: Jurnal Ilmiah Teknologi Informasi* (2025), pp. 13–28.

[750] W. Li, W. Huan, B. Hou, Y. Tian, Z. Zhang, and A. Song. Can Emotion be Transferred? A Review on Transfer Learning for EEG-Based Emotion Recognition. *IEEE Transactions on Cognitive and Developmental Systems* 14.3 (2022), pp. 833–846. DOI: 10.1109/TCDS.2021.3098842.

[751] WSJ Staff. Inside TikTok's Algorithm: A WSJ Video Investigation. The Wall Street Journal. 2021. URL: https://www.wsj.com/articles/tiktok-algorithm-video-investigation-11626877477.

[752] D. Affsprung. The ELIZA Defect: Constructing the Right Users for Generative AI. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, 2023, pp. 945–946. DOI: 10.1145/3600211.3604744.

[753] L. Walker. Belgian man dies by suicide following exchanges with chatbot. The Brussels Times. 2023. URL: https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt.

[754] R. Irvine, D. Boubert, V. Raina, A. Liusie, Z. Zhu, V. Mudupalli, A. Korshuk, Z. Liu, F. Cremer, V. Assassi, et al. Rewarding Chatbots for Real-World Engagement with Millions of Users. 2023. arXiv: 2303.06135.

[755] J. Dwivedi-Yu, Y.-C. Wang, L. Qin, C. Canton-Ferrer, and A. Y. Halevy. Affective Signals in a Social Media Recommender System. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '22. Washington DC, USA: Association for Computing Machinery, 2022, pp. 2831–2841. DOI: 10.1145/3534678.3539054.

[756] N. Helberger, M. Sax, J. Strycharz, and H.-W. Micklitz. Choice architectures in the digital economy: Towards a new understanding of digital vulnerability. *Journal of Consumer Policy* (2022), pp. 1–26.

[757] V. Bakir. Psychological Operations in Digital Political Campaigns: Assessing Cambridge Analytica's Psychographic Profiling and Targeting. *Frontiers in Communication* 5 (2020), p. 67. DOI: 10.3389/fcomm.2020.00067.

[758] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences* 114.48 (2017), pp. 12714–12719. DOI: 10.1073/pnas.1710966114.

[759] A. R. Hochschild. Stolen Pride: Loss, Shame, and the Rise of the Right. The New Press, 2024.

[760] S. Kundu, Y. Bai, S. Kadavath, A. Askell, A. Callahan, A. Chen, A. Goldie, A. Balwit, A. Mirhoseini, B. McLean, C. Olsson, C. Evraets, E. Tran-Johnson, E. Durmus, E. Perez, J. Kernion, J. Kerr, K. Ndousse, K. Nguyen, N. Elhage, N. Cheng, N. Schiefer, N. DasSarma, O. Rausch, R. Larson, S. Yang, S. Kravec, T. Telleen-Lawton, T. I. Liao, T. Henighan, T. Hume, Z. Hatfield-Dodds, S. Mindermann, N. Joseph, S. McCandlish, and J. Kaplan. Specific versus General Principles for Constitutional AI. 2023. DOI: 10.48550/arXiv.2310.13798.

[761] O. Klingefjord, R. Lowe, and J. Edelman. What Are Human Values, and How Do We Align AI to Them? 2024. DOI: 10.48550/arXiv.2404.10636.

[762] A. Findeis, T. Kaufmann, E. Hüllermeier, S. Albanie, and R. Mullins. Inverse Constitutional AI: Compressing Preferences into Principles. 2024. DOI: 10.48550/arXiv.2406.06560.