ENVISIONING IDENTITY: THE SOCIAL PRODUCTION OF COMPUTER VISION

by

Morgan Klaus Scheuerman

M.S., University of Colorado, 2021
M.S., University of Maryland Baltimore County, 2018
B.A., Goucher College, 2016

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado
in partial fulfillment of the requirement
of the degree of
Doctor of Philosophy
Department of Information Science
2023

Committee Members:
Dr. Jed R. Brubaker, University of Colorado
Dr. Casey Fiesler, University of Colorado
Dr. Robin Burke, University of Colorado
Dr. Mary L. Gray, Microsoft Research
Dr. Allison Woodruff, Google (serving in personal capacity)

ABSTRACT

Scheuerman, Morgan Klaus (Ph.D., Information Science)

Envisioning Identity: The Social Production of Computer Vision

Dissertation directed by Dr. Jed R. Brubaker

Computer vision technologies have been increasingly scrutinized in recent years for their propensity to cause harm. Computer vision systems designed to interpret visual data about humans for various tasks are perceived as particularly high risk. Broadly, the harms of computer vision focus on demographic biases (favoring one group over another) and categorical injustices (through erasure, stereotyping, or problematic labels). Prior work has focused on both uncovering these harms and mitigating them, through, for example, better dataset collection practices and guidelines for more contextual data labeling. This research has largely focused on understanding discrete computer vision artifacts, such as datasets or model outputs, and their implications for specific identity groups or for privacy. There is opportunity to further understand how human identity is embedded into computer vision not only across these artifacts, but also across the network of human workers who shape computer vision systems.

This dissertation focuses on understanding how human identity is conceptualized across two different "layers" of computer vision: (1) at the artifact layer, where the classification ontology is deployed, in the form of datasets and model inputs and outputs; and (2) at the development layer, where social decisions are made about how to implement models and annotations by traditional tech workers. Specifically, I examine how identity is represented in artifacts and how those representations are derived from human workers. I demonstrate how human workers rely on their own subjective positionalities—the worldviews they hold as a result of their own identities and experiences.

I present six studies that identify the subjectivity of computer vision. Three studies focus on artifacts, both model outputs and datasets, to discuss how identity is currently implemented and how that implementation is embedded with specific disciplinary values that often clash with more sociocultural lenses on identity. The fourth and fifth studies focus on how human workers shape these artifacts.

Through interviews with both traditional tech workers (like engineers and data scientists) and contingent data workers (who apply requirements given to them by traditional tech workers), I uncover how the positionality of human actors shapes identity in computer vision. Finally, in the sixth study, I examine how power operates between these two types of workers, traditional tech workers and data workers. Identity, as a concept, is treated as an infrastructure for which to build products. Workers attempt to uncover some underlying truth about identity and capture it in technical systems. However, in reality, workers reference the nebulous and intangible concept of identity to implement their own positional perspectives. I demonstrate that traditional tech workers have a positional power in the development of identity in computer vision; traditional worker positionalities are viewed as expert perspectives to be solidified into artifacts. Meanwhile, data worker positionalities are viewed as risks to the quality and trustworthiness of those artifacts. Thus, traditional tech workers attempt to control data worker positionalities, instilling in data workers their own positional perspectives.

By synthesizing insights from these six studies, this dissertation contributes a theory on identity in developing technical artifacts. I argue that identity concepts in the process of computer vision development move from open—filled with nuance, complexity, history, and opportunity—to closed—narrowly defined and embedded into artifacts that are deployed to reify a specific worldview of identity. I describe how workers pull from the intangible meta-concept of “Identity” to shape, through the process of development, specific Attributes to embed into technologies. I show how workers transform these Attributes through the development process into narrower and narrower definitions. These definitions of identity thus become Technical Attributes, highly specific implementations of identity which are no longer malleable to different perspectives.

ACKNOWLEDGMENTS

Visions of academia produce a pervasive myth: the notion of the lone academic, toiling away alone to produce profound scholarship. Yet the reality is, my dissertation could not have been possible without the support of countless others throughout my career as a graduate student. There are no words sufficient to express the gratitude I feel towards all of the amazing people in my life who made this possible—but I am going to try to put it into words, nonetheless.

First, to my advisor, Jed Brubaker. Through every step of my PhD, you were there to support and guide me down my chosen path. For all of the advice, the hours of dedication, and the talking me down off of ledges, thank you. Additionally, I would love to acknowledge the dedication and support of my dissertation committee. To Casey Fiesler, Robin Burke, Allison Woodruff, and Mary Gray, thank you so much for the time, effort, and knowledge you contributed to this project. I also want to thank Iva Gumnishka; without her collaboration, this project would not have been possible.

I have had the pleasure of being in a lab filled with fantastic and supportive colleagues. Thank you to Aaron Jiang, Anthony Pinter, Katy Weathington, Kandrea Wade, Katie Gach, Dylan Burke, Casey Paul, Jes Feuston, Mally Dietrich, and Michael Ann DeVito. I also want to especially thank a few individuals within my lab. First, I want to thank Aaron Jiang for his collaboration on numerous exciting projects, and for being my buddy at our Facebook internship and beyond. Second, I want to thank Kandrea Wade, not only for her contribution to the “How We’ve Taught Algorithms Identity” project but for her support and friendship outside academia. And finally, I want to thank Katy Weathington, who, after joining our lab, quickly became one of my closest friends.

Beyond my lab, I also had the support of many other wonderful colleagues. Thank you to those in the INFO department, particularly Brianna Dym, Ellen Simpson, Mikhaila Friske, Jacob Paul, Lucy van Kleunen, Janghee Cho, Shiva Darian, Jordan Wirfs-Brock, Janet Ruppert, Samantha Dalal, Shamika La Shawn, and more—for providing me everything from friendship, support, collaboration, ideation, and job tips. Outside of my department, I would also like to thank my many collaborators, including Alex Hanna, Emily Denton, Madeleine Pape, Caitlin Lustig, Deb Raji, Katta Spiel, Cynthia L. Bennett, Cole Gleason,

Jeffrey P. Bigham, Anhong Guo, Alexandra To, Razvan Amironesei, Angelina Wang, Solon Barocas, Jared Katzman, Su Lin Blodgett, Hanna Wallach, Kristen Laird, Ali Abdolrahmani, William Easley, and more. I especially want to thank Alex Hanna and Emily Denton for being integral mentors to me from the beginning of my work with Google through my PhD.

Thank you also to my family, especially my mother, Shelly, who, despite not knowing exactly what I do, has always supported me through my college career. Similarly, thank you to my friends outside of academia for their continued support and understanding. Finally, I want to thank my cats, the little monsters who kept me company as I wrote, always offered unconditional love, and constantly interrupted my work by screaming and knocking things over throughout my dissertation.

From first-generation college student to doctor, it feels unreal to finally reach the end (or the start of many new beginnings).

CONTENTS

TABLES

FIGURES

1
INTRODUCTION

Identity, the qualities that define each individual, is salient to every aspect of the human experience. Identity can be theorized in a myriad of ways and is often viewed as a culmination of both the mind and the body. Internal invisible identities are often construed as one's own perception of the individual self, who a person believes they are on an individual level (e.g., Chalmers, 1996; Descartes, 1993; Locke, 1689). Such conceptions of the self might also tie to interpersonal connections, identities tied to some collective group or social roles, which might include concepts like family roles, professional roles, or a member of a specific academic discipline (Hogg, 2016). Both individual and collective, sociocultural identities—such as gender, race, ethnicity, sexuality, religion, and class—entangle to shape how humans relate to themselves and others, on an individual interpersonal level and on a broader societal scale. In particular, *visible identities,* such as race and gender, where identity is tied to and often inscribed onto the body, influence one's sense of self and how that self relates to the world (Alcoff, 2006). The worldview that one develops as a result of this intricate web of identities is often called *subjective positionality*—how the position a subject occupies in the world shapes their experiences and perspectives, and thus decision-making processes and agency that subject may have access to in specific contexts (Anthias, 2008; da Silva & Webster, 2018; Merriam et al., 2001).

Despite the perspective that they are visible, sociocultural identities are fluid, changing both temporally and culturally (Lamont, 2001). Yet, sociocultural identities are often naturalized or calcified within specific times and contexts (Reicher, 2004). For example, perspectives on race in the United States have changed significantly over time (Alim et al., 2016). Further, sociocultural identities are often highly consequential: racism, cissexism, and misogyny operate at intrapersonal, interpersonal, institutional, and structural levels (Pincus, 2019; Risman, 2018). Sociocultural identities operate as technologies themselves, tools from which to derive and assign meaning and agency (Coleman, 2009; Sheth, 2009). As technologies of classification, sociocultural identity attributes, like race and gender, are commonly used as variables in a variety of technical systems. From the U.S. Census (Rodriguez, 2001) to Facebook (Bivens & Haimson, 2016), we see technical representations of these identities everywhere. Seemingly stable systems of identity classification have been critiqued across a number of domains, from library cataloging (Roberto, 2011) to government identification (Gehi & Arkles, 2007), for portraying politicized values as neutral, natural, or inflexible.

With the growing momentum of machine learning, a branch of artificial intelligence aimed at learning patterns from prior data, identity has become a crucial topic in computing. While identity has always been crucial to interface design (e.g., Haimson & Hoffmann, 2016), machine learning methods present unique challenges due to the use of historical data to drive decisions about identity-specific tasks at a massive scale. In particular, the domain of computer vision (CV), a subset of machine learning for visual pattern recognition, regularly utilizes subjective identity characteristics—from sociocultural identities like race and gender to internal characteristics like emotion and intelligence—in both academic research and commercial application. From facial analysis techniques for classifying the race and gender of human faces (Fu et al., 2014; Ng et al., 2015) to auto-captioning images with racialized and gendered concepts (Barlas et al., 2019; J. L. P. Díaz et al., 2020), to more subtly using demographic information in data to make decisions (Eubanks, 2018; Klare et al., 2012; Richardson et al., 2019; Vayena et al., 2018), computer vision regularly intersects with human identity.

The use of human identity characteristics in both the inputs (the training data) and the outputs (the inferences) of computer vision have been increasingly scrutinized by computing researchers (e.g., Hamidi et al., 2018; Keyes, 2018; Raji et al., 2020). Race and gender have become two of the largest concerns regarding bias in machine learning fairness literature—particularly, how systems are biased against certain races and genders (Agüera y Arcas, 2017; Klare et al., 2012; Ngan & Grother, 2015) and how to mitigate those biases (Buolamwini & Gebru, 2018; Gong et al., 2019; T. Wang et al., 2019). These concerns include bias in the databases used to train and evaluate machine learning algorithms (e.g., Danks & London, 2017; Mehrabi et al., 2019; Tommasi et al., 2017) and biases in the outcomes that might arise from numerous areas of the pipeline post-training (Suresh & Guttag, 2019). Sample selection bias—bias resulting from what subjects are included in a database—is a known issue in machine learning datasets (e.g., Mehrabi et al., 2019; Torralba & Efros, 2011), leading computer scientists to try to mitigate it using various methods. For example, scholars have proposed algorithmic methods for both "undoing" dataset bias in existing biased datasets (e.g., Khosla et al., 2012) and for creating "unbiased" models using biased data (e.g., Kamiran & Calders, 2009).

Much of the current research on ethics in computer vision has focused on technical artifacts—the datasets (e.g., Peng et al., 2021), the documentation (e.g., Miceli et al., 2020), and the models (e.g., Barredo Arrieta et al., 2020). Beyond bias measurement and mitigation strategies, interdisciplinary researchers are also conducting more critical analyses of identity in computer vision. Concern for using predictive modeling that might implicitly or explicitly result in systemic racial and gender discrimination and ableist assumptions about emotions and innate characters has led to escalating discussions on the very morality of facial analysis use cases (e.g., Bacchini & Lorusso, 2019; Marciano, 2019; Wevers, 2018). There has been less research on the actual work processes and positionalities of those involved in shaping computer vision, resulting in obscurity about the role of human subjectivity in defining identity in computer vision. One reason for this lack of research may be the challenges of accessing industry tech workers, which I will discuss in detail in my methods. Beyond the specific field of computer vision, Holstein et al. identified numerous challenges that traditional tech workers[1] face when trying to implement machine learning fairness, such as a lack of fairness auditing resources and presumed biases of humans involved in the processes (Holstein et al., 2019). Madaio et al. similarly identify the subjective challenges of AI fairness workers, such as which demographic groups to focus on and how to evaluate bias (Madaio et al., 2020). Gray and Suri highlight the unfair work practices that gig workers[2]—or, as they coin, ghost workers—face (Gray & Siddharth, 2019). In terms of work practices around computer vision, in the case of either traditional tech workers or data workers[3], Miceli et al. highlight the power that traditional tech workers enact over the data annotation process (Miceli et al., 2020; Miceli & Posada, 2021). Not only do

[1] E.g., engineers, ethicists, researchers, etc. working in "high tech," focused on computing

[2] Workers hired for short term work on online platforms, many of which contribute to machine learning datasets

[3] Often "gig workers" who work on short term contracts, or "gigs" (Vallas & Schor, 2020)

traditional tech workers define the taxonomies underlying data labeling in ways that reflect their own specific worldviews regardless of the cultural site of data annotation (Miceli & Posada, 2021), annotators regularly view traditional tech workers as having more expertise and knowledge than themselves, and therefore do not question labeling guidelines (Miceli et al., 2020).

Such prior work on data workers has highlighted a need to better understand the role of workers, including both traditional tech workers and data workers, in shaping the subjective outcomes of machine learning, including computer vision. It is critical to understand the *social* production of computer vision artifacts by examining how the positionalities of tech workers in the many roles across the computer vision development pipeline embed their values and perspectives into them. Thus, in this dissertation, I plan to more deeply explore how identity is embedded into computer vision and how the humans involved in the processes of computer vision development shape these identity-based outcomes in computer vision systems. The following research questions guided the work in this dissertation:

- How is identity embedded into computer vision and what does that communicate about the values of the field?
- How do industry professionals working on computer vision pipelines make decisions about incorporating human identity into computer vision systems?
- How do individuals' positionality impact the way they do their work and how does that influence the outcomes of computer vision models?
- What are the relationships that shape and constrain workers when developing computer vision models?

To answer these questions, I present a series of interconnected studies focused on examining how human identity characteristics are operationalized for computer vision across multiple layers of human and computer actors. All of the work presented in this dissertation is guided by an Interpretivist methodology, regardless of the methods employed in each study. Given that I will argue that identity in computer vision is constructed through the positionality of workers, I adopt an Interpretivist lens which proposes a multidimensional worldview that is shaped by human agency, experiences, and perceptions (Mackenzie & Knipe, 2006).

I argue that identity in computer vision systems is conceptualized across two infrastructural layers: (1) the artifact layer, where the classification ontology is deployed; and (2) the development layer, where social decisions are made about how to implement the classification ontology via models and annotations. The way identity is conceptualized across these two layers reflects a diverse web of positional approaches, synthesizing personal, community, and societal values. My dissertation presents research in these two areas:

Artifacts. I present work on what computer vision artifacts communicate about identity. I show that identity—like race and gender—is presented in computer vision artifacts as obvious, static, and apolitical; it is presented as truth. In this presentation, computer vision actively marginalizes and erases identities that do not fit into the narrow definitions embedded into datasets and models. The current status quo of commercial facial analysis technologies cannot contend with either binary transgender faces or non-binary faces—misgendering transgender individuals while erasing non-binary genders by forcing them into a binary classification system. I also show that examining artifact documentation communicates the implicit values of their designers. While artifacts are presented as objective and neutral, attending to artifacts—how they are constructed and how they fail for certain identities—reveals the underlying subjectivity of identity. The disciplinary values embedded into current computer vision practices insinuate a devaluing of human positionality and place data work—work to collect and label data for use to train computer vision, a core necessity of the field—as less important than model work, engineering and tuning the models.

Development. I examine the subjective practices of developing computer vision artifacts. I analyze how the positionality, in terms of experiences, cultures, identities, and worldviews, of humans—from engineers building out models to annotators contracted to label image data—subjectively shape the outcomes of identity in computer vision systems. I demonstrate how both traditional tech workers and data workers embed their own positional perspectives into identity concepts for computer vision artifacts. Traditional tech worker positionalities are shaped by the context of the companies they work in and their negotiations with their fellow workers. Meanwhile, data worker positionalities often reflect their exposure or lack thereof to certain types of identities. In both workers, positional gaps during development result in unforeseen and undesirable outcomes for computer vision artifacts. Further, traditional tech workers and data workers have different levels of positional power in defining identity, and thus traditional worker approaches are prioritized over data worker approaches.

In this dissertation, I present my work in two parts as determined by these two areas, Artifacts and Development. In Part One: Artifacts, I describe how identity has been historically implemented in both datasets and models. However, examining artifacts raises underlying questions about *how* identity has become historically implemented in specific ways. In Part Two: Development, I show how different workers attend to identity during the development of computer vision artifacts. This work attends to the open questions about how and why identity in artifacts has been designed the way it historically has been. The two parts of this dissertation, taken together, illuminate how identity permeates every aspect of computer vision.

Study Summary	Publication (As Applicable)	Chapter
Content analysis of race and gender labels in computer vision image datasets focused on understanding how race and gender are represented, what sources are used to define them, and how they are annotated	Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. Proc. ACM Hum.-Comput. Interact. 4, CSCW1, Article 58 (May 2020), 35 pages. https://doi.org/10.1145/3392866	3. How identity shows up in datasets
Audit of 5 commercial computer vision services' gender classification and image labeling models focused on understanding how diverse genders are classified and labeled	Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 144 (November 2019), 33 pages. https://doi.org/10.1145/3359246	4. How identity shows up in models
Content analysis of computer vision datasets with attention to how specific values are communicated through documentation	Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 317 (October 2021), 37 pages. https://doi.org/10.1145/3476058	5. How creator values are communicated through artifacts
Interviews with 24 traditional tech workers about how they approach implementing identity concepts in computer vision	N/A	7. How traditional workers implement identity
Interviews with and observations of 27 data workers about how they approach collecting and annotating identity concepts for computer vision	N/A	8. How data workers implement identity
Analysis of how traditional tech workers and data workers both enact their positionalities in the development of computer vision with particular attention to power	N/A	9. How work practices reflect power

Table 1. A table describing each study in this dissertation.

Together, the results from the six studies provide answers to the proposed research questions and demonstrate that identity is differentially embedded in three layers of computer vision. I provide computer vision developers and researchers with a more nuanced and complex perspective about identity—not only the implementation of identity categories in artifacts, but the role of positionality in shaping those categories and artifacts in the first place. My work assesses computer vision infrastructures, not just as sociotechnical systems, but as inherited and interconnected layers with the goal of understanding how they reference, leverage, and constrain one another. I connect the dots of positionality across the computer vision pipeline, from traditional tech worker to data worker to the final product, both dataset and model.

Based on all of these studies, I contribute a theoretical framework of identity development for technical artifacts. I argue that concepts of identity in technology development are increasingly constrained through the process of development. Identity moves from something that is open—nebulous, intangible, and multidimensional—to something that is closed—calcified, narrow, and one-dimensional. The contributions of this dissertation not only highlight how the positionality in the development layer shapes identity in the artifact layer, but showcase how identity is transformed from something open to something closed through the development process.

Structure of the Dissertation

This dissertation is organized into ten chapters. It is also organized into two parts. The first part of the dissertation covers Chapters 3-5 and focuses on identity in artifacts. The second half of the dissertation covers Chapters 6-9 and focuses on how identity is implemented through development.

Following the current Introduction chapter, Chapter 2 provides an overview of background work central to this dissertation. First, I provide an introduction to computer vision, including its applications and approaches to documentation. Then, I provide literature on theorizing identity, including how identity is applied in technical artifacts, like computer vision. Finally, I review scholarship on how values and positionalities are instilled into artifacts by their human creators.

Chapters 3, 4, and 5 all focus on how identity is represented in artifacts—fully developed computer vision datasets and models. Chapter 3 shows how identity, specifically race and gender, are conceptualized in computer vision datasets, the underlying data used to train computer vision models how to classify. It introduces the major problem underlying this dissertation: that identity can be conceptualized in numerous ways, and those ways are laden with social decisions. I describe how race and gender are often poorly documented, presented as static and apolitical, despite the sociohistorical reality of race and gender categories.

Chapter 4 focuses on how identity is classified by models. Specifically, it shows how gender is treated in commercial computer vision models and the consequences current representations of gender have for transgender and non-binary individuals. It proves how current approaches to identities like gender actively marginalize those who do not fit into the normative constructions of identity employed by model creators.

Chapter 5 examines how artifacts like datasets actively reflect the implicit values of their creators. It describes how the authors of computer vision datasets value technical skills and misplaced notions of objectivity over social knowledge or experiences. This chapter explicitly discusses how the values of an artifact's human creators reflect specific worldviews and beliefs.

Chapters 6, 7, 8, and 9 focus on the development of artifacts, centering the humans involved in embedding identity into computer vision. Chapter 6 outlines the methods underlying the three studies presented in Chapters 7, 8, and 9. All of these chapters focus on the development of industrial-scale computer vision, rather than academic research (as demonstrated in Chapters 3 and 5). I chose to focus on industrial-scale computer vision to understand the work practices of those involved in products which will be deployed in the real world.

Both Chapters 7 and 8 focus on examining the role of individual positionality in computer vision development practices. Positionality describes how an individual's experiences, values, beliefs, and identity impact how they view the world. Chapter 7 focuses on the role of traditional tech workers, like engineers, data scientists, and researchers, while Chapter 8 focuses on the role of data workers, like data collectors and annotators. Both chapters examine how different workers' positionalities influence how identity is represented in computer vision artifacts.

Chapter 9 then attends to the relationship between traditional tech worker positionality and data worker positionality. Specifically, it showcases how these two types of workers have different levels of power in applying their positionalities to computer vision data. Traditional tech workers are given more positional power, and thus able to shape identity outcomes in computer vision more acutely than data workers, whose positionalities are viewed as undesirable.

Finally, in my concluding chapter, Chapter 10, I propose a theory based on my analysis of identity in both finalized artifacts and development practices. I argue that identity in technical artifacts reflects an open-to-closed development model, where identity concepts become increasingly constrained through both the processes of development and the differential power of the actors involved in the process. Identity in technologies is always transformed into something calcified, rigid, and closed.

Methodology of my Dissertation

My dissertation includes six studies spanning the duration of my doctoral work—from 2019 to 2023. In this section, I briefly describe the methodological underpinnings of my research in its entirety.

I use mixed methods throughout my dissertation, ranging from algorithmic audits (Chapter 4) to content analyses (Chapters 3, 4, and 5) to interviews and ethnographic observations (Chapters 7, 8, and 9). However, underlying these many methods, I adopt an overwhelmingly interpretivist orientation. Identity is complex, messy, socially constructed and historically meaningful, and thus, even when engaging with quantitative data, my approach to scientific “truth” is one that is multiply constructed and situationally meaningful. Specifically, I adopt a feminist epistemology focused on centering the socially meaningful and contextually situated “truths” of those most marginalized (E. Anderson, 1995). I take what Anderson calls a “value-laden inquiry” (E. Anderson, 1995), one which actively acknowledges and engages with both researcher and researched as significant actors in undercovering relevant and contextual truths.

No researcher approaches a problem from nowhere, and thus my own positionality—just like that of my research subjects—has led me to specific questions, specific ways of interpreting, and specific conclusions. I will describe my methods as they pertain to individual studies throughout the rest of this dissertation. I describe my methods and positionality individually in Chapters 3, 4, and 5. Meanwhile, I describe my methods and positionality in Chapter 6 as it pertains to Chapters 7, 8, and 9.

As Holloway suggests, good research involves good storytelling (Holloway & Biley, 2011). The work in this dissertation is not presented sequentially (either in terms of the order in which I conducted them or in the order in which development occurs), but rather, presented in the form of a “story.” I begin by presenting work on artifacts to ground readers in understanding how identity is actually represented in computer vision. I begin with datasets because understanding how identity is represented in the data helps to understand how identity is represented in models. I present work on values in artifacts third because it begins to bridge the gaps between finished artifacts and development practices.

I present work on development after work on artifacts because I desire readers to ask, much like I did: *so how did we get here?* Once readers understand the status quo of identity in artifacts, I then present work that uncovers how those artifacts were created, and the complexity of social practices that went into them. I showcase first how different workers shaped the artifacts presented in Chapters 3 through 6. Then, I discuss how those workers aren't equally represented in those constructions.

My hope is that readers can understand the perspective, if not believe for themselves, that artifacts can be constructed differently, and that identity in computer vision is *designed*, not innate. In the following chapters, I will tease apart how identity is designed, and then consider how it can be designed radically differently.

2
BACKGROUND

The focus of this dissertation is on how the concept of "identity" becomes embedded into human-centric computer vision models and what those conceptualizations communicate about the values of industry practices. Given this focus, I present background on three broad areas of scholarship: (1) computer vision; (2) identity; and (3) values.

In the background section on computer vision, I define computer vision technologies, including those specific to the analysis of human faces. In this section, I also briefly describe current concerns about issues of transparency and diversity in computer vision and how some scholars are attempting to address those concerns. The goal of this section is to provide the reader with a base understanding of the site of inquiry (computer vision) and why computer vision is a valuable domain for interrogating how identity becomes conceptualized and embedded into technical artifacts.

A crucial aspect for how identity becomes embedded into computer vision is through the positionality of those developing it. Thus, in the following section on identity, I present three areas of scholarship for both explaining positionality and identity. The first is focused on social theories of identity, where scholars conceptualize the notion of human identity as something either invisible and internal or visible and external. Here, I describe how I approach identity in this dissertation, not as a search for some inherent source of identity but as a means of understanding how individuals enact positionality to classify the identities of others in computer vision work. Next, I describe how identity has been conceptualized in technical infrastructures, like databases for demography. The purpose of this subsection is to establish a background on current approaches in social computing to understanding identity. In the final subsection on identity, I describe how identity has been conceptualized in computer vision, specifically. I describe how computer vision technologies are currently conceptualizing human identity as visible and stable, and therefore classifiable. I also describe how FATE (fairness, accountability, transparency, and ethics) has become a focus in computer vision due to differing perspectives and concerns about the current state of identity in computer vision. I use this subsection to provide a background of the current state of identity in computer vision and FATE, and also to highlight that my dissertation seeks to understand how the current state of identity shows up in computer vision.

The final section focuses on values and how scientific practices and the resulting artifacts are laden with specific disciplinary values that reflect a specific sociocultural context. I describe how scholars have used documentation to understand scientific values in technical infrastructures, like databases, and how prior work on computer vision datasets have begun to unearth values around data collection and documentation practices. I use this section to surface the areas where values are embedded that I plan to focus on: in the artifacts themselves.

Computer Vision

Applications and Approaches

Computer vision is a domain of machine learning focused on training computer systems to analyze patterns in visual data, like images and videos. Computer vision models might be trained to complete a variety of different tasks, such as recognizing specific objects (e.g., that there is a stop sign present in a live feed for autonomous vehicles), counting the number of objects (e.g., the number of plastic bottles in a garbage pile), or classifying types of objects (e.g., the type of clothing a person is wearing). Human-centric computer vision refers to computer vision that centers humans, in terms of data and tasks. Human-centric computer vision might include data featuring whole human bodies, or portions of human bodies, such as the hands (e.g., for gesture recognition) or the face.

The face is featured in the vast majority of human-centric computer vision systems and is often the site of inquiry and critique from those concerned about bias in computer vision. Computational facial analysis technologies, such as facial detection, recognition, and classification, first emerged in the 1960s but have rapidly advanced in the last decade (Bledsoe, 1964; Raviv, 2020). Such technologies are now embedded into everyday life in many countries across the globe. Facial recognition is commonly used in airports, by local police departments, and by government agencies, like the FBI and ICE in the United

States (Ghaffary & Molla, 2019). In everyday commercial activity, facial classification is often sold to businesses for determining demographic information about potential customers for targeted advertising (e.g., Kuligowski, 2019). Almost every large technology company in the United States—Google, Microsoft, IBM, Amazon—offers its customers access to a facial analysis service.

Facial analysis technologies are built using a number of machine learning approaches (e.g., Gargesha & Panchanathan, 2002; Lien et al., 1998; Szlávik & Szirányi, 2004), generally to accomplish two goals: facial recognition and facial classification. Broadly, facial analysis systems work much like anthropometric measurements of facial features performed by human beings, but through automated, large-scale pattern recognition. Facial *recognition*, like other biometric technologies, attempts to match an individual's identity to records in a database based on images of their face. Recognition is now used by social media platforms (e.g., Facebook's tagging system (Narayanan, 2019), consumer electronics (e.g., iPhone's Apple ID), and police departments (e.g., Valentino-DeVries, 2020). Facial *classification*, on the other hand, attempts to classify individuals according to categorical schemas. For example, the perceived gender (e.g., Ramey & Salichs, 2014; Rodríguez et al., 2017; Santarcangelo et al., 2015) or ethnicity (e.g., Gutta et al., 1998; Lu & Jain, 2004; Mansoor Roomi et al., 2011) of a face; whether the face is considered beautiful or not (e.g., Eisenthal et al., 2006; Whitehill & Movellan, 2008); and even what the face can tell us about a person's criminality, sexuality, or intelligence (e.g., McFarland, 2016; Y. Wang & Kosinski, 2017; Wu & Zhang, 2016). Automated bodily analysis techniques work similarly but are focused on analyzing the entire body rather than solely the face. A simplified diagram of facial recognition and classification pipelines can be seen in **Figure 1**.

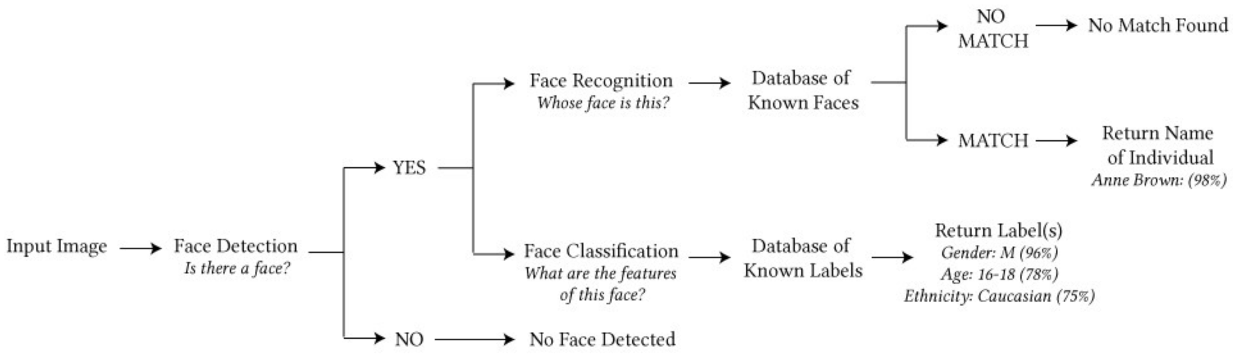


Figure 1. A diagram of facial analysis tasks: one branch represents facial recognition, the other facial classification. The diagram represents one simple approach; there are numerous other approaches developers might take, including using facial classification to aid facial recognition (e.g., Mahalingam & Kambhamettu, 2011). Figure originally published in Chapter 4 (Scheuerman et al., 2020).

All automated analysis systems are premised on their data: the hundreds to thousands of images that are used to train the model to complete a specific task. For human-centric computer vision, this means data that features human beings. In the case of body detection tasks, these might be images of human beings annotated by bounding boxes or semantic polygons, highlighting the human shape to teach a system what a human looks like. In gesture recognition, this might be images of hands with joint annotations. In the case of facial detection tasks, images of faces are used to train a model to detect what sorts of patterns in an image equate to a human face. In the case of facial recognition tasks, the system is trained to distinguish an individual's face from others in a database. The more diverse the images of a single individual available to the system to process, the more successful it should become at accurately recognizing that individual.

Classification systems, on the other hand, are trained to recognize predefined features of a face or body. Such systems use these facial features to classify an individual in terms of specific demographic categories of interest, such as those associated with race or gender. To do so, (human) annotators assign categories to each image in the database, often with little explanation as to how they made their determination. More recently, however, efforts have been made in computer vision research to allow individuals in biometric datasets to self-identify their gender (Hazirbas et al., 2021). The model then reads those images to learn visual patterns in the data, such as what patterns are present in those images that are labeled as "female" by those annotators. A system then uses this information to classify new, previously unseen images.

Posterior to data collection, the images are often annotated, or labeled, with specific information that is then useful to facial analysis. Databases may be annotated differently depending on the task the database was intended to be used for. For example, some databases may be annotated with identifiable information for each individual person in an image, such as a name. Another may be annotated with characteristic information about an individual, such as their race or gender. Annotation may be done via a number of methods. Smaller databases may be annotated by the original creators, as can be seen in databases like Pilot Parliaments Benchmark (PPB) (Buolamwini & Gebru, 2018). Increasingly, databases are being annotated through crowdsourcing or by scraping associated image information on the web (Roh et al., 2019).

I posit human-centric computer vision as a domain that offers unique insights into how identity is conceptualized and embedded into technical artifacts across a number of dimensions. Firstly, the artifacts—the datasets and the models—produced for human-centric computer vision tasks are inherently imbued with aspects of human identity, whether through explicit classification (e.g., demographics, emotions) or through efforts of diversification (e.g., collecting data from specific demographics). Secondly, given that these conceptualizations of identity are sociohistorical (as I will describe in more detail in Conceptualizing Identity), there is an opportunity to understand *how* they are conceptualized by the workers developing them. Thus, in this dissertation, human-centric computer vision is both an avenue for understanding identity in computer vision artifacts and the values those communicate and how the positionality of workers shapes those artifacts.

Computer Vision Data Documentation

Given the importance of computer vision datasets to model outputs, there has been increasing attention on the process of gathering and documenting computer vision data. A number of scholars have noted a lack of consistency in documenting data, which has also made it difficult to assess data gathering practices (e.g., Holland et al., 2018; Miceli et al., 2021; Paullada et al., 2021). Stemming in part from the lack of transparency of both the provenance and the contents of many machine learning datasets, several dataset documentation frameworks have been proposed in recent years. These different proposals have varied goals and stem from different academic communities, with many different monikers: datasheets, data statements, dataset nutrition labels, and dataset requirement documents. However, they are united in understanding the different ways that dataset development can affect the outcomes of machine learning systems.

Gebru et al. take the inspiration for their framework, datasheets for datasets, from datasheets in the electronics industry. The authors provide a long list of questions to ask of each dataset, including motivations, composition of the dataset, the data collection process, the preprocessing and labeling processes, and the use and distribution of the data (Gebru et al., 2021). Holland et al. remodel the nutritional label used to report information about the nutritional value of foods, and provide a web tool to facilitate the creation of data nutritional labels (Holland et al., 2018). Bender and Friedman propose a similar documentation method: data statements, which is documentation specifically geared towards natural language processing (NLP) datasets(Bender & Friedman, 2018). Drawing on value-sensitive design (Friedman, 1996), Bender and Friedman compel NLP data authors to include language variety, speaker and annotator characteristics (such as “presence of disordered speech,” “native language,” and “training in linguistics”), speech situation, text characteristics, and recording quality. Geiger et al., meanwhile, do not provide a formal diagnostic or checklist for the construction of dataset documentation but form one implicitly by analyzing a set of papers focused on social computing (Geiger et al., 2020). Coming from the tradition of structured content analysis from the social sciences (e.g., Krippendorff, 2018), Geiger et al. code papers for several items, including whether the data used human annotation, if they had come from in-house or crowdsourced annotators, if the annotators were compensated, if they had training and if the instructions are available, if they used an interrater reliability metric, and if the dataset is available. They found that most of this information is not available for the datasets reported. Hutchinson et al. have proposed a data reporting framework that follows engineering principles of iteratively creating design, requirement, and maintenance documentation with the participation of many stakeholders across the data lifecycle (Hutchinson et al., 2021). Miceli et al. and Afzal et al. provide summarizations of different documentation frameworks’ defining characteristics (Afzal et al., 2021; Miceli et al., 2021).

The above scholarship is primarily focused on improving the documentation practices for computer vision so that those practices can be replicated, understood, and scrutinized, ideally opening doors to understanding data requirements and the technical components of collection and annotation. Scrutinizing the methods by which data was collected and annotated should help researchers better understand how the data being used might affect the outcomes of modeling decisions. However, the above scholarship is also largely focused on *doing* documentation, rather than studying how datasets, and other computer vision artifacts, are created. In its focus on producing documentation for computer vision artifacts, this approach to scholarship also focuses on specific *types* of documenting. Namely, facts about the process of creating artifacts, rather than the subjectivity of producing them. In this dissertation, I focus on human-centric computer vision as not only a set of artifacts—models and datasets—but as a set of value-laden practices shaped by the positionality of the tech workers developing them.

Machine Learning in Industrial Contexts

Building on the rich history of research on corporate settings in computing (e.g., Kogan & Muller, 2006; MacKay, 1999; Woolgar & Suchman, 1989), many scholars have begun to examine the practices of machine learning practitioners in corporate settings. For example, some have focused on current state-of-the-art uses of machine learning in specific industries, like oil and gas (Pandey et al., 2020) and building (Hong et al., 2020). Paleyes et al. mapped the challenges to deploying machine learning products (Paleyes et al., 2022). Kumar et al. interviewed industry practitioners about practices for securing their machine learning assets, identifying a lack of established practices for dealing with security threats (R. S. S. Kumar et al., 2020). Such work focuses on the perspectives of industry stakeholders, to identify the barriers facing industrial machine learning and developing mechanisms and frameworks for improving development processes. As I will show in this dissertation, working directly with industry practitioners is a fruitful method for uncovering their practices and identifying unknown challenges.

Given the power industry has over the AI landscape (L. Irani et al., 2019), it is unsurprising that scholars have also focused on the implications of corporate machine learning. Corporate models are not only more powerful than research models, given the economic power of big tech companies, they are also deployed in real world scenarios (e.g., Crawford & Schultz, 2019; Hawkins, 2017; Slota et al., 2020; Valentino-DeVries, 2020). Through examinations of these models, many researchers have discovered troubling outcomes. In particular, biases against certain groups have become a major area of concern.

For example, Buolamwini and Gebru famously uncovered bias against women with darker skin tones in corporate computer vision gender classification models (Buolamwini & Gebru, 2018). I similarly found biases against transgender and non-binary people in the same types of model (see Chapter 4). Noble critiqued Google's search algorithm for reinforcing racist stereotypes in search results about Black girls and women (Noble, 2018). Chen et al. found that men had significantly improved results when it came to job rank in resume search engineers (Chen et al., 2018). Many other machine learning biases have also been discovered by users themselves, such as the notorious examples of Google Photos labeling Black faces as "gorillas" (Barr, 2015) and women's resumes being passed over for men in Microsoft's resume parsing system (Jeffrey Dastin, 2018). Beyond biased outcomes, issues of transparency (e.g., Ananny & Crawford, 2018), accountability (e.g., M. Khan & Hanna, 2022), and data rights (e.g., Contractor et al., 2022) are at the forefront of conversations about industrial AI.

In response to these concerns, companies have increased their focus on FATE (fairness, accountability, transparency, and ethics) for machine learning. Beyond simply building machine learning models and deploying them, companies have developed dedicated teams aimed at researching and developing fair machine learning methods. Representative of this trend are Google's Responsible AI, Microsoft's FATE, and IBM's Trustworthy AI. As such, research on machine learning in corporate settings is also increasingly focused on the practices around ethics and fairness. For example, Holstein et al. identified the technical and organizational barriers preventing industry practitioners from effectively improving machine learning fairness (Holstein et al., 2019). Rakova et al. similarly identified constraints to enacting fairness initiatives, offering aspirational future processes to better enable effective initiatives (Rakova et al., 2021). Given the lack of trust in machine learning, Passi and Jackson conducted an ethnographic investigation of a large new media organization to understand how data scientists foster trust in applied data science in corporate settings (Passi & Jackson, 2018).

However, perhaps due to issues of access to industry settings, particularly large company settings, research on fairness practices in companies is still sparse. The study at hand builds on the growing body of work focused on industrial level machine learning practices. However, rather than focusing solely on the concept of preventing bias or improving fairness in developed products, this work focuses on how industry practitioners conceptualize identity characteristics for computer vision throughout the development lifecycle.

Conceptualizing Identity

Social Theories of Identity

Identity is a complex phenomenon—so complex, that it has been theorized and re-theorized by scholars across a wide range of disciplines, from feminist theory (e.g., Butler, 1988) to philosophy (e.g., Alcoff, 2006) to biomedical sciences (e.g., Repo, 2015). Scholars and theorists from diverse fields define identity in different, sometimes divergent ways. Identity can refer to an individual's personal identity, their social identity, their professional identity, or their cultural identity. Governments often characterize identity as measurable, as in demography, the sociological study of populations (Veron et al., 2006)—though this too varies across different nations and cultures. Others focus on the multiplicity of ways complex human identities form, as seen in Erikson's theory of psychosocial development (Erikson & Erikson, 1982) or in Marcia's identity status theory (Marcia, 1966). Other theorists focus more acutely on the social constructions of identity defined through cultural discourse. For example, Judith Butler posits that gender exists as an inescapable yet socially constructed entity, upheld through discursive beliefs about how gender is and ought to be performed (Butler, 1988). Gender is not a fixed or biological category; rather, it is a performative act rooted in "social regulation and control" (Butler, 1988). Similarly, trans scholars such as Jack Halberstam have explored the nuances of masculinity and its construction in society, and how certain masculinities are socially and infrastructurally policed when they don't ascribe to a gender binary (Halberstam, 1998). The multitude of theories, contrasting and overlapping, showcases that identity—whether at the individual level of the self or the broader social level of communities and groups—cannot be collapsed into a single universal definition. Yet, as I will argue in Identity in Computer Vision and Machine Learning FATE, computer vision has largely attempted to collapse human identity into simplistic visual data representations.

When considering identity in the context of computer vision, it is useful to consider how identity can be divided into two perspectives: the invisible and the visible. The invisible focuses on the individual self in the form of the mind or human consciousness. Some scholars argue a mind-body dualism, often attributed to Descartes and called Cartesian dualism (Cunning, 2011). The mind-body dualism philosophy argues that the mind and body can be viewed as entirely separate and non-influential to one another. For example, early philosopher John Locke argued that one's personal sense of self might belong to the "consciousness," having nothing to do with visible physical embodiment (Locke, 1689). Modern philosophers have built off early theories of mind-body dualism to argue, for example, that human consciousness is so separate from the human body, that it is autonomous of physical properties (Chalmers, 1996). In contrast to those who argue for the mind-body dualism, is physicalism: the belief that human identity and consciousness is purely physical, although the processes are still invisible to the eye. The contrast in theories about the source of internal invisible identities showcase the difficulty, if not impossibility, of understanding the true essence of the self. However, even while the source of "invisible" human characteristics is debated, different theories showcase the importance of the internal characteristics of human identity that are otherwise viewed as unobservable: thoughts, feelings, opinions, logic, and so on are core to how humans navigate the world.

In contrast to both early scholarship on individual identity and philosophical debates around mind-body dualism, scholars such as Husserl, Butler, and Alcoff assert that the embodied self is central to the development of internal awareness and one's relationship with the world (Alcoff, 2006; Butler, 1988; Monticelli, 2002). Unlike physicalists, they do not seek to argue that internal characteristics are inherently biological. Instead, they focus on the social meanings assigned to bodies and how those social meanings shape the invisible identities we hold. This realm of work focuses on the connection that the visible self has to the internal experience. Alcoff, in particular, embraces notions of visual embodiment, discussing the significance of visible markers of identity for race and gender in discussions of social identity, particularly in opposition to perspectives that seek to erase race and gender from political discussions (Alcoff, 2006). Like theorists focused on the source of invisible identity, theorists focused on embodiment also disagree about how divorced theory should be from the body. For example, Namaste critiques the overly philosophized perspective on gender, centering trans people's lived and embodied experiences which are often erased, or made invisible, from gendered theory (Namaste, 2000). Namaste criticizes prominent Anglo-Saxon feminist theorists, including Butler, for using the transgender body as a tool to ask epistemological questions about gender as a concept (Namaste, 2000). Scholars focused on the embodiment of identity consider how race and gender, as well as ability, sexuality, and other appearance-based markers intersect with and shape the human experience. Both Alcoff and Namaste argue that embodiment—how the experience of living in a certain body—is fundamental to how we interpret and interact with the world. Feminist scholars have called this perspective *positionality.* Positionality is reflective of one's sense of self and how their embodied form acts as a point of intersection between the self and the world (da Silva & Webster, 2018). One's sense of identity, perspective, values, and reasoning are informed by a complex set of both the visible and the invisible.

Stuart Hall assesses how anti-essentialist identity theories fail to provide any better theory for human identity (Hall, 2012). For example, many feminist theories, which often disagree with the theories I presented as "invisible identity," still fall short of providing any reasonable explanation for human identity. However, Hall also argues that perhaps trying to understand human identity at its essence is not possible or useful to meaningfully engaging with the outcomes and experiences of identity (Hall, 2012). Other domains, such as demography—or computer vision—are not focused on understanding any essence of identity or the self, whether invisible or visible. Instead, they are focused on classifying identity for instrumental means. Hall thus gives us the lens of *identification* through which to problematize identity classifications (Hall, 2012). To Hall, identification signifies the "process of articulation, a suturing" (Hall, 2012).. The suturing occurs between the discourse attempting to hail every individual into specific social practices (e.g., a specific gender group and expected gendered behaviors) and the processes which produce individuals as subjects (with consciousness and agency). It can mean an identification *with* (for example, a shared history) or an identification *of* (for example, an assigned sex). He describes the construction of identity categories within "specific modalities of power," for which identities like race and gender are employed for discursive means. While the theories of identity presented above showcase the diversity of theories on identity, and thus the overlapping and contrasting perspectives from which identity can be defined, Hall's identification provides a lens for understanding how positionality is embedded within computer vision for specific discursive means. We can use identification to understand how human actors define identity from a specific modality of power, or positionality.

The above theories are useful for understanding the overlapping and contradicting schools of thought on the essence of identity, if only to showcase how applied domains like computer vision simplify and conflate identity into simplistic instrumentations. The lens of identification opens interesting avenues for viewing perspectives of identity as it pertains to digital technologies. Specifically, how technologies are designed with the identification of human identity features in mind. Identification can be used to understand how identity is defined for specific uses in computer vision. Meanwhile, individual positionality might represent the suturing between the visible and invisible, how one's experience embodying a specific identity group (e.g., a sociocultural group like gender or a professional group like engineer) might lead to that person seeing the world in specific ways (e.g., how they would classify a person's gender for data annotation). In the development layer, an engineer might decide that classifying gender is useful for a computer vision application. The engineer then passes their conception of utility, informed by their positionality as an engineer, to an annotator, who labels a person in an image as male or female. The annotator, who assigns the label using their own conception of gender, engages in a moment of identification, suturing a person into a specific identity classification for the engineer's use. However, unlike interpersonal means of identification, like perceiving a person as male or female on the street, this identification is then applied in large scale to a number of unknown strangers when computer vision is deployed.

In this dissertation, like Hall, I am less concerned with the *source* of either invisible or visible identities. Instead, I focus on how individuals enact those identities within the workplace to classify the identities of others for technologies. In the context of individuals assigning meaning to human-centric computer vision, the invisible is relevant not only to how individual workers make subjective judgments and decisions (e.g., by deciding gender should look like a binary in a system), but how some identities which are otherwise invisible (e.g., gender as an internally held identity or emotion) are treated as visible when designing computer vision classifiers. The visible identities become an anchor point through which external actors—tech workers—enact their own positionality and assign an identity—what Hall calls "an identification *of*." These identifications communicate a specific modality of power, where human identity is constructed and deployed for computational purposes and to meet specific, usually capital desires. Thus, rather than focus on uncovering any specific source of identity (a task that philosophers continue to attempt) in computer vision, I turn attention to understanding how identity is conceptualized in computer vision artifacts and how those identifications are shaped by the positionalities of those developing them.

Identity in Technology

I am seeking to explore how classifications of identity, as seen in computer vision, coalesce in the development process. Specifically, I am interested in positionality—how identity affinities, like gender and race, but also their experiences in their work roles, their values, and their relationships with others and their companies, shape their approaches to identity classifications for computer vision systems. Understanding how human identity intersects with technologies has been central to explorations of experiences with a variety of digital technologies. Within social computing scholarship, two perspectives of socio-technical identity have emerged: social identity and technical identity. Social identity work has focused largely on the experiences users have when interacting with technologies. For example, Ammari et al. explored the performance of fatherhood in online do-it-yourself communities (Ammari et al., 2017). Haimson et al. examined the practices of disclosure trans users engaged in on Facebook, and the stress associated with it (Haimson et al., 2015). Similarly, I previously investigated how trans users navigate safe and unsafe social spaces online (Scheuerman et al., 2018). On the other hand, technical identity research concentrates on how identity is represented through system affordances—or, as Leavitt defines in (Leavitt, 2015), the "technical implementation of an individual's presence within a sociotechnical platform." Leavitt explored the concept of temporary technical identities by examining the practices of Reddit users who make temporary accounts for the explicit purpose of later abandoning them (Leavitt, 2015). Brubaker and Hayes examined the different uses of persistent identities on Facebook vs. "single-use" identities on Craigslist, documenting how system representations supported social interactions, representing user relationships through differing affordances (Brubaker & Hayes, 2011).

At the center of both the social and the technical lenses adopted in social computing is the reality that, as we build technologies that intersect with humans, whether to be directly used by them or to output information to be used by them, the complexity of identity must be packaged into more simplified data. To make identity useful for accomplishing specific goals, identity requires classification. For example, birth certificates are generally issued with a gender based on the visual determination of gender from a doctor when a baby is born. Alongside this is a long history of gender taxonomies in health defining gender categories by visible characteristics (Fausto-Sterling, 2000). Yet, as demonstrated by the theories about human identity presented in the prior section, categorizing, classifying, and databasing the complexity that is human identity into information systems is laborious and often muddled.

Transforming identity into something interpretable is a suturing (Hall, 2012), or what Koopman refers to as a *fastening* (Koopman, 2019). Koopman argues that human beings are now inherently informational persons, as we are all “inscribed, processed, and reproduced as subjects of data” (Koopman, 2019). Human identity is so tied to data that the loss of data means the loss of access to human rights. In their work, Bowker and Star similarly highlight the cultural, political, and historical decisions underlying the creation of classifications and standards, showing how political agendas have allowed governments to deny rights to certain classifications, like Black South Africans during Apartheid. They describe the process of fitting complex human identity into simplistic databases as *torquing* (Bowker & Star, 2000). Torque describes how classification systems introduce tension into the lives of the individuals being classified—when “the ‘time’ of the body and of [its] multiple identities cannot be aligned with the ‘time’ of the classification system (page 190)” (Bowker & Star, 2000). As humans are fastened to databases, becoming tied in integral ways to their data, their identity is torqued, simplified and forced to fit into political agendas made technical in uncomfortable and often painful ways.

Numerous social computing researchers have inspected the experiential results of classification in computing architectures. For example, Blackwell et al. describe the marginalization of users whose experiences with harassment are invalidated when they meet rigid classifications (Blackwell et al., 2017). Harrell similarly examined how stereotypes and stigmas are reified in games and social media websites (Harrell, 2009). Phillips connected past physiognomic practices to shaping racial models in video games (A. Phillips, 2020). These social computing studies highlight how suturing identity to computational systems minoritizes specific experiences and results in torque. In an attempt to flip the power of the majority often imbued in technical infrastructures, Feinberg et al. employed a critical design perspective to privilege the “others” that fall between categories in database infrastructures (Feinberg et al., 2014). Yet even such critical designs, in their attempt to undermine classical identity infrastructuring, reinforce the power embedded in classifying human identity at all.

Even while human identity needs to be simplified and made stable for technical infrastructures, institutional classifications of identity are constantly shifting; they are not solidified remnants of singular past decisions that we continuously utilize. For example, in the United States, previous United States president, Donald Trump, issued an executive order to extend racial and ethnic classification to those with Jewish ancestry in an effort to curb anti-Israel protests on college campuses under Title IX (Dias et al., 2019). Similarly, Rodríguez discuss the fluid history of race/ethnicity classification of Latinx people in the United States census (Rodriguez, 2001). Furthermore, classification systems are often implemented unevenly, differing across jurisdictions. For example, while some states in the United States allow for non-binary genders on birth certificates (e.g., Savage, 2019) and driver's licenses (e.g., Dance, 2019; Schmelzer, 2018), it remains impossible to change gender markers on any identifications in other states. One might be able to change their gender marker on their birth certificate if they were born in Colorado but be unable to change their gender marker on their birth certificate if they were born in Tennessee.

Computer vision, like other digital technologies, has become a site of critical analysis in regard to human identity. Concern about computer vision largely stems from the degree to which computer vision technologies intersect with human identity, at the level of the individual and larger sociocultural groups. Beyond other digital technologies, like social media websites or video games, computer vision is of particular concern, given its fundamental reliance on vast amounts of historical data, the difficulty of interpreting blackbox inferences, the ability to deploy it opaquely, and the attempt to deploy general models across multiple contexts. Identity in computer vision technology has largely focused on one level of the infrastructural layers proposed in this dissertation: the artifact layer. Specifically, researchers are concerned with issues of individual privacy and fairness for different visible sociocultural identity groups, like race and gender. In the next section, I discuss current approaches to identity in computer vision and attempts to make it fairer, and how those approaches might be extended to include examinations of how positional subjects fasten specific identities to computer vision infrastructures.

Identity in Computer Vision and Machine Learning FATE

Human-centric computer vision focuses on making visual data about humans interpretable to computational models. Identity in computer vision, like in other technical infrastructures, is portrayed as static and immutable. While computer vision learns from new examples, it only learns how those new examples fit into a set schema. Given that computer vision uses visual data, like images and videos, identity in human-centric computer vision largely focuses on the visible. For example, facial recognition is designed to recognize an individual person from their facial images, indicating a tie between a person's facial structure and their identity, in terms of their name and, in some cases, like those of police surveillance, associated records. In other cases, computer vision systems are designed to make the invisible visible. Some commercial facial analysis companies have adopted notions of physiognomy, such as Faception, which attempts to tell internal characteristics, like IQ and criminality, from facial morphology (McFarland, 2016).

Different perspectives on certain social identities have also emerged which highlight the tension between the visible and the invisible. For example, gender in computer vision is often binary, male and female, and built on the assumption that gender is a visible characteristic. Such a view aligns with how gender has been portrayed historically, in sexology (Fausto-Sterling, 2000) and even second-wave feminism (Repo, 2015). However, gender is often, and increasingly, theorized as an internally held concept, which is tied to embodiment in that it is enforced as a visible construct which certain bodies are hailed to perform (Butler, 1988). Race, also, is operationalized in computer vision as visually obvious, but has historically been difficult to classify, even during regimes of oppressive classification (Bowker & Star, 2000).

An increasing focus on FATE (fairness, accountability, transparency, and ethics) in machine learning has led scholars to critique certain approaches to identity in computer vision. In some cases, the focus is still on the visible, accepting that identity can be classified from visual data but ensuring that systems work fairly for every classification. Fairness research, in particular, is concerned with classification parity and mitigating bias in human-centric computer vision. Bias has been conceptualized in various ways by researchers—for example, statistical bias that leads to skewed results (Das et al., 2019; Jacobs & Wallach, 2019) and representation bias stemming from historical prejudice (Howard & Borenstein, 2018; Mehrabi et al., 2019).

Bias in algorithmic contexts can cause widespread, real-world harm. The Future Privacy Forum released a report categorizing the numerous harms that can result from algorithmic bias, grouped by individual-level harms (e.g., employment discrimination) and societal-level harms (e.g., differential access to job opportunities (Future of Privacy Forum, 2017)). Already, much research has been done to uncover bias in facial analysis systems. NIST conducted an evaluation of face recognition in 2019, finding that recognition systems tend to perform better on men and older people, than on women and younger people (Ngan & Grother, 2015). Klare et al. similarly discovered that models performed worse on women, as well as people who are Black (Klare et al., 2012). Buolamwini and Gebru found that facial analysis services, like those provided by Microsoft and IBM, had significantly higher gender misclassification rates for women with dark skin tones (Buolamwini & Gebru, 2018). Stereotype aligned correlations between gender and the activities being depicted in images have also been identified in several computer vision datasets (e.g., overrepresenting women in images depicting cooking and shopping (M. L. Hendricks & Testa, 2012; Zhao et al., 2017)). A recent audit of ImageNet found the dataset contained significant gender biases, and even the inclusion of non-consensual pornographic imagery, depicting predominantly women (Birhane & Prabhu, 2021). The American Civil Liberties Union (ACLU) exposed Amazon's Rekognition algorithm for incorrectly matched Black members of the U.S. Congress with mugshots of people who had committed a crime (Snow, 2018). The long list of examples of representation and performance bias in computer vision showcases that, when adopting a visible approach to classifying identities into discrete categories, how the data is organized to represent those categories is still crucial to ensuring equitable system performance.

To address issues of performance disparity for certain classificatory groups, solutionist approaches to bias such as bias auditing and bias mitigation have become more common. Machine learning researchers have proposed numerous statistical approaches (e.g., Das et al., 2019) and toolkits (e.g., Bellamy et al., 2019) for mitigating bias. As bias can manifest in numerous ways and in numerous places within a machine learning system, many scholars have begun considering its consequences. For example, Danks and London present a taxonomy of where algorithmic bias might appear in the pipeline (Danks & London, 2017): in the training data, in the focus of the algorithm, in the processing of information, and in the use of a single algorithm from one context to another. Such tools provide researchers and practitioners with clear steps to improve systems within the confines of the existing infrastructure.

Those concerned with fairness for different identity categories focus acutely on the representations of sociocultural groups in the data used to train and evaluate computer vision. In particular, they seek to increase the diversity of data across identity groups. One method that researchers use to assess whether the data being fed into automated systems is sufficiently diverse is to measure variation in facial landmarks: features of the face that are believed to be commonly associated with particular racial and gender categories. For example, IBM's Diversity in Faces dataset, which was created as a response to the lack of diversity in prior datasets, employs what researchers call "craniofacial science:" "[T]he measurement of the face in terms of distances, sizes and ratios between specific points such as the tip of the nose, corner of the eyes, lips, chin, and so on" (Merler et al., 2019). Such facial landmarks are not generated by facial analysis technologies themselves but rather are identified and used by human researchers to compare facial variation across categories of age, race, ethnicity, and gender. In doing so, researchers imply that certain craniofacial distances and shapes are objectively associated with specific races and genders, and that this diversity has thus been adequately accounted for by the model.

Other scholarship in the FATE space critiques computer vision not for identity bias between the classified groups, but for the method of classifying identity in the first place. Certain examples of classifications clearly showcase the subjective decisions being made about human identity. In an examination of the "person" categories within ImageNet—derived from the WordNet hierarchy (Fellbaum, 2012)—Crawford and Paglen found the inclusion of misogynistic terms, racial slurs, and otherwise offensive labels (Crawford & Paglen, 2019).[4] Birhane and Prabhu extended this analysis to other image datasets that have derived their categorical structure from WordNet, and found the TinyImages dataset also contained slurs and other offensive labels (Birhane & Prabhu, 2021).[5] Offensive labels, such as racial epithets and sexist denigrations reflect how socially constructed power structures can easily become embedded in technical infrastructures.

[4] Following the release of Crawford and Paglen's article, the ImageNet creators removed a subset of the person categories from the dataset.

[5] TinyImages has since been removed from the web.

Intentionally using computer vision to denigrate and harm marginalized groups has actually been put into practice. In China, Uyghur Muslim minorities, who are increasingly being detained in re-education camps, are subject to government surveillance by facial classification and recognition, trained explicitly to attempt to classify and track people who appear Uyghur (Mozur, 2019). Hamidi et al. also interviewed transgender individuals and found their participants were largely concerned about how facial analysis could be used for discrimination, due to known histories and contemporary political agendas that seek to discriminate against transgender communities, particularly trans communities of color (Hamidi et al., 2018).

The examples of race and gender, portrayed as inflexible and obvious from visual data, highlight how the very worldviews on identity in computer vision have become contentious. Scholars and practitioners are questioning the underlying social and moral judgments being made when detecting and classifying core human identities. Benthall and Haynes critique machine learning researchers and practitioners for approaching race as an "inherent property of a person," rather than as a social or political category (Benthall & Haynes, 2019). They argue that racial bias is likely to come from subjective human decisions during the collection and labeling process (Benthall & Haynes, 2019). Hanna et al. similarly critique current approaches to classify race in machine learning datasets for not accounting for the "socially constructed nature of race," also arguing that current fairness approaches fail to account for contextual and complicated social categories like race (Hanna et al., 2019)). Keyes, focusing on gender classification, states that computer vision that purports to read gender visually inherently erases the existence of transgender and non-binary people, given the conception of gender as immutable, visual, and situated wholly in the body sits in opposition with trans realities (Keyes, 2018). I argue in prior work that current constructions of gender in computer vision are reliant on discriminatory colonialist histories that reify racist and transphobic worldviews (Scheuerman et al., 2021). At the level of data classifications more broadly, Sen et al. problematize the broad use of so-called "universal gold standards" in benchmark datasets at all (Sen et al., 2015). In other words, there is no such thing as a universal and objective classification system.

To reiterate, identity in computer vision artifacts is only able to communicate via visuality, relying on visible aspects of identity that have been historically mapped to certain visible features, and, in some cases, making claims about what parts of identity are visible at all. The epistemologies of computer vision and FATE are often in contention. Whether FATE approaches sometimes embrace the visuality of identity to improve representation and diversity or oppose the essentializing of identity and view identity as incompatible with computer vision systems, FATE researchers disagree with the treatment of identity in computer vision as technical rather than social. Critiques about the lack of social context and variability in identity classifications map to identity theories that position identity as an embodied positional experience; identity is neither always visible nor can it be separated from larger sociocultural power structures.

In alignment with the perspective that identity is sociocultural, this dissertation focuses on the social context of identity in computer vision. In extending prior FATE research on artifacts, I focus on the worldviews about identity that computer vision practitioners have embedded into models and datasets. I examine artifacts not only as a source for understanding biases around identity attributes, but also as a reflection of the identities held by their creators. Thus, I also steer away from focusing solely on the visuality of identity, like in both computer vision research and many FATE approaches. I also examine the invisible aspects of identity, how the positionality of computer vision workers indicate how their underlying values about identity become embedded into computer vision systems. In the next and final section, I explicate the scholarship describing the role of values in artifact creation, highlighting examples of prior work on the values embedded in computer vision artifacts specifically.

Values and Positionality Instilled into Artifacts

Values in (Machine Learning) Artifacts

Scientific disciplines are encultured with their own specific practices and values. Philosophers of science have examined the relationship between practices and values of certain scientific disciplines for some time (Becher, 1987; Breeze, 2011; Cooper & Bowers, 1995), as well as how those values are shaped by and shape social life (Goldman, 2000; Salter & Martin, 2001; R. Smith, 2001; Winner, 1986). In this section, I first describe how science, broadly, has been philosophized as value-laden, often imbued with the specific agendas of those in a relevant discipline. I then describe how researchers have applied this lens to specific artifacts, like data structures, to demonstrate how artifacts can communicate deeper subjective values beyond what is presented. The purpose of this section is to present the perspective on disciplines from which I am conducting the work of this dissertation, and to showcase how prior researchers have adopted this perspective to analyze values in disciplines. I also highlight some work on computer vision artifacts focused on excavating values.

Philosopher Michel Foucault defined knowledge as a practice imbued with discursive power, in which a scientific "truth" is constructed through classificatory practices (Powell, 2015). He argued that the boundaries of truth, including language and thought, were confined to the specific cultural and temporal contexts (Foucault, 1970). Through complex social processes, scientific discourse might evolve, to allow for new accepted truths—what Kuhn calls a "paradigm shift" (Kuhn, 1962). Foucault would argue that paradigm shifts support the claim that every historical period—or perhaps, more granularly, sociocultural context—has an underlying "episteme," an acceptable form of discourse and truth.

Foucault's perspective is not universally accepted by all philosophers and theorists. For example, Foucault was focused on discursive power, rather than any analysis of the inherent improvement of science and might disagree that knowledge is always inherently improving, while Kuhn took a more positivist stance that science is inherently improved through paradigm shifts. However, Foucault's perspective underlines an oft shared worldview: that science is not inherently objective, but also shaped by and interpreted through human subjectivity (e.g., Kuhn, Polanyi, Habermas, Heidegger, etc.). The intricate webs connecting science with social and political institutions shape how we see and interact with the world, and what we accept as truth. In some cases, science has been used as a tool of social and political intervention with profound consequences. Foucault's analysis of sexuality (Foucault, 1976), Repo's analysis of gender (Repo, 2015), and Stoler's analysis of race (Stoler, 2014) reveal historical genealogies in which science has intervened to shape how certain groups of people are viewed, treated, and interacted with—in everything from academic theory to family life to medicine. Whether basic or applied, scientific and technological production is necessarily an exercise of judgment that reflects a specific episteme of the time, requiring a series of value judgments on what should be made visible and invisible, documented and undocumented.

Given that scientific disciplines have their own epistemes, science and technology studies are rich with analysis of the underlying values embedded in different technical infrastructures. Documentation has been viewed as a source of understanding these underlying values, through an analysis of both what is and what is not documented. Data, the categories used to organize the data, and how those categories are then applied all communicate the worldviews and values of those who constructed them.

At the level of data, Bowker examined how the data considered most interesting and classifiable by scientists shapes how data is then stored, and how data deemed uninteresting or difficult to classify can become underspecified or entirely lost in larger data catalogs (Bowker, 2006). Bowker argues that some data entities are overlooked “because they do not lead to spectacular science or good funding opportunities” (pg. 146). For example, in biodiversity, there are “charismatic” species that appeal to the public and scientific funding agencies, like the koala being more appealing to funders than a species of seaweed. The type of data collected for knowledge production is shaped by a number of social factors—from disciplinary interest to funding availability—situated within a specific context, such as whether the animal to collect data about is endangered in the Anthropocene.

At the level of categories, Ásta describes the process of categorizing social groups as inherently reliant on specific social contexts where “individual agents create and maintain social categories by the conferral actions of classifying and placing people in the contexts they travel” (Ásta, 2018). Categories are bestowed upon people, or data, and those categories then constrain the available actions or uses of those people, or data. In the context of system design, Edwards et al. argue that the categories upholding technical infrastructures constrain the available uses of those infrastructures (Edwards et al., 2010). Once more, the construction of categories, even for applied use, reflects a specific worldview of what the constructors deem relevant and important.

Finally, at the level of application of categories, Suchman argued that categorical decisions in system design had the potential to hold up specific social orders (Suchman, 1993). For example, Bowker and Star discussed the politics imbued in technical classification systems, like those of the International Classification of Diseases (ICD) and identity documentation employed in Apartheid South Africa (Bowker & Star, 2000). Through these two examples, they discuss how the categories chosen and utilized reflect specific epistemes. For example, the racial classifications used during Apartheid reflected sociohistorical and culturally situated beliefs about the racial superiority of whites. “Data” in the form of visual assessments of race, socioeconomic status, and family ancestry determined how individuals were classified and documented. Both the classifications and the underlying sociopolitical beliefs about race were predicated by scientific racism, previously accepted as scientifically valid.

While the above examples are not exhaustive, they show how prior researchers have approached understanding the underlying values and social contexts of research and development through documentation, and what those values and contexts might communicate about the episteme of the fields or disciplines specific documentation is situated within. Computer vision, as a data-driven field, is rife with opportunities to understand the underlying values of the discipline being communicated through documentation of the data and practices of researchers and engineers. For example, examinations of the practices of dataset development have exposed the widespread devaluation of data work. Sambasivan et al. found high-quality dataset development to be one of the most undervalued components of machine learning practice (Sambasivan et al., 2021). Dataset development is often omitted entirely from machine learning curriculums and textbooks (e.g., Goodfellow et al., 2016), upholding disciplinary norms that devalue dataset work, which is reflected in peer review processes that make it difficult to publish work focusing exclusively on datasets (Heinzerling, 2019). Jo and Gebru characterize the resulting culture of dataset development as one that embodies a *laissez-faire* attitude, which they contrast with the careful and critical curatorial practices of archivists (Jo & Gebru, 2020). Paullada et al. discuss how dataset culture within machine learning prioritizes speed for the achievement of algorithmic performance on a fixed set of benchmarks with little regard to the implications of data reuse, data management, and legal issues (Paullada et al., 2021).

Given the critiques of current computer vision documentation practices for being opaque, underspecified, and focused solely on algorithmic development, it could be argued that computer vision is undergoing its own paradigm shift. Technical researchers who have been focused on creating new methods for accomplishing computer vision tasks and improving the state-of-the-art of models through statistical approaches are coming up against increasingly constructivist and critical analyses of computer vision artifacts, practices, and impacts, which are now shifting the field away from purely technical to more interdisciplinary focus on real world implications (e.g., Ashurst et al., 2021). For example, Agre criticized approaches to artificial intelligence for undervaluing critical social theories (Agre, 1997). In some cases, this disciplinary shift has caused division amongst the old (technical) and the new (interdisciplinary) (e.g., Soper, 2020). Of course, even FATE approaches propose specific worldviews that impose specific values in designing computer vision artifacts. The earlier example in Computer Vision Data Documentation describing Holland et al.'s decision to reappropriate nutrition labels for machine learning documentation (Holland et al., 2018) are rooted in documentation that prioritizes a specific viewpoint and value judgment about health and dieting. Such disagreements about the role of human values in shaping the social machine learning systems, rather than simply technical ones, indicate a fundamental value misalignment between traditional computer vision and FATE-focused researchers.

Given the necessity of data to modern computer vision research and applications, this dissertation work focuses specifically on examining the underlying values of computer vision datasets—the practices of collecting, classifying (often, through annotating), and disseminating data to be used by researchers and practitioners for building and evaluating computer vision models. Vertesi and Dourish argue that the values in data arise in the nature of data production itself, and that such values are often only available through understanding the context of production (Vertesi & Dourish, 2011). Values reflect the positionality of those who are instilling them; in the case of this dissertation, values reflect decisions people make about how to represent identity. Computer vision practitioners might approach identity as objective, visually evident, drawing on positivist notions of observation as reflective of reality. In contrast, I adopt the theoretical perspective that scientific disciplines, researchers, and artifacts are imbued with specific subjective values. As such, I focus on both the artifacts to understand the values communicated through documentation and how the workers involved in creating those artifacts conceptualize those values.

Positional Values in (Machine Learning) Artifacts

Positionality refers to how an individual's "position" in the world shaped their outlook—how the complex web of identities like race, gender, nationality, location, sexuality, class, and more influence their experiences and thus beliefs, values, and relationships (da Silva & Webster, 2018). Such positions are not static or necessarily chosen, but mutually constituted through one's relationship with others and also themselves (Collins, 1990; Crenshaw, 1991). As described by Iris Marion Young, "one *finds oneself* as a member of a group, which one experiences as always already having been" (Young, 1990) (emphasis in original). From positionality comes a specific epistemic standpoint, a socially situated way of viewing the world (Haraway, 1988; Harding, 2004; Rolin, 2009).

From positionality comes a specific epistemic vantagepoint, a socially-situated way of viewing the world (Collins, 1998; Haraway, 1988; Harding, 2004; G. Rose, 1997). Such epistemic vantage points are theorized through feminist standpoint theories, the theory that all views come from somewhere and no view stems from nowhere (da Silva & Webster, 2018; Rolin, 2006). Standpoint theorists argue that those occupying some positions may be more knowledgeable on certain subjects than those occupying different positions. As Wylie writes, some individuals "may know different things, or know some things better than those who are comparatively privileged (socially, politically), by virtue of what they typically experience and how they understand their experience" (Wylie, 2003). Generally, standpoint theorists approach situated knowledge through power. Women understand misogyny in ways that men cannot; Black women understand misogynoir in ways that white women and Black men cannot (Collins, 1998; Crenshaw, 1991). In contrast, a view from nowhere would posit some objective and observable truth about the world—and human identity—that can be captured in an unbiased manner.

Linda McDowell argues that "we must recognize and take account of our own position, as well as that of our research participants, and write this into our research practice" (McDowell, 1992). Feminist scholar Rose posits a "reflexivity that aims, even if only ideally, at a full understanding of the researcher, the researched and the research context" (G. Rose, 1997). Rose highlights the uncertainties in this goal—to fully account for the positionalities of all actors in a research project—are not deficits, but rather opportunities for more transparently understanding the limitations of research and why they might occur (G. Rose, 1997).

When it is attended to in computing research, positionality is largely attended to through reflexivity—a methodological process of self-reflection on how researcher and research mutually construct one another. Positionality statements have become increasingly commonplace in social computing (Liang et al., 2021). Understanding how research subjects express their positionalities is often implicit. In the realm of machine learning, many scholars have examined how human values have become embedded in and shape data categories. For example, I document the types of values driving the creation of datasets for computer vision (see Chapter 5). Hanley et al. interrogate the normative dilemmas when describing identity categories in alt text for people who are blind (Hanley et al., 2021). Metcalf et al. found that many practitioners at the forefront of ethics in industry tolerated corporate values, like market fundamentalism and technological solutionism, for the sake of minimal ethical impact (Metcalf et al., 2019).

Numerous scholars have also critiqued the underlying values governing identity characteristics in machine learning. Much like Suchman argues (Suchman, 1993), language categories are explicitly designed to maintain current status quo social orders, not challenge them or provide space for social action. As I explicated in Chapter 3, how the construction of race and gender categories for computer vision datasets reflects normative beliefs that identity is "insignificant, indisputable, and apolitical." Hanna et al. similarly argue that the treatment of race as categorical erases the reality that race is socially constructed and meaningful (Hanna et al., 2019). Such expressions of values are reflective of the larger positional standpoints occupied by the designers of machine learning artifacts, like datasets.

As Davis writes, a "dataset is a worldview" (H. Davis, 2020). The worldview that a dataset holds is the result of the worldviews of the humans working to produce it. In the case of computer vision datasets, how an identity like gender is represented reflects how the humans working with the data think about gender. The presence of identity characteristics in computer vision—like gender classifications, labels applied to images, inferred racial demographics from textual data—indicates that, at some point, identity characteristics are *developed* for machine learning. Going beyond Davis's statement, the humans instilling worldviews into datasets are also not a homogenous group; they bring different perspectives and privileges to the table during their work. Thus, beyond building on studies of industrial practice, I also contribute to burgeoning research on how different stakeholders structure and define identity categories for machine learning. Specifically, I extend this area of inquiry by examining how industry practitioners embed their own perspectives about identity in the process of defining it.

Many scholars have examined the role of human subjectivity in shaping data (e.g., Cheng & Cosley, 2013; Feinberg, 2017; Hanna et al., 2019; Muller et al., 2021). For example, Vertesi and Dourish propose a data economy framework for CSCW aimed at exposing how the context of dataset production instills datasets with specific values and meaning (Vertesi & Dourish, 2011). They examine how social relationships between organizations are shaped by the ways data is produced. In Chapter 5, I describe how computer vision dataset authors in research contexts sacrifice data work in the name of model work, prioritizing efficiency over care and erasing contextuality and positionality in data practices. In Chapter 3, I also argue that the lack of engagement with documenting how identity-based labeling decisions are made during data work makes datasets less trustworthy and promotes an unbiased worldview about social categories like race and gender. Denton et al. propose a genealogical methodology for researching machine learning datasets, "for investigating how and why these datasets have been created, what and whose values influence the choices of data to collect, the contextual and contingent conditions of their creation" (Denton et al., 2020). Promoting reflexivity among machine learning researchers and practitioners, a core tenant of this genealogical methodology is focusing on the role of human values in the creation of datasets.

Other scholars have focused on understanding the role of data worker subjectivity on datasets, specifically. For example, Hube et al. found that workers with "strong opinions" tended to produce biased annotations (Hube et al., 2019). Patton et al. examined the differences between domain experts and graduate students who annotated Twitter data from African American and Latino youth and young adults; they argue that disagreements between the two annotator groups emphasize the importance of annotator background, particularly "nuances in culture, language, and local concepts" (Patton et al., 2019). Sen et al. conducted a survey study to see whether Amazon Mechanical Turk workers from different cultural communities produced different ratings on the same data (Sen et al., 2015). Beyond finding that different communities produce different rating labels, they also found that algorithms trained on datasets sourced from different communities perform dramatically differently. Similarly, Dong et al. compared how Chinese and American participants apply image tags to movies and found distinct differences between these two nationalities (Dong et al., 2021). While Americans largely applied what the authors refer to as “factual tags” relevant to describing the films, Chinese participants preferred to apply “subjective tags” more relevant to personal opinion. Participants were also more likely to choose tags sourced from their own culture in the survey design. The authors use their results to suggest that designers attend to the cultural “deficits” introduced by culturally contingent taggers. Litman et al. found that lower compensation rates for platform tasks led to lower quality data work, particularly in India (Litman et al., 2015), showcasing that the geolocation context and economic conditions of data workers is highly influential on their data production. These studies on datasets engage with how the worldviews of dataset authors and annotators shape dataset outcomes, even if they don’t explicitly label these worldviews as positional. Data, no matter how simple it appears, does not simply exist—it is designed.

Positional perspective is generally bound up in tacit knowledge. In contrast to explicit knowledge, tacit knowledge makes up the skills and ideas we gain from our experiences yet have difficulty articulating or formalizing (Polanyi, 2009). Despite the importance of tacit knowledge in work practices, it is often difficult to capture. Understanding the role of tacit knowledge in work practice has thus been a major focus of CSCW (e.g., Mtsweni & Mavetera, 2018; Reeves & Shipman, 1996; Tavanti et al., 2006). In this work, I attend to the explicit and tacit knowledge that workers had of their own positionalities and the role it played in conducting their work. I align with Rolin’s perspective on positional standpoints (Rolin, 2009), rejecting that knowledge of a subject is neither biased or unbiased, correct or incorrect, but instead operates from a specific social position that influences how they view the world in the data and models they are working with.

PART ONE:

ARTIFACTS

The work presented in this section is focused on how identity is represented in artifacts and how artifacts can communicate the underlying values of their creators. Here, I present three studies:

1. In Chapter 3, I show how identity is embedded into computer vision datasets. Specifically, I present work that I conducted on how race and gender are defined and explained in datasets. I show that authors rarely engage with underlying sources for defining identity categories, justify their use, or explain how they went about labeling each image. I critique how dataset authors ignore the sociohistorical context and fluidity of race and gender, and instead present race and gender as obvious, static, and apolitical.
2. In Chapter 4, I show how identity is embedded into computer vision models. Specifically, I present work that I conducted on how diverse genders are classified by commercial computer vision models. I examined gender in both the gender classification outputs of these models, but also in the image labeling outputs. I show that: (1) transgender individuals are disproportionately misclassified in commercial computer vision; (2) non-binary individuals can never be accurately classified due to an underlying gender binary in all commercial computer vision models; and (3) that a binary and reductive gender perspective also shows up in image labeling.
3. Finally, in Chapter 5, I examine how technical artifacts communicate the values of their human creators. I present work that I conducted on how computer vision dataset documentation showcases the implicit values of the authors who created them. I argue that computer vision dataset documentation communicates the values of efficiency, universality, impartiality, and model work. I argue that the oppositional values of care, contextuality, positionality, and data work are thus silenced.

3
HOW IDENTITY HAS BEEN EMBEDDED IN DATASETS

Image detection and classification represents a pertinent domain where I see a tight coupling of human identity and computation. Perhaps the most salient example is *automated facial analysis technology* (FA) (Buolamwini & Gebru, 2018), an umbrella term for computer vision methods that use machine learning (ML) techniques to automate problems related to reading the human face (S. Z. Li & Jain, 2011). FA is often discussed in the context of two specific tasks: facial detection and facial recognition. Both use computational methods to measure the human face, whether simply to detect that a face is present (i.e., facial detection) or to detect a specific individual's face (i.e., facial recognition).

Facial analysis technology—the machine learning (ML) approach to determining information about human faces—is just the latest tool in a long history of tools used for classifying human identity. Facial classification is used to target marketing campaigns at specific demographics (e.g., Pomranz, 2017; Robitzki, 2019; Sharma et al., 2007) and to track physical consumer behavior inside stores (e.g., (e.g., *Clarifai*, 2019; Huang et al., 2006; A. Lin, 2017). In this chapter, I have chosen to focus specifically on facial analysis systems in the broader realm of computer vision because race and gender are regularly embedded into these systems.

Race and gender have become two of the largest concerns regarding bias in machine learning fairness literature—particularly, how systems are biased against certain races and genders (Abdurrahim et al., 2018; Grother et al., 2018; Klare et al., 2012) and how to mitigate those biases (Buolamwini & Gebru, 2018; Gong et al., 2019; T. Wang et al., 2019). These concerns include bias in the databases used to train and evaluate machine learning algorithms (e.g., Danks & London, 2017; Mehrabi et al., 2019; Tommasi et al., 2017) and the very morality of facial analysis use cases (e.g., Bacchini & Lorusso, 2019; Marciano, 2019; Wevers, 2018). Sample selection bias—bias resulting from what subjects are included in a database—is a known issue in machine learning databases (e.g., Mehrabi et al., 2019; Torralba & Efros, 2011), leading computer scientists to try to mitigate for it using various methods. For example, algorithms for exploiting database bias to improve those databases (e.g., Khosla et al., 2012) and for creating "unbiased" models from known biased data (e.g., Kamiran & Calders, 2009).

Such attempts to mitigate bias and build more diverse databases are invaluable to creating fairer outcomes. However, despite increasing attempts to diversify databases, approaches remain simplistic and lacking in critical and social theories. ML and human-computer interaction (HCI) communities do not have an agreed upon approach to how diversity is being operationalized in training and evaluation databases. New databases are seeking to fill gaps with more images without a deeper engagement of the categories of race and gender themselves or the ethics of collecting that information (e.g., Google contractors targeting homeless people of color for face images (Hollister, 2019)). While some scholars are questioning the politics of identity representations in facial analysis classification infrastructures (e.g., Keyes, 2018) and myself (Chapter 4)), there has been little inquiry into the assumptions authors of facial analysis databases have made when collecting and annotating data. This is a major obstacle in meaningfully representing race and gender in databases, resulting in databases that are opaque and inconsistent. To truly understand the available outcomes of facial analysis models, it is imperative to understand the underlying decisions embedded into the construction of training and evaluation databases.

Like Benthall and Hayes (Benthall & Haynes, 2019), I examine race—and gender—as socially constructed categories machine learning has failed to critically engage with. I approach my analysis from a critical discursive perspective. Specifically, I investigate *how* race and gender are codified into image databases. To do this, I analyze how race and gender are represented in image databases and how those representations are derived. I focus on answering the following research questions:

- What *purposes* do the authors of image databases intend their databases to be used for and how does that shape their use of race and gender?
- What information about race and/or gender are *implicit* (i.e., the authors describe the demographic distribution in a database, but each image is not annotated with race and/or gender)

and what information is *explicit* (i.e., each image in a database is annotated with race and/or gender information)? What are the *categories* being used to define race and gender?

- What *sources* are being used to derive race and/or gender in both implicit and explicit databases?
- How are database authors describing the *annotation procedures* for explicitly annotated race and/or gender categories?

To identify relevant databases for analysis, colleagues and I[6] created a corpus of machine learning literature on facial analysis technologies and manually coded them for which databases are referenced. We used this corpus to identify a sample of 92 image databases, whose documentation we examined to answer the outlined research questions. We started by analyzing the database documentation to identify the motivations authors provided for the use of each database—in other words, what each database was created for. Understanding these motivations provided context for the intended uses of race and gender in facial analysis systems. I found three database use cases: (1) individual face recognition and verification; (2) image labeling and classification; and (3) providing diversity for model training and evaluation. I then surveyed both (1) implicit race and gender information; and (2) explicit race and gender annotations. I chose to analyze both implicit and explicit descriptions as both have implications for database use, value, and potential bias.

Within both implicit and explicit race and gender categories, I found two diverging themes. For race, I observed no consistent classification schema; the classification of race and the way race is discussed by authors varies greatly. For gender, I observed numerous instantiations of the same “male” and “female” binary categories. I found that the vast majority of image databases (1) do not utilize sources (e.g., make use of existing resources, like prior literature) for defining race and gender categories; and (2) do not document the process of annotating images for race and gender categories.

Given that image databases are used as a resource on which facial analysis systems are built and evaluated, I argue that the field of computer vision needs to adopt more standardized methods for using and documenting race and gender. I posit facial analysis as a digital form of otherwise familiar classification technologies to critique current approaches in image databases for their lack of critical

[6] My colleagues in this work were Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker.

engagement with racial and gender histories. I discuss race and gender categories in technical databases through a multidisciplinary lens, synthesizing theory from critical race studies, gender studies, infrastructural studies, and identity scholarship. I build on previous fairness scholarship to specify options for the field of computer vision, and machine learning more broadly, to evolve its approach to human identity and embrace new lines of research. My findings highlight opportunities for more human-centered methods that will improve both the representation of race and gender and the validity of annotations in image databases.

Methods

Researcher Positionality

Reflexivity in research practice establishes the researcher as a lens through which research is conducted; what Attia and Edge call an "on-going mutual shaping between researcher and research" (Attia & Edge, 2017). In other words, the research is shaped by the positionality—the social and political context—of the researcher. In alignment with the feminist practice of reflexively examining one's relationship to one's research, I want to highlight how the positionality of myself and my colleagues may have shaped this work.

Our approaches to examining both gender and race are informed by our collective experience—and many other experiences that make up our perspectives as researchers. The second author is Black, while the remaining three authors are white. Every author has a different gender. All of the authors are based in the United States. As such, our experiences are rooted in a Western-centric point of view. Each author comes from a multidisciplinary background, including HCI, computer science, psychology, gender studies, communication, and the arts. Our synthesized experiences with critical theory, race studies, and gender studies are shaped by education (both formal and informal) related to our U.S. nationalities. Our decision to examine computer vision practice with a critical lens stems from our scholarly upbringings. Our privilege as academics awarded us access to the resources to conduct this work, while our power as differentially marginalized individuals gave us the perspective to develop our research questions and interpret our data.

Defining the Language of Analysis

Colleagues and I discovered that different terms were being used interchangeably to describe race. This was true for gender as well. For that reason, I will outline how colleagues and I have decided to define our use of the terms “race” and “gender” in the context of this chapter. I return to practices of defining race and gender in more depth in Design Considerations.

Defining “Race”

Documentation of image databases often used both “ethnicity” and “race” to refer to a single concept: an annotation based on phenotypic attributes like skin tone and facial features perceived to be relevant to differing ancestral histories. Critical race theory and histories of racial stratification have outlined the use of “marked difference” for making visual determinations of racial categories (Hirschman, 2004; Lewis, 2003; Raengo, 2013), rather than traceable ancestral history. In other words, I suppose these annotations are linked to notions of visuality of making racial determinations, rather than determinations of ethnic origin. While critical race theory is primarily rooted in Western discourse, born of the desire to dismantle racism and white supremacy in the United States (Brown & Jackson, 2013), I believe it offers a lens for critiquing the structural issues of race in globally-sourced facial analysis databases. Thus, in this chapter, I chose to embrace the focus of “race,” as defined by critical race scholarship, in the analysis and discussion of image databases. With this in mind, I use the terms “race” and “racial.”

Defining “Gender”

Database documentation also used both “gender” and “sex” to refer to the singular concept of an externalized gendered appearance. In this chapter, I chose to use the term “gender” to refer to the analysis and discussion of gendered presentation in image databases, aligning ourselves both with trans scholars and concerns that the sex/gender dichotomy is often used to disavow trans identities (Chase, 1998; Fausto-Sterling, 2000; Serano, 2017).

Data Collection and Cleaning

I identified relevant facial analysis databases by examining which databases are being used in recent academic papers. I chose to identify databases used in academic papers because I could then assume they are viewed as useful databases in facial analysis research. To do this, I first created a corpus of 277 research papers that have studied facial analysis. To create this corpus, I started by identifying relevant papers in both the Association for Computing Machinery Digital Library (ACM DL) and Institute of Electrical and Electronics Engineers (IEEE). I chose to use papers published in the ACM and IEEE as they are both two of the largest associations of computing research, and thus contain a great deal of technical research on facial analysis. I scraped 18,661 manuscripts from the ACM DL for papers using the keyword "facial recognition" using Selenium WebDriver (OpenQA, 2016) on 12/12/2019.

I also downloaded 4,000 manuscripts from IEEE's Xplore library using the using search terms "facial recognition" and "facial classification" using IEEE's export functionality (12/12/2019)[7]. I aggregated the two datasets from the ACM and IEEE, removing all duplicates, resulting in 16,505 unique manuscripts.

Next, I filtered the manuscripts by "author keywords" for "facial recognition," "face recognition," and "face classification" to ensure the papers were directly relevant to facial analysis research. This resulted in 781 manuscripts. I decided to narrow the corpus to papers published within the last five years, between 2014 and 2019, to ensure both the manageability of the corpus and the modern relevance of the research. This left me with 277 manuscripts from which to manually identify the use of facial analysis databases.

Identifying Image Databases

Colleagues and I manually coded the remaining 277 manuscripts for referenced databases. Colleagues and I coded any mention of databases, including database creation, use for training, use for evaluation, and databases referenced in literature reviews. Through this process, we identified 160 different databases. We then manually coded each of these 160 databases for what types of media were included in each. This process revealed 15 different types of databases: Image; Video; 3D Model; Image and Video; Image and Sketch; Image, Audio, and Video; Video and 3D Model; Image (Eyes);

[7] The "export" feature can be found on the upper right-hand side of the search result page of https://ieeexplore.ieee.org/xplore/

Electrocardiogram (ECG); Audio; Image and Eyes; Image and Audio; Video and Audio; Image and 3D Model; and Image, Sketch, and 3D Model.

Image databases were the most common media type, referenced about five times as often than the next most popular media type, video. Accordingly, I chose to focus on image databases in the analysis of race and gender in databases.

As colleagues and I analyzed each database, we also eliminated those that did not contain human subjects (e.g., Common Objects in Context (COCO) (T. Y. Lin et al., 2014)). We also eliminated two databases for which we could not find explanatory documentation: 5NJJ-YN63 and PCSO_LS Mugshot. Finally, we chose to combine multiple versions of databases together for analysis, treating them as one database (e.g., Yale and Extended Yale B (Belhumeur et al., 1997); faces94, faces95, faces96, and grimace (*Face Recognition Data*, n.d.)) unless there were significant differences between versions (e.g., the introduction of new race and gender categories). This resulted in a final corpus of 92 image databases. For each entry in the corpus, we analyzed original sources such as research papers, websites, and additional supporting documentation.

Codebook for Analysis

I sought to understand *what purpose* race and gender were meant to serve for facial analysis systems; *which*, if any, sources race and gender were built on; and *what* processes were used to annotate race and gender information. In an initial review of the database sources, colleagues and I noticed the presence of this information was sporadic, at best. Thus, I developed a simple codebook that was iteratively updated in phases to capture four aspects of the databases:

1. Whether the database included information about race and/or gender
2. If race/gender was defined
3. Whether race/gender was either (1) only provided in the form a summary for the entire database or (2) race and/or gender was annotated at the level of individual images
4. How race/gender was annotated (when annotations were provided)

I developed this codebook to quantify trends across the databases and to focus my qualitative investigation. I coded the databases using the available documentation about them. Documentation included original research publications, auxiliary materials, websites, posted slide decks, and the databases themselves. The first author developed a codebook by first open-coding the types of identities present in the corpus of databases (Forman & Damschroder, 2007). Through regular discussion between the first author and the fourth author, the first author went back to develop tighter codes for the how race and gender showed up in the databases (what I refer to as “database types”) and how those identities were explained (what I refer to as “sources” and “annotation processes”). After finalizing this codebook, the first, second, and third authors then coded all database materials using the codebook. All coders regularly discussed their thought processes when coding, at the end of which the first author verified the coding of each codebook entry.

Codebook			
Concept	***Code***	***Description***	***Example***
Attributes Present	Race	Race, ethnicity, or skin color	Asian, white, dark skin
	Gender	Gender or sex	Man, female
Database Type	Implicit	The database includes race and/or gender information in the form of demographic distributions	56% female and 44% male
	Explicit	Every image in the database is annotated with race and/or gender information	Images of men are marked with an \`M’
Source	Present	An explanation for how race/gender was defined or derived	A formal citation defining the selected race classification
	Absent	No sources were used to explain the definition of race/gender	
Annotation Process	Present	An explanation for how explicit annotations were conducted	A description of how race categories were annotated by crowdworkers
	Absent	No explanation describing the annotation process	

Table 2. A table showing the codebook. Every database was marked as either Implicit, Explicit, or Neither. All databases were coded with either a 0 (absent) or 1 (present) for both gender and race.

I break down the codebook in **Table 2** in the following sections. For clarity, I provide a number of examples of coded documentation to highlight both the diversity of ways that information was represented across these databases, and how that information was coded.

Race and Gender

For each database, colleagues and I coded the presence and absence of race and gender. We coded “present” if the database contained race and/or gender. We coded “absent” if it did not.

Race: We coded race as any mention of racial categories, ethnicity, or skin color. The following example represents a snippet I would code as containing race:

> *“We also manually annotated the basic attributes (gender, age (5 ranges) and race) of all RAF faces. ... For racial distribution, there are 77% Caucasian, 8% African-American, and 15% Asian.” —Real-World Affective Faces Database (RAF-DB) (S. Li & Deng, 2019)*

Gender: Colleagues and I coded gender as any mention of gender or sex categories (e.g., gender, sex, men, male, etc.). We also coded proxies of gender, such as familial relationships like mother and daughter. For example, AR Database was coded as including gender based on the following text:

> *"Men's image names start with an `M' symbol and women's images start with an `W'."*
> *—AR Database (Martinez, 1998)*

Some databases discussed race and gender, but only in so far as to explicitly state race/gender were not accounted for. For example:

> *"Some questions were raised about the age, racial, and sexual distribution of the database. However, at this stage of the program, the key issue was algorithm performance on a database of a large number of individuals." —FERET (Jonathon Phillips et al., 2000)*

In these instances, I did not code the database as including race and/or gender.

Implicit and Explicit

I identified two ways human race and gender are included in databases. Some databases explicitly annotate every image with race and/or gender information. Others, however, only provide race and/or gender information in high-level descriptive statistics for the dataset as a whole. I coded these databases as "explicit" and "implicit" respectively.

Source for Definition of Race/Gender

After identifying which databases included race and gender, colleagues and I coded each for whether the database provided a source for how race and/or gender were defined. I accepted both formal citations and claimed reflexive expertise. I required the source justify the *definitions* of race and gender categories. For example:

> *"As prior work has pointed out, skin color alone is not a strong predictor of race, and other features such as facial proportions are important (Goldstein, 1979; Karras et al., 2018; Porcheron et al., 2017; Porter & Olson, 2001). Face morphology is also relevant for attributes such as age and gender (Ramanathan & Chellappa, 2006). We incorporated multiple facial coding schemes aimed at capturing facial morphology using craniofacial features (Farkas, 1994; Farkas et al., 2005; Ramanathan & Chellappa, 2006)." —IBM Diversity in Faces (DiF)* (Merler et al., 2019)[8]

Annotation Practices

Finally, colleagues and I coded each explicit database for explanations of how the authors conducted annotations. My criteria was some form of explanation to how race and/or gender was explicitly

[8] Note: Citations in quotes throughout this chapter have been altered to map to the correct references in the quoted paper.

annotated. For example, if the authors explained that they visually evaluated each image to make a determination about a race classification:

> *"Demographics for the 10k US Adult Faces Database were determined by an Amazon Mechanical Turk demographics study involving 12 workers per face. Amazon Mechanical Turk worker demographics were assembled from demographics surveys attached to the main tasks of Experiments 1 and 2." —10k US Adult Faces (Bainbridge et al., 2013)*

Based on an analysis of the coded databases, I present findings in three sections: (1) the purpose and intended use of databases (see Contextualizing Race and Gender by Understanding the Intended Purpose of Databases); (2) the implicit race and gender features found in databases (see Implicit Features); and (3) the explicit annotations of race and gender found in databases (see Explicit Features).

Public Availability of My Dataset

Given that the image databases are publicly available for academic use, I felt it was ethically responsible to publish the corpus of databases examined in this study for the benefit of other researchers, engineers, and the public. I have created an open access spreadsheet of the 92 databases (and associated versions) examined in this paper. This spreadsheet contains the codebook, including tabs for databases I classified as "implicit" and "explicit." I included the titles of original research papers, links to their Google Scholar entries, and the number of citations at the time of data collection. I also included quotes relevant to how race/gender are defined, sourced, and annotated, when available.

I encourage other researchers to use the dataset for additional research and to add new database entries. The dataset is available for download using the following DOI: 10.5281/zenodo.3735400

Contextualizing Race and Gender by Understanding the Intended Purpose of Databases

Image databases serve as an important resource for facial analysis research. Each time a new database is released, the authors are looking to fulfill some need within this community. I observed numerous justifications for the creation of new databases. For example, many databases were looking to continuously expand the number of individual faces available within a single database. I observed three major categories for which image databases were intended to be used for: (1) individual face recognition and verification; (2) image labeling and classification; and (3) for diverse training and evaluation. In these three categories, race and gender were included—or not included—for different reasons. Understanding these reasons contextualizes why I see race and gender manifest in both implicit and explicit ways.

Face Recognition and Verification

A great deal of database authors described the utility of their database for individual face recognition and verification tasks—that is, tasks meant to match a single individual to a database of images. Race and gender were often implicit in these recognition databases. It is likely that many database authors did not view explicitly annotated information as relevant to the task of face recognition. Matching a single face to a single identity is often viewed as an individual-level task, not requiring additional labels beyond a unique identifier (e.g., a subject's name, a number ID). However, some databases did include race and gender information—typically, along with other attributes like age and facial expressions. These often described the inclusion of such information for the sake of more expansive, more various data. Motivations for variety, however, are not necessarily the same as improving demographic diversity. Databases which sought to increase variety did not mention diversity as a motivation for identity information.

Image Labeling and Classification

Some databases were meant to aid with image labeling and classification—the assignment of labels to an image based on a database of images with annotations. Explicit race and gender annotations were common in such databases and were used to train a system to classify those annotated race and gender categories. While race and gender classification literature did not make up the majority of the original corpus, databases like Cohn-Kanade (CK) (Kanade et al., 2000) (which did not have explicit annotations) were still used for identity classification tasks (e.g., Anusha et al., 2017), suggesting a gap between the intended and actual use of this database.

Diverse Training and Evaluation

A number of image databases were created to improve the diversity of available faces for training and evaluation, and thus, ideally, mitigate potential representation biases within facial analysis models (e.g., DiF (Merler et al., 2019), PPB (Buolamwini & Gebru, 2018)). Such databases were looking to improve conditions for both face recognition and image labeling tasks, but their explicit contribution to the field is motivated by addressing known biases and underrepresentation, allowing for systems to recognize a wider variety of human faces and identity attributes. I saw that most databases utilized gender as a means to "balance" racial diversity—that is, to ensure there are a comparative number of women of a certain race to the number of men of a certain race. Explicit race and gender annotations may or may not be present in databases created for diversity—some chose to provide implicit demographic distribution information instead. Implicit demographic distributions still described the diversity of people represented in a database, while explicit annotations of that diversity could also allow for improved image classifications.

Implicit Features

Approximately 64% (*n*=59) of the 92 image databases colleagues and I coded did not contain explicit annotations. About 37% (*n*=34) contained no information about race or gender whatsoever (see **Table 3**). Approximately 27% (*n*=25) databases contained implicit information about race and/or gender in the form of demographic distributions. These implicit databases contained descriptive statistics about the race and/or gender of the people featured in the database but did not annotate that information for each image in the database. Only 4% (*n*=1) of databases with implicit data included source information for where demographic categories came from. The other 96% (*n*=24) databases did not contain any source information underlying demographic descriptions; I generally assumed that the database authors gathered this information directly from subjects or determined subject race/gender themselves.

Race and Gender Representation in Image Database Corpus				
Explicit Annotations			***Implicit Data***	***Neither***
35.9% (33)			27.2% (25)	36.9% (34)
Race	*Gender*	*Both*		
45.5% (15 of 33)	100% (33 of 33)	Both (45.5%) 15 of 33		

Table 3. The number of databases that contained (1) explicit annotations; (2) implicit demographic information; or (3) neither explicit or implicit race and gender information. Each count is out of 92 total image databases.

Types of Race and Gender Categories

Demographic descriptions manifested semantically in numerous ways. I observed both "*gender*" (e.g., CASIA-WEBFACE (Yi et al., 2014)) and "*sex*" (e.g., NUAA Photograph Imposter Database (Tan et al., 2010)), as well as both being used interchangeably to refer to the same concept (e.g., HRT Transgender Face Database (Mahalingam & Ricanek, 2013)). Similarly, I observed "*race*" (e.g., Sheffield (previously UMIST) (D. B. Graham & Allinson, 1998)) and "*ethnicity*" (e.g., VGGFace2 (Cao et al., 2018)), as well as both being used interchangeably (e.g., CMU Pose, Illumination, and Expression (CMU PIE) (Sim et al., 2002)), Compound Facial Expressions of Emotion (CFEE) (Du et al., 2014)). Underlying these concepts, I also observed numerous instances of categorical labels. For example, CFEE stated:

> *"A total of 230 human subjects (130 females; mean age 23; SD 6) were recruited from the university area, receiving a small monetary reward for participating. Most ethnicities and races were included, and Caucasian, Asian, African American, and Hispanic are represented in the database." —CFEE (Du et al., 2014)*

In this instance, "female" is used as a default gender, implying there must be another gender accounted for in the demographic distribution (most likely "male," given binary trends). *"Most ethnicities and races"* also similarly insinuates some races are not accounted for, but the authors believe their subject pool accounts for "most" of them. I also found some troubling descriptions, which relied on otherwise criticized or contentious terminology. One of these was in the documentation of the now unavailable Microsoft Celeb (MS-CELEB-1M) (Guo et al., 2016), which employed the categories "Caucasian," "Mongoloid," and "Negroid." The authors refer to these terms as encompassing "*all the* *major races in the world*" (Guo et al., 2016). It is possible the use of such terms is tied to historic scientific uses of the term to describe physiological differences between races; however, this was the only facial analysis database I saw use this term, indicating it is likely uncommon in computer vision literature. Such descriptions imply author determinations tied to cultural notions about race. Further, MS-CELEB-1M did not provide detailed distributions of these categories, stating:

> *"The diversity (gender, age, profession, race, nationality) of our celebrity list is guaranteed by the large scale of our dataset." —MS-CELEB-1M (Guo et al., 2016)*

Other databases also claimed to include different genders and races but did not describe what terms or categories they used. For example, the authors of NUAA wrote in their publication, "*Note that [the database] contains various appearance changes commonly encountered by a face recognition system (e.g., sex, illumination, with/without glasses)*" (Tan et al., 2010). However, they did not describe what "*sex*" looked like in the database.

Sources for Race and Gender Categories

The only database to contain source material for implicit demographic information was VGGFace2. The authors describe using Freebase knowledge graph to determine the *"attribute information such as ethnicity"* for the images of IMDB celebrities in their database (Cao et al., 2018).

Other databases, which I might expect to contain source material based on the outlined methodology, did not. For example, although the Facial Expression Recognition 2013 (FER-2013) Database described using Google search with keywords, they did not describe the process they undertook to define the keywords (Moreno & Sánchez, 2004). So although they stated that "*keywords were combined with words related to gender, age or ethnicity*" (Moreno & Sánchez, 2004), it is impossible to tell how the keywords were determined. Similarly, the HRT Database contained no source material on trans people in their definition of transgender as "*someone who under goes a gender transformation via hormone replacement therapy; that is, a male becomes a female by suppressing natural testosterone production and exogenously increasing estrogen*" (Mahalingam & Ricanek, 2013).

Choosing Implicit Demographics over Explicit Annotations

As I reported in Contextualizing Race and Gender by Understanding the Intended Purpose of Databases, databases meant for facial recognition and verification often do not need explicit race or gender annotations to function for their intended purpose, even when they sought to improve diversity. Even databases, which were built specifically in response to human characteristics, did not contain annotations. For example, the HRT Database, which was created for the purpose of identifying individuals across gender transition, was not annotated with gender information about individuals. The HRT Database was particularly unique in comparison to other databases in its treatment of "gender" and "sex." It was also the only database which discussed transgender identities:

> *"Gender transformation occurs by down selecting the natural sex hormone of a person in replacement for its opposite. This is known medically as hormone replacement therapy; however, more broadly this can be described as hormone alteration or medical alteration." —HRT Database (Mahalingam & Ricanek, 2013)*

In the HRT Database, transgender faces are problematized for recognition and verification tasks. Gender presentation is thus described not as an identity, but rather a challenge to facial analysis systems.

Often, individual subjects in the database were documented by race and gender to such a degree that descriptive statistics were possible, yet that documentation was never translated into explicit annotations. The NimStim Set of Facial Expressions Database, which also did not explicitly annotate race, but included demographic information in its documentation, stated:

> *"A number of features of the set are advantageous for researchers who study face expression processing. Perhaps the most important is the racial diversity of the actors. Studies often show that the race or ethnicity of a model impacts face processing both behaviorally and in terms of the underlying neurobiology of face processing. This modulation by race or ethnicity is not identified for all populations and may be driven by experience and bias." —NimStim (Tottenham et al., 2009)*

The above snippet outlines the reasoning for why NimStim authors intentionally included individuals of multiple racial categories into their database, despite not annotating those features: to improve accuracy and precision for facial recognition tasks. It is possible racial categories were not explicitly annotated, because the authors did not find that information relevant to recognition. FERET, one of the oldest databases, dated to 1993, was the only database I found to provide reasoning for the lack of annotations. On their website, they wrote:

> *"Some questions were raised about the age, racial, and sexual distribution of the database. However, at this stage of the program, the key issue was algorithm performance on a database of a large number of individuals." —FERET ("Face Recognition Technology (FERET)," 2017)*

I also witnessed an interesting example of identity-specific licensing agreements in the Iranian Face Database (IFDB), which prohibited the use of women's images in publication. They stated:

> *"Some female's images are also provided in this database. These images will never appear in any document of any form." —IFDB (Nik et al., 2007)*

IFDB was interesting in this regard, as they displayed some concern over the misuse of women's images by third-party researchers and commercial interests. However, their approach to choosing implicit demographic labels of gender also leaves the gender of each image up to interpretation from those same third parties. I return to the concept of identity-specific licensing in the Design Considerations.

Some databases would include explicit annotations for one feature but not another. For CMU PIE, which included "*sex*" but did not include race, they also stated: "*At the time of writing, we have not decided whether or not to include the "race" or "ethnicity" of the subjects in the personal attributes*" (Sim et al., 2002). The authors did not detail why they had not decided to include race in their annotations. I rarely found explanations about why demographics were included, or not included, annotated, or not annotated.

I did observe instances of third-party researchers annotating databases which originally did not contain explicitly annotated features. For example, Afifi et al. annotated gender for Labelled Faces in the Wild (LFW) (Afifi & Abdelhamed, 2019). While I did not include third-party annotations of databases in the official analysis, I return to them in the Design Considerations.

Explicit Features

As shown in **Table 3**, approximately 36% (*n*=33) of the 92 databases colleagues and I analyzed included explicit annotations—either of race or of gender, or of both. About 45% (*n*=15) of the 33 databases with explicit features included explicit race annotations, while 100% of the 33 databases with explicit annotations included explicit gender annotations.

Of the databases annotated explicitly with racial features, none of the databases with race annotations contained *only* sources (with no annotation information) for how racial determinations were made; 20% (*n*=3 of all databases with race annotations) contained explanations for how annotation practices were conducted, but no sources; and 20% (*n*=3) contained both sources and annotation documentation (see **Table 4**).

Similarly, no databases contained sources without descriptions of the annotation process for gender (of all databases with gender annotations); 6% (*n*=2) contained descriptions of the annotation process by which images were labeled with gender; 6% (*n*=2) contained both a source and a description of the annotation process: IBM DiF and PPB. For both race and gender, the databases which provided sources and/or descriptions of the annotation processes did so with varying levels of rigor. I further illustrate the observed source material and annotation documentation in the database corpus in the following sections.

Sources and Annotation Descriptions in Explicit Databases		
	Race	***Gender***
Source	0%	0%
Annotation Description	20% (3 of 15)	6% (2 of 33)
Both Source & Annotation Description	20% (3 of 15)	6% (2 of 33)
Total # Databases with Source/Annotation Description	40% (6 of 15)	12% (4 of 33)

Table 4. The above table shows a count of sources and annotation descriptions in explicitly annotated databases. Databases marked as containing sources do not contain annotation information, and vice versa. Only databases marked as including both source and annotation information contain both. Only 2 databases—DiF and PPB—contained both sources and annotation explanations for both race and gender.

Explicit Race Annotations

Types of Race Categories

Like I found in implicit demographic descriptions, race varied widely across the image databases I analyzed. Once more the concept of race was visually described using numerous concepts, like "*race*" (e.g., MORPH (I & II) (Ricanek & Tesafaye, 2006), Sheffield) and "*ethnicity*" (e.g., Annotated Facial Landmarks in the Wild (AFLW) (Köstinger et al., 2011), FER-2013) and "*skin type*" (e.g., PPB) or "*skin color*" (e.g., IBM DiF). Also like with implicit results, I once more observed numerous instances of categorical labels. For example, Radboud Faces Database (RAFD) contained only two racial categories: "*Caucasian*" and "*Moroccan*" (Langner et al., 2010). Such categories seem more explicitly tied to origin, than visual characteristics of race; yet their annotations imply visually determinable information. PUBFIG employed four categories: "*White*," "*Asian*," "*Black*," and "*Indian*" (N. Kumar et al., 2009). PUBFIG's categories seem to be determined by visual racial categories, like white, as well as notions of origin, like Indian. I also found that "*other*" was sometimes utilized as a category (*n*=3; MORPH-II, KINFACEW, 10K US Adult Faces).

Sources and Annotation Processes for Race Categories

The authors of the 10K US Adult Faces Database defined their racial categories as "*White*," "*Black*," "*Hispanic*," "*East Asian*," "*South Asian*," "*Middle Eastern*," and "*other*." They explain that they sourced these categories from “common” Amazon Mechanical Turk demographics found in experiments they conducted with Mechanical Turk workers. They compare the demographics of the workers and their database to the United States Census:

> *“Demographics for the 10k US Adult Faces Database were determined by an Amazon Mechanical Turk demographics study involving 12 workers per face. Amazon Mechanical Turk worker demographics were assembled from demographics surveys attached to the main tasks of Experiments 1 and 2 ... The 1990 U.S. Census asks about Hispanic origin as a separate question from race, so there is likely overlap with other races.” —10K US Adult Faces (Bainbridge et al., 2013)*

The U.S. Census categories and the selected 10K US Adult Faces Categories do not perfectly align in the paper, making it difficult for readers to discern what decisions were made in collapsing “*East Asian*,” “*South Asian*,” and “*Middle Eastern*” into simpler categories. Presumably, these categories were used to pre-define their annotation guidelines, which were also conducted using Mechanical Turk. The authors describe the process for which workers were asked to annotate relevant facial attributes:

> *“To collect the facial attributes, we conducted a separate AMT survey similar to (N. Kumar et al., 2009), where each of the 2222 face photographs was annotated by twelve different workers on 19 demographic and facial attributes of relevance for face memorability and face modification. We collected a variety of attributes including demographics such as gender, race and age, physical attributes such as attractiveness, facial hair and make up, and social attributes such as emotional magnitude and friendliness.” —10K US Adult Faces (Khosla et al., 2013)*

The PPB database was the first database to explicitly annotate skin tone as a proxy for race, annotating images with two different skin tones: darker and lighter. They explain their decision to use skin tone due to the instability of racial categories:

> *“Since race and ethnic labels are unstable, we decided to use skin type as a more visually precise label to measure dataset diversity. Skin type is one phenotypic attribute that can be used to more objectively characterize datasets along with eye and nose shapes. Furthermore, skin type was chosen as a phenotypic factor of interest because default camera settings are calibrated to expose lighter-skinned individuals (Roth, 2009)... By labeling faces with skin type, we can increase our understanding of performance on this important phenotypic attribute.” —PPB (Buolamwini & Gebru, 2018)*

PPB uses the Fitzpatrick skin type scale as both a source of categorizing and a guide for annotation (Buolamwini & Gebru, 2018). In their annotation practice, they apply the Fitzpatrick scale to each image in their database. The authors write:

> *"For the new parliamentarian benchmark, 3 annotators including the authors provided gender and Fitzpatrick labels. A board-certified surgical dermatologist provided the definitive labels for the Fitzpatrick skin type." —PPB (Buolamwini & Gebru, 2018)*

IBM DiF was the only database which attempted to conceptualize race using multiple phenotypic categories. They employed both skin color and numerous facial coding schemes to determine racial annotations. They derive diverse racial categories from multiple sources: skin tone from Chardon et al. (Chardon et al., 1991); craniofacial distance from Farkas et al. (Farkas et al., 2005); and craniofacial ratios from (Ling et al., 2010). Like PPB, these different sources also drove the annotation process:

> *"As prior work has pointed out, skin color alone is not a strong predictor of race, and other features such as facial proportions are important (Fu et al., 2014; Goldstein, 1979; Porcheron et al., 2017; Porter & Olson, 2001). Face morphology is also relevant for attributes such as age and gender (Ramanathan & Chellappa, 2006). We incorporated multiple facial coding schemes aimed at capturing facial morphology using craniofacial features (Farkas, 1994; Farkas et al., 2005; Ramanathan & Chellappa, 2006)." —IBM DiF* (Merler et al., 2019)

Not all databases which provided annotation information also provided source information. The photos making up MORPH are taken from public records, but they do not describe where they source their concepts of race from. The authors of the MORPH database described their race annotation process in auxiliary materials. The authors, who conducted the annotation, evaluated images of those determined to have "inconsistent race" on a case-by-case basis:

> *"Each of the 33 people with inconsistent race was evaluated on a case by case basis. A final decision was made according to one of the following criteria:*
>
> - *Simple Majority: All images for a given person were assigned the race that appeared at least 50% of the time.*
> - *Visual Estimation: Each person's images were inspected one at a time. We decided the race only if there was a wide consensus among our team members.*
> - *Other: For some people (e.g. [sic] those of mixed race) it was difficult to guess their race from the photos, and there was substantial variation in the original dataset. We set the race of all images to Other." —MORPH (Bingham & Yip, 2017)*

From this description, I can only assume the MORPH authors made subjective decisions about each "*inconsistent race*" image based on their own perceptions about what race should look like. While this approach is presumably true for many databases with annotated race characteristics, it is made apparent through MORPH's description of their consensus-driven approach.

Explicit Gender Annotations

Types of Gender Categories

In all gender annotations, gender was only categorized in two ways: "*gender*" (e.g., LFW, CMU PIE) and "*sex*" (e.g., NUAA). Similarly, categorical variables took two variations: "male/female" or "man/woman." The majority of the databases used the same schema as CAS-PEAL: "*male*" and "*female*" (Gao et al., 2008). AR Database annotated each image with an "*M"* to indicate men and a "*W*" to indicate woman (Martinez, 1998). I also determined proxies for gender to be explicit gender labels. For example, FAMILY101 used the labels "*son*," "*daughter*," "*father*," "*mother*," "*wife*," and "*husband*" (Fang et al., 2013). These annotations could also easily be used for gender classification tasks; thus, I determined them to be gender annotations.

A particularly interesting example was PUBFIG, which had two gendered annotations: "*male*" and "*attractive woman*," of which there was no associated "female" (N. Kumar et al., 2009). The absence of annotations for "female" or "attractive man," however, highlights the culturally situated values around gender that can emerge within an annotation schema (see Chapter 4).

Given previous literature documenting that automated gender classification only exists as a binary (See Chapter 4 and (Keyes, 2018)), I was not surprised to find almost all databases used only two categories that equated to "male" and "female." The only outlier in this regard was RAF-DB, which contained *"5% remains unsure"* in their description of gender annotations (S. Li & Deng, 2019). While "*unsure*" still insinuates the individuals in the database should fit into the gender binary, it showcases something beyond "male" and "female." Overall, the following section focuses primarily on how gender categories were determined and how the process of labeling them was discussed.

Sources and Annotation Processes for Gender Categories

Two databases described the annotation process they used to apply gender labels. PUBFIG, which used the labels "*man*" and "*attractive woman*," used Amazon Mechanical Turk to crowdsource gender labels:

> *“Each job was submitted to 3 different workers and only labels where all 3 people agreed were used. In this way, we collected over 125,000 confirmed labels over the course of a month…” —PUBFIG (N. Kumar et al., 2009)*

FAMILY101, which uses gendered labels to describe family members (e.g., “*son*,” “*daughter*”), used a similar process. The authors selected pre-determined celebrity families, then asked the crowdworkers to find specific images of those families (Fang et al., 2013). Other databases used web-based information to determine gender labels. For example, Indian Movie Face Database used IMBD to annotate for the gender of different actors and actresses (Setty et al., 2013). However, it is notable that none of the above databases provided sources for how they defined gender for the purposes of annotation.

Only two databases included source information justifying their gender annotation categories: PPB and IBM DiF. These databases also happened to be the only two detailing their gender annotation process. PPB, employed binary “*male*” and “*female*” labels, explained their choice (Buolamwini & Gebru, 2018). The authors wrote: “*In this work we use the sex labels of `male’ and `female’ to define gender classes since the evaluated benchmarks and classification systems use these binary labels*” (Buolamwini & Gebru, 2018). PPB used existing database literature to source their binary gender category selection. The authors describe using a combination of “*the name of the parliamentarian, gendered title, prefixes such as Mr [sic] or Ms [sic], and the appearance of the photo*” to label gender for each image (Buolamwini & Gebru, 2018).

IBM DiF happened to be the only database which incorporated gender into a “*facial coding schema*” based on sourced craniological features. They reference Rothe et al. (Rothe et al., 2018) as driving their coding scheme for gender. They also describe multiple approaches for annotating gender. In one step, they used automated gender estimation as described by Rothe et al. (Rothe et al., 2018). In another step, they employed human labelers using the crowdsourcing platform Figure Eight. All of the coding schemes, including both automated gender classification and human labeling, are available for each image.

Both PPB and DiF explicitly describe the binary gender annotations in their respective databases as a limitation. The authors of IBM DiF wrote:

> *“Note also that the gender categorization in Table 3, as in much of the prior work, uses a binary system for gender classification that corresponds to biological sex – male and female. However, different interpretations of gender in practice can include biological gender, psychological gender and social gender roles. As with race and ethnicity, over-simplification of gender by imposing an incomplete system of categorization can result in face recognition technologies that do not work fairly for all of us.” —IBM DiF* (Merler et al., 2019)

DiF tried to mitigate the binary effect by using an average score of 0 to 1 for each image to *“to predict a continuous value score for gender between 0 and 1, and not just report a binary output”* (Merler et al., 2019). However, they also used a “*male*” versus “*female*” scale for subjective human-labeled annotations.

Discussion

Moments of Identification: A Machine Learning Approach to Human Identity

Human identity characteristics have become increasingly operationalized and scrutinized within machine learning literature. On one hand, there are attempts to classify attributes like ethnicity (Gutta et al., 1998; Lu & Jain, 2004; Mansoor Roomi et al., 2011) and gender (Ramey & Salichs, 2014; Rodríguez et al., 2017; Santarcangelo et al., 2015). On the other, there are growing concerns about how identity is being represented (Benthall & Haynes, 2019; Keyes, 2018), whether it is fair (Grother et al., 2018; Mehrabi et al., 2019; Raji & Buolamwini, 2019), and what the outcomes are when it is not (Bacchini & Lorusso, 2019; Noble, 2018; Suresh & Guttag, 2019). Underlying these concerns are the massive amounts of data that machine learning models require for their training and evaluation. For facial analysis models, in particular, this data must contain visual information about human faces. Thus, I see database authors primarily relying on *moments of identification* (Hall, 2012), using the visible, external appearances of faces to make determinations about race and gender. Yet, the nature of identity is shaped by both cultural and historical factors; it is sociohistorical. This nature reveals challenges to assumptions that racial and gender categories throughout these databases are objective in the first place. Given the lack of engagement with sociohistorical theory or deeper notions of invisible, internal race and gender identities, I thus observed race and gender categories—that are socially and historically complex—*portrayed as obvious, static, and apolitical*.

Underlying the three purposes of image databases—recognition, classification, and image diversity—is the utility, reliability, and accuracy of the ground truth data provided. I observed very few instances where database authors documented how they determined information about race and gender in facial images. However, I observed extensive documentation by database authors on the mechanics of lighting, image quality, and camera angles. Such mechanics are generally rooted in agreed upon standards. For example, no one is disputing that angles exist on a 360-degree plane. The lack of similar engagement with ground truth race and gender data undermined the very purpose of image databases to be usable, reliable, and accurate.

Yet, considering the sociohistorical nature of these categories, standardization and benchmarking remains challenging, if not impossible. Race on the one hand demonstrates the challenges when there is a lack of agreed upon standards. However, in contrast, gender shows how even when there is an agreed upon standard, it can be problematic for sociohistorical identity attributes. Instead of aiming for objective standards for classifying race and gender, clear documentation of the tradeoffs and decisions would be a more reasonable and effective approach.

When I examined race, I quickly saw the notion of objectivity degrade. I observed inconsistent practices for identifying race categories, which were otherwise portrayed as objective or obvious. When it came to race, I saw disparate schemas for classifying race and ethnicity. I observed notions of visible race markers (e.g., "*Black*") and notions of origin (e.g., "*Moroccan*"), as seen in RAFD (Langner et al., 2010)). Such inconsistency actually revealed the inherently subjective nature of identifying race from images, and therefore the apparent lack of a benchmark standard as seen in measuring camera angles.

Gender was somewhat opposite compared to race. There is a longstanding practice in database construction to adopt a physiological binary perspective of "male" and "female." I did observe a small number of recent and emerging efforts to address documentation of race and gender in image databases—in particular, with PPB (Buolamwini & Gebru, 2018) and IBM DiF (Merler et al., 2019). Both databases justified their decisions about the race and gender categories they chose and provided explanations about how they went about annotating them. They also weighed the benefits and tradeoffs of various approaches, as seen in their discussions on the limitations of binary gender categories. However, their different schemas highlight the lack of a shared standard for (in this case) race. The complexity they detail in arriving at their classification, however, further highlights the situatedness of sociohistorical attributes, and the need for better guidelines around how to produce such datasets, and what contextual constraints (e.g., cultural context, time, etc.) should be considered when using them. Yet, authors of newer databases (i.e., DiF and PPB) have begun to question how gender is typically viewed in machine learning, even if they have not figured out ways to annotate images beyond it. HCI researchers have also criticized its erasure of trans experiences, as seen in my prior work (Hamidi et al., 2018), in Chapter 4, and in work by other scholars (e.g., Keyes, 2018). These criticisms highlight a gap between facial analysis databases and the social realities they are attempting to capture. It may not be possible, or desirable, to develop a shared standard.

When the nature of identity attributes is approached as "common sense," this may also lead to their portrayal as something neutral, objective, obvious, or even irrelevant. As the classifications of race and gender in these databases become folded into actual facial analysis systems, potentials for assessing the impact of classification decisions become increasingly opaque. As other scholars have noted, continuing to embed such limited perspectives into technical systems has the potential to reinforce harmful historical practices of exclusion (Bacchini & Lorusso, 2019; Keyes, 2018) and even fortify pseudoscientific practices of asserting invisible internal characteristics from visible identifications, like physiognomy (Agüera y Arcas et al., 2017; Valla et al., 2011).

While the sociohistorical nature of race and gender make straightforward and universal database construction unreasonable, these limitations highlight the importance of entirely new lines of scholarship in computer vision addressing how race and gender are encoded into our systems. Specifically, I suggest that facial analysis researchers embrace two practices for situating their databases: (1) embracing positionality; and (2) adopting a sociohistorical perspective when making decisions around classifications of race and gender. In the rest of this section, I provide more details specific to race and gender and then concrete examples as to how researchers may adopt these practices in the Design Considerations.

Positionality: Author and Annotator Classifications of Race and Gender

When annotating image databases, identification was often conducted through an analysis of the visible characteristics of a subject—usually by an author or a crowdworker. Of the databases which did not provide this information, I can assume that race and gender categories are also assigned primarily by visually assessing each image or subject in a database. Through the work of Stuart Hall, I see these as moments of *identification*, the situated construction of identity categories, to include or exclude specific groups of people (Hall, 2012). Importantly, identifications of race and gender is conducted through the lens of the person doing the identifying and is situated with "*specific modalities of power*" (Hall, 2012). Identification is colored by one's experiences, interpretations, and perspectives. Without a doubt, the differing perspectives of database authors are why I observed such a wide variety of taxonomies in the data—"gender" versus "sex," and "ethnicity" versus "race."

These categories are "*constructed within, not outside, discourse... in specific historical and institutional sites within specific discursive formations and practices*" (Hall, 2012). We see in current database practices the collapsing of human identity—which consists of both visible external and invisible internal aspects—into solely the visible. For example, visible gender expression is being used as a proxy for internal gender identity. Beyond the general lack of source material and annotation descriptions in databases, I also observed a lack of acknowledgment of external identification as a subjective process, informed by one's own position and perspective in shaping race and gender categories. Without understanding the position of the author or annotator, the collapse of subject identity is made to appear neutral or objective. Statements like "*most ethnicities and races were included*" (CFEE (Du et al., 2014)) and "*the diversity ... is guaranteed by the large scale of our dataset*" (MS-CELEB-1M (Guo et al., 2016)) are written into database documentation as if the comprehensive diversity of race is objectively possible.

Due to vague documentation, it was also not apparent what visible markers were most salient when authors or annotators were making race and gender determinations. The physical embodiment of identity manifests in numerous ways, both physiological—like skin color, face shape, body type—and expression—like clothing, makeup, and mannerisms. Such visible embodiment is crucial to identity, but it is also not always as simple and static as portrayed in the neutral and unexplained language of most database documentation. It is intertwined with social, cultural, and historical aspects of gender and race—

the otherwise invisible aspects of identity. Prior scholars have critiqued the collapse of both the visible and the invisible. For example, I assessed the binary output of commercial gender classification systems, noting that, in facial analysis software, "presentation equals gender" (see Chapter 4). I saw the "objective" portrayal of visible identifications fall apart in cases like MORPH, where the authors described the difficulty "*to guess [some subjects'] race from the photos*" (Bingham & Yip, 2017). As facial technologies are often deployed beyond the locale they were initially developed in—often even globally—they enact forms of identification that do not necessarily align with the cultural reality of race and gender of other cultures and histories. Further, they do provide any methods for ensuring that they do align with localized cultures and histories.

The practice of embracing positionality and acknowledging one's perspective as a researcher has been a longstanding practice in feminist epistemologies (e.g., Attia & Edge, 2017; England, 1994; G. Rose, 1997). Positioning race and gender classifications within a discipline, a theory, a history, or oneself would increase the transparency and utility of the database itself. The practice is meant to inform others of the context the research was conducted in and instill trust in the researcher's perspective. In detailing the subjective perspective of identifying race and gender features in images, database authors would make their decisions more transparent to potential users of their databases. This would also allow third parties to better understand how the database might fit their specific use case. I discuss one approach for doing this in Embrace .

Sociohistorical Sourcing: Tying Historical Approaches of Race and Gender Identification to Database Documentation

Identification in technical systems has spanned centuries and geographies. Race and gender have largely been defined by *visible markers of difference*. These differences have then been encoded into numerous technical systems of identification. Given that the premise of facial analysis databases is to enact moments of identification based on the visible features of people, it is imperative that we understand how race and gender have been operationalized in technical infrastructures that predate machine learning. After all, categories like "*Negroid*," "*Black*," and "*African American*" hail from historically evolving notions of both the physicality of race and countries of origin (Hirschman, 2004; Takezawa, 2012); categories like “*male*” and “*female*” have largely been derived from historically entrenched notions of biological sex that erase trans realities while producing normative cisgender ones (Butler, 1988). This is particularly important given the necessity of incorporating race and gender into system design for mitigating bias and ensuring equal representation of marginalized groups (Corbett-Davies & Goel, 2018; Lambrecht & Tucker, 2016; Obermeyer et al., 2019).

In order to understand the potential misuse of classifications in facial analysis technologies, I review how race and gender identifications have been used to enact discriminatory political actions. For race, I discuss the evolution of the Census in the United States (M. M. Smith, 2006) and the sordid practice of physiognomy (Krüger, 2010). For gender, I observe an ongoing battle with gender classification schemas in the trans rights movement (Kunzel, 2014; Namaste, 2000). I review these examples, connecting them with what I observed in the identification practices in image databases. My goal in drawing these connections is to ensure facial analysis research avoids replicating problematic practices and mitigates the creation of technical systems that repeat unjust histories.

Histories of Race Identification Embedded into Databases

Some databases in the findings utilized U.S. Census categories for their definitions of race; however, none engaged with how Census categories have been used to count and erase certain people from political participation. This history is crucial to the evolution of the Census categories in the first place. The initial classification of race only had two distinctions—free or slave—which then evolved into the first census groups: European, African, and Native American (Prewitt, 2005). Even now, the Census is constantly evolving alongside shifting political agendas and social change. Who gets recognized on the Census determines who is literally counted. Alongside Census-informed categories, I also observed classifications for “*Other*,” which otherwise erase the racial identities of non-classifiable subjects.

Embedding terms like “*Negroid*” and “*Mongoloid*” into database documentation, which have associations with histories of scientific racism based on such visible differences (Mukhopadhyay, 2018; Takezawa, 2012), insinuates a lack of understanding of categorical oppression. I observed some databases attempting to move away from politicized racial categories, as found in the census, to more static notions of racial affinity: visible skin tone and facial morphology. This form of identification potentially allows for more accurate categorizations of people than subjective racial categories. However, there are tensions between more accurate measurements and the historical practice of physiognomy: the procedure of asserting one's internal character from visible racial and ethnic characteristics. For example, both the segregation in Apartheid South Africa (Bowker & Star, 2000) and the extermination of Tutsi people in the Rwandan genocide were based on racial difference through the selected codification of cranial features (Krüger, 2010). Such tensions—between attempts to more objectively increase image diversity and histories of scientific racism—are problematic to actually address. Explanation for choosing to rely on visible difference would help make author intent visible to third parties.

Beyond critically rethinking when and how to incorporate race categories into databases, I further encourage critical questions as to how those categories may be used in working facial analysis technologies. As I discussed in the related work, annotations of race and gender can aid the classification of minorities, making it easier to track them using facial analysis.

While I encourage thoughtful inclusion of racial diversity into databases, I caution database authors to consider potential physiognomic ties and oppressive outcomes that might result from operationalizing race; I discuss opportunities for mitigating such uses in the Design Considerations.

Histories of Gender Identification Embedded into Databases

As I discuss in Chapter 4, the identification of gender in image databases can reify notions of gender as binary, visible, and obvious. I also observed notions of gender tied to archaic notions of women's appearances. Categories like "*attractive woman*" (PUBFIG) places additional weight on the visibility of images of women subjects, emphasizing beauty standards that are more often applied to women than men (Ponterotto, 2016). Such perspectives sit in direct opposition of trans activists, as well as feminist scholarship that seeks to imagine gender as an internal, social, and cultural phenomenon (Butler, 1988). Trans rights movements have been built on fighting restrictive gender markers on government documentation. Mismatches between identity documents can restrict movement between countries (Currah & Mulqueen, 2011; James et al., 2016); result in differential healthcare access, particularly due to gendered insurance restrictions (L. Khan, 2011); and may result in concerning and risky encounters with police and other officials (Muth, 2018; National Center for Transgender Equality, 2015). Yet legal documentation has changed in response to trans movements. For example, non-binary options are becoming increasingly available in certain U.S. states (e.g., Colorado driver's licenses (Schmelzer, 2018)).

Much like with race, such shifts also showcase the fluid and political nature of gender identification. Moreover, they show the mismatch between the standards being employed for gender in databases with the changing standards of gender in other technical systems. The choices authors embrace when making gender identifications necessitate a thoughtful questioning of the role a database will play when incorporated into working facial analysis systems. Like race, gender should be handled with care when operationalized into databases; in particular, I encourage more nuanced ways of representing gender that neither exclude trans identities nor put them at risk.

Given the replication of historical race and gender categories in databases, and their generalizability for uses in unforeseeable large-scale systems, researchers must critically imagine potential misuses. Such imagination is necessary; research has shown that the omission of explicit consideration of race and gender results in disparities (e.g., Corbett-Davies & Goel, 2018; Lambrecht & Tucker, 2016; Obermeyer et al., 2019), and thus, race is still a necessary construct to consider when building databases for machine learning systems. Examining sociohistorical identifications gives researchers the tools to do just that. Given the historical mistreatment of race and gender reviewed in this section, I encourage computer vision researchers to begin incorporating historical perspectives into their documentation. The status quo of database construction and annotation disservices computer vision research. I seek to envision new, human-centered lines of scholarship aimed at capturing the invisible, internal aspects of identity.

Design Considerations

I found that there is a general lack of documentation for how race and gender categories are designed and annotated in training and evaluation databases for facial analysis technologies. This finding exposes numerous opportunities. Specifically, there are abundant opportunities to address the two main issues in the previous sections: to provide clear documentation that addresses both (1) the positionality of the authors and annotations and (2) the sociohistorical context of race and gender categories. I then provide three additional design considerations: (1) revisit, revise, or retract existing image databases; (2) creatively incorporate "invisible" aspects of identity; and (3) explicitly define identity-specific limitations of use. In this section, I detail these five interventions towards promoting growth in the field of computer vision, and machine learning broadly.

Provide clear and transparent documentation of race and gender that includes positionality and sociohistorical context

I urge database authors to provide more rigorous documentation about their database creation processes. First, they should expound on the decisions they make to include and exclude certain races and genders in their databases. Second, they should describe how they are defining race and gender. Third, they should write rich descriptions of how they annotate race and gender information. For example, whether they annotated images based on participant self-identification or based on appearance. They should also provide any guidelines they follow for conducting annotations. For example, what features made an annotator label an image with "woman." Providing documentation on these decision-making processes would make databases more transparent, and thus more usable to third parties.

Embrace positionality

As race and gender are sociohistorically situated, so too are the perceptions authors and annotators introduce into databases. Including the perspectives, training, and identities authors and annotators bring to image databases would increase the level of transparency currently absent in decision-making processes. Knowing the demographic distribution of authors and annotators is just as useful as knowing the demographic distribution of subjects in the database. There is detailed precedent in other fields for including positionality statements in research. Sociology, anthropology, and increasingly HCI introduce positionality (or reflexivity) into research to ground the choices researchers make when defining research questions, methods, and findings (e.g., Attia & Edge, 2017; Dombrowski et al., 2016; Rode, 2011). Database authors can include small statements on positionality in their work. Beyond positionality statements, database authors might also consider weaving in smaller nods of positionality into descriptions of the methods used to define and annotate classifications. To accomplish this, database authors might explain their relevant expertise to the task of identity classification. They might also explicitly ask annotators to describe their own race and gender identities, for the sake of understanding how that might shape their annotations. Ogbonnaya-Ogburu, Smith, To, and Toyama also suggest that in addition to these endeavors, researchers ought to be “other-conscious” (Ogbonnaya-ogburu et al., 2020); in other words, to consider how their work will be viewed by people in other groups than their own, particularly racial minority groups.

It is important to note I am not advocating for the compulsory disclosure of sensitive experiences or marginalized identities. I acknowledge the increased burden of researchers from marginalized identities in self-disclosure, which may risk their personal and professional lives (LaSala et al., 2008). There is also the valid concern that work conducted by marginalized individuals on topics of identity will be viewed as less scientific and less valid (Honeychurch, 1996; Serrant-Green, 2002). Rather, I am advocating for increased context setting around the decisions researchers make when constructing and documenting databases, and a deeper attention to documenting the identities of annotators as seen in more research practice (e.g., the reporting of participant demographics).

Incorporate sociohistorical context

Documentation is crucial to understanding the sociohistorical context in which databases are created and annotated. Database authors should explicitly detail the decisions they have made in defining race and gender, as well as the categories they have chosen to represent. As I have demonstrated through my discussion in Sociohistorical Sourcing: Tying Historical Approaches of Race and Gender Identification to Database Documentation, not only do cultural definitions of race and gender change over time, but they are linked to sociohistorical modalities of power. How race and gender concepts are distinguished is contextual to how such categories are being used. Race is often seen as having shared physical traits, whereas ethnicity is seen as shared cultural traits; furthermore, nationality is sometimes is conflated with both (Morning, 2008). Such classificatory distinctions can also disguise heterogeneity within both “*race*” and “*ethnicity*” (Gordon, 2007). Similarly, “*gender*” and “*sex*” are often distinguished as two separate concepts. Furthermore, sexologists and activists have criticized the distinction between “sex” and “gender” for erasing intersex bodies (Chase, 1998; Fausto-Sterling, 2000). Many trans and gender scholars have since rebutted the gender and sex distinction altogether (e.g., Butler, 1988; Cealey Harrison & Hood-Williams, 2002).

When making distinctions, database authors should review theories of race and gender as part of their literature review, and discuss the limitations and tradeoffs of their decisions in their documentation.

Revisit, revise, or retract existing image databases

During analysis, I came across databases which third-party researchers had subsequently annotated. For example, although the original LFW database did not have explicit gender annotations, Affi et al. have subsequently annotated this database with gender annotations (Afifi & Abdelhamed, 2019) on their official website. This is a creative way to extend existing databases, which may not contain race and gender information, and improve them. Given the difficulty in accessing the original subjects for most existing databases, it’s likely third parties would still need to adopt methods for external identification of race and gender. In this case, Affi et al. adopted a binary approach to gender in their adaptation of LFW. However, I encourage third parties to embrace opportunities to introduce alternative, more inclusive annotations. Much like I recommended in Chapter 4, third-party annotators might instead choose to annotate varieties of gender expression (e.g., feminine, masculine; long hair, short hair; makeup, no makeup), rather than perceived gender identity. They might also embed feminist theories into empirical methods. For example, by introducing Standpoint Theory (Rolin, 2009) into crowdworking annotations, purposefully recruiting crowdworkers of diverse races and genders to label images, and situating those identities within annotation reports.

I also encourage the research community to revisit and question the utility and inclusivity of previously published databases. For example, I saw that both the MS-Celeb-1M and the HRT Transgender Faces database had been retracted from their websites. They are no longer available for download. It is likely this is due to critical commentary from both researchers (e.g., Keyes, 2018) and the media (e.g., Murgia, 2019; J. Vincent, 2017). Reassessing the validity of databases should be encouraged by the research community, through encouraging both original authors and third parties to publish work evaluating existing databases. The retraction of databases should not be seen as a failure, but instead a contribution to improving computer vision.

Creatively engage “invisible” aspects of identity

I believe that a true representation of race or gender cannot be ascertained without explicitly coupling both the visible—physical embodiment—and the invisible---social and historical realities that shape the internal sense of self. Yet, it is extremely difficult to incorporate the invisible aspects of identity into such a visual medium as computer vision. Those few databases that attempted to fold in the social and internal aspects of gender (PPB and DiF) did so only insofar as to say their databases were limited by being unable to capture this complexity.

Moving forward, I encourage database authors to re-imagine how identity in image databases can also embrace the invisible characteristics of racial and gender identities. Already, there have been attempts to incorporate self-identification into image databases. For example, as described in Chapter 4, I built an image database using self-annotated gender. This work attempts to encapsulate the invisible in visible images using self-identification of the subjects. Database authors might also consider methodologies for collecting self-identification directly from subjects—in studio or by survey. We might also consider looking beyond prevailing one-sided moments of identification—where facial analysis systems classify or verify an individual without consent—to interactive systems which allow individuals agency over identification. Such techniques could greatly improve the depth and accuracy of the representations of the identity of the subject.

I acknowledge such approaches are also faced with a large number of limitations. It may be technically infeasible to gather enough images this way; it may also be technically infeasible to use self-annotated annotations, which may be too complex or varied for computer vision systems. As demonstrated by other researchers, the ethical critiques of facial analysis technologies highlighted in this work cues can be extended to other “invisible” aspects of human identity, like personality traits and emotions (e.g., Wouters et al., 2019)—many of which were also present in the databases I analyzed. Using invisible characteristics, whether that is gender identity or emotion, may still enable problematic applications of facial analysis and must always be critically examined.

Explicitly define identity-specific limitations of use

I encourage authors to think through the potential implications of their databases being used in facial analysis technologies—in particular, the misuses. One way of taking caution against oppressive uses of race and gender categories in databases, as demonstrated through the discussion in Sociohistorical Sourcing: Tying Historical Approaches of Race and Gender Identification to Database Documentation, is to create licensing agreements that delimit what kinds of uses are acceptable. Many of the databases I examined in this study had licensing agreements for the type of use (e.g., commercial) that were allowed. In some cases, people who want to use these databases must contact the creators before they are given permission to use it.

However, database authors could go further. Specifically, licensing agreements should address acceptable and unacceptable identity-specific uses of databases. I saw this in one database which I examined, the Iranian Face Database (IFDB), which detailed terms of use for the images of women included in the database (Nik et al., 2007). While I acknowledge it is not possible to predict all problematic uses of a database, licensing agreements present a first step to protecting both the subjects in the database and the potential targets of facial analysis systems. A significant benefit of licensing agreements is that authors will more easily be able to identify who is using their database and can inform them of changes or even retractions. When it is unclear how much or whether to restrict use of a database, authors should engage with community groups and advocates to determine what uses are appropriate (e.g., as seen in my own prior work (Hamidi et al., 2018) and in (Woodruff et al., 2018)).

Limitations and Future Work

I acknowledge the limitations of my methods. When database authors did not clearly state how race and gender were derived or how they were annotated, I was left no choice but to read between the lines. Deeper understanding of the motivations and decisions being made by database authors requires an insider perspective; future work would benefit from interviews with database authors, both in academia and in industry contexts. Furthermore, given the propensity of utilizing crowdworking solutions for large scale database annotations, there are immense opportunities for new research on crowdsourced annotation practices. I seek to conduct future work on the positionality of diverse crowdworkers as annotators of race and gender features.

Furthermore, given the sociohistorical power structures that impact groups at the intersection of both race and gender, more work is needed to address theories of intersectionality in facial analysis databases. I observed race and gender being treated as highly disparate identities; they were not addressed as relevant to one another, except when gender was used to balance out racial categories. Future work on more theory-driven approaches to addressing intersectionality (Crenshaw, 1991; Hill Collins & Bilge, 2016; Rankin & Thomas, 2019) in image databases could help alleviate the lack of sociohistorical context I observed in this study. This work should also consider less U.S. and Westernized views of identity and structural inequality, as the theories used to examine databases in this paper are largely based on Western theories of gender and race.

Additionally, the classification of subjects through computer vision, whether the labels are derived from subjects themselves or otherwise, may simply be viewed as morally objectionable in many circumstances. In particular, when we consider the historical—and contemporary—operationalization of race and gender classification for political means. As such, even in attempts to fold in complex and self-held invisible identities, we must always consider how those databases may be appropriated to accomplish the types of oppression I discussed in this study.

Conclusion

Emerging research in the realm of FATE (fairness, accountability, transparency, and ethics) has yielded unique insights in improving equity in facial analysis technologies. Specifically, researchers have called for increased engagement with critical scholarship (Hanna et al., 2019) and complex realities of identity (Benthall & Haynes, 2019). This study embraces these calls, adopting a critical sociohistorical perspective of race and gender classification to analyze current identity documentation practices in image databases. I examined (1) for what purposes are race and gender included in image databases; (2) what information about race and gender is implicit and what is explicit; (3) what sources are being used to define categories of race and gender; and (4) what annotation practices are being used to identify race and gender in images.

To accomplish this, colleagues and I analyzed 92 image databases popularly cited in facial analysis literature. I developed a codebook to examine how database authors described their definitions and annotation practices of race and gender. I found that the majority of database authors neither provided sources for which gender and/or race were derived, nor described the annotation practice of identifying race and/or gender in database images. While a small subset of newer image databases are aimed at increasing, in particular, racial diversity and engage more deeply with literature on race and gender identities, they still rely on moments of visible identification that could be used to augment sociohistorical practices of oppression if adopted inappropriately.

I discussed the current state of the art in database construction: apolitical and obvious approaches that erase the subjective reality of external identifications. I highlighted the politicized history of race and gender identifications, including how facial analysis systems are being adapted to expand harmful and oppressive agendas. Throughout this discussion, I highlighted the role of *visible* difference, that otherwise erases the internal experiences of race and gender. I concluded with recommendations for improving approaches to database construction and documentation, including opportunities for authors to mitigate harm to subjects of facial analysis.

4
HOW IDENTITY HAS BEEN EMBEDDED IN MODELS

In the previous chapter, I showed how race and gender are built into image datasets for training facial analysis (FA) models, a subset of computer vision. This chapter now drills down into a specific subset of FA models: *automatic gender recognition (AGR)*, a facial analysis (FA) task for classifying the gender of human subjects (for overviews see my prior (Hamidi et al., 2018)). Focusing so acutely on AGR allows me to actively *evaluate* how an identity like gender is represented in existing models.

Evaluation studies are increasingly common. For example, Buolamwini and Gebru evaluated FA services from Microsoft, IBM, and Face++ and found higher gender classification error rates for dark-skinned women than for white men (Buolamwini & Gebru, 2018). Muthukumar et al. sought to isolate the reason that facial analysis systems often worked disproportionately worse on women of color than other groups, highlighting lip, eye, and cheek structure as a primary predictor for gender in FA systems (Muthukumar et al., 2018). Within CSCW and HCI, one major thread of scholarship addresses concerns over how gender is conceptualized. In Chapter 3, I discussed the potential harms when gender is collapsed into a single gender binary—"male" or "female"—rather than approaching gender as socially constructed, non-binary, or even fluid. Colleagues and I also conducted a qualitative examination of transgender and/or non-binary[9] (which I abbreviate to "trans" from here on (H. F. Davis, 2017)) users' expectations and impressions of AGR systems, uncovering widespread concern about the ramifications of AGR (Hamidi et al., 2018). This prior work highlights an opportunity to understand how gender is represented in specific facial analysis models, rather than solely in datasets.

[9] We use the term "trans and/or non-binary" to acknowledge and respect both non-binary individuals who *do* identify as trans and non-binary individuals who *do not* identify as trans. We acknowledge this difference in our shortened umbrella use of "trans" as well.

In this chapter, I speak to these concerns by studying how computers *see* gender. I intentionally focused my attention on cloud-based providers of computer vision services, focusing on how these services operationalize gender and presented as a feature to third-party developers. Specifically, I focused on answering the following research questions:

1. How do commercial FA services codify gender characteristics (through facial classification and labels)?
2. How accurate are commercial FA services at classifying images of diverse genders (including binary and non-binary genders)?
3. How do individuals self-describe internal gender identity and how does this compare with the descriptions FA infrastructures provide?

To answer these research questions, I present results from a two-phase study. I share results from a technical analysis of commercial facial analysis and image labeling services, focusing on how gender is embedded and operationalized in these services. Building on my technical analysis, I then present my evaluation study of five services. Using a manually constructed image dataset of 2450 faces with diverse genders from Instagram, colleagues and I[10] conducted a performance evaluation to determine the success rate of commercial classifiers across multiple genders. I then share my analysis of image labeling services, focused on how gender is detected by labeling services and embedded in the labels they provide. Finally, I compared these services with the content Instagram users provided in their own captions and hashtags, revealing clashes between social and technical perspectives about gender identity.

I reflect on my findings in relationship to three different perspectives of gender: internal self-held gender, gender performativity, and systematic demographic gender. I discuss how these three different perspectives emerge and are omitted through layers of infrastructure and third-party applications, resulting in people experiencing what Bowker and Star call *torque*, especially when they reside in the *residual* spaces that are unrecognizable to these systems (Bowker & Star, 2000).

[10] Colleagues in this chapter included Jacob M Paul and Jed R. Brubaker.

If researchers are going to propose design and policy recommendations, it is critical to understand how gender is currently classified in available commercial services. My findings build on previous scholarship to provide an empirical analysis of FA services. Where prior work on AGR and gender diversity has focused on academic AGR literature, I provide an in-depth analysis of the infrastructure that supports existing commercial systems already widely available for third party use. My research demonstrates how gender is conceptualized throughout multiple layers of FA systems, including an analysis of both classification and labeling functionality. I enumerate what options are currently available to third party clients, providing insight into the implications of the underlying infrastructure. Finally, I detail ethical decisions I made that may be of benefit to scholars conducting similar research—particularly as it pertains to minimizing misuse of user data when working with cloud-based services. The findings presented in this paper can lower barriers for stakeholders evaluating blackbox systems, support current approaches to fairness research, and open doors to more creative ways of imagining gender in algorithmic classification systems.

Phase I: Technical Analysis of Facial Analysis and Image Labeling

Facial Analysis and Image Labeling Services				
Name	***Service Name***	***HQ***	***Gender Class. Terms***	***Prob. Score***
Amazon	Rekognition	United States	Male/Female	Incl.
Baseapp	DeepSight	Germany	Male/Female	Not Incl.
Betaface	Betaface API	India	Male/Female	Incl.
Clarifai		United States	Masculine/Feminine	Incl.
Face++		China	Male/Female	Not Incl.
Google	Cloud Vision	United States	N/A	N/A
IBM	Watson Visual Recognition	United States	Male/Female	Incl.
Imagga		Bulgaria	N/A	N/A
Kairos		United States	M/F	Incl.
Microsoft	Azure	United States	Male/Female	Not Incl.

Table 5. The set of facial analysis and image labeling companies (and their service name, if it is different) whose documentation I analyzed. The "Gender Classifier Terms" column represents the language used to describe gender classification in the service. The "Probability Score" column indicates whether the gender classifier includes a probability score. Bolded names represent the services I studied during Phase II: Evaluating Five Facial Analysis and Image Labeling Services.

My investigation began with an in-depth technical analysis of commercially available computer vision services that included facial analysis and image labeling functionality. I start by detailing how these services function, paying close attention to what forms of data are provided and how they are organized. In line with Edwards et al. (Edwards et al., 2010), I argue that these services provide an infrastructure that empowers system designers and developers that make use of said infrastructure, but also constrains the possibilities for their designs. Specifically, Edwards et al. call attention to *interjected abstractions*—the risk of low-level infrastructural concepts becoming part of the interface presented to end-users (Edwards et al., 2010). I argue that this can occur, but that these abstractions can also be uncritically adopted by developers of third-party applications as well. Developers working with computer vision services may accept the APIs and data produced by these services as representative of the actual world (cf. (Brubaker

& Hayes, 2011; Woolgar & Suchman, 1989)). Even as computer vision services provide tools that empower designers and developers to create new applications, the data provided by these services also represent a set of affordances that designers can use, naturalizing categories and specific data values.

To select a set of services to study, I reviewed several dozen commercially available computer vision services. I initially identified services to review based on my existing knowledge, previous scholarship on these services, and online articles comparing providers. I compared these services using information from their public facing websites, including advertising and promotional content, technical documentation, tutorials, and demos. I eliminated services that (1) did not classify attributes about a human face or body and (2) did not have publicly available demos to test. This process helped me narrow the list down to the ten services I studied during this phase (see **Table 5**).

Functionality of Facial Analysis and Image Labeling Services

The computer vision services I analyzed bundled their features in different ways, but the features I analyzed can be broadly understood as falling into two categories—facial analysis and image labeling.

- *Facial analysis* employs specific feature detection functionality trained for faces. Notable for the current analysis, most services bundled a predetermined set of classifiers that were not solely focused on facial recognition. Services typically classified and categorized the image relative to other concepts (e.g., gender, age, etc.).
- *Image labeling* (or "tagging" on some platforms) provides a set of labels for objects detected in the image (e.g., young lady (heroine), soul patch facial hair). In contrast to the consistent data schema provided by FA, the specific labels and how many are included varies, depending on what was detected in the image.

To better understand the features of each of these ten services, I used free stock images including people, animals, inanimate objects, and scenery. I included images other than people at this stage to identify differences in the data returned for human versus non-human images. My analysis of the data returned from each service focused on two levels: the schema of the response and the range of values contained in that schema. To this end, I analyzed technical documentation, marketing materials, and the results from my own tests with these services.

The Schema of a "Face"

The schemas for facial analysis services were elaborate (see Figs. 3-5), but highly varied. Some services provide robust detection of the location of facial "landmarks" (e.g., eyes, nose, mouth, etc.) and orientation of the face within the image (i.e., roll, yaw, and pitch). While facial features may be indicative of gender (e.g., facial morphology (Mahalingam & Ricanek, 2013; Ramey & Salichs, 2014)), in this analysis, I focus on the classification data that would most commonly be used by third-party developers making use of these services.

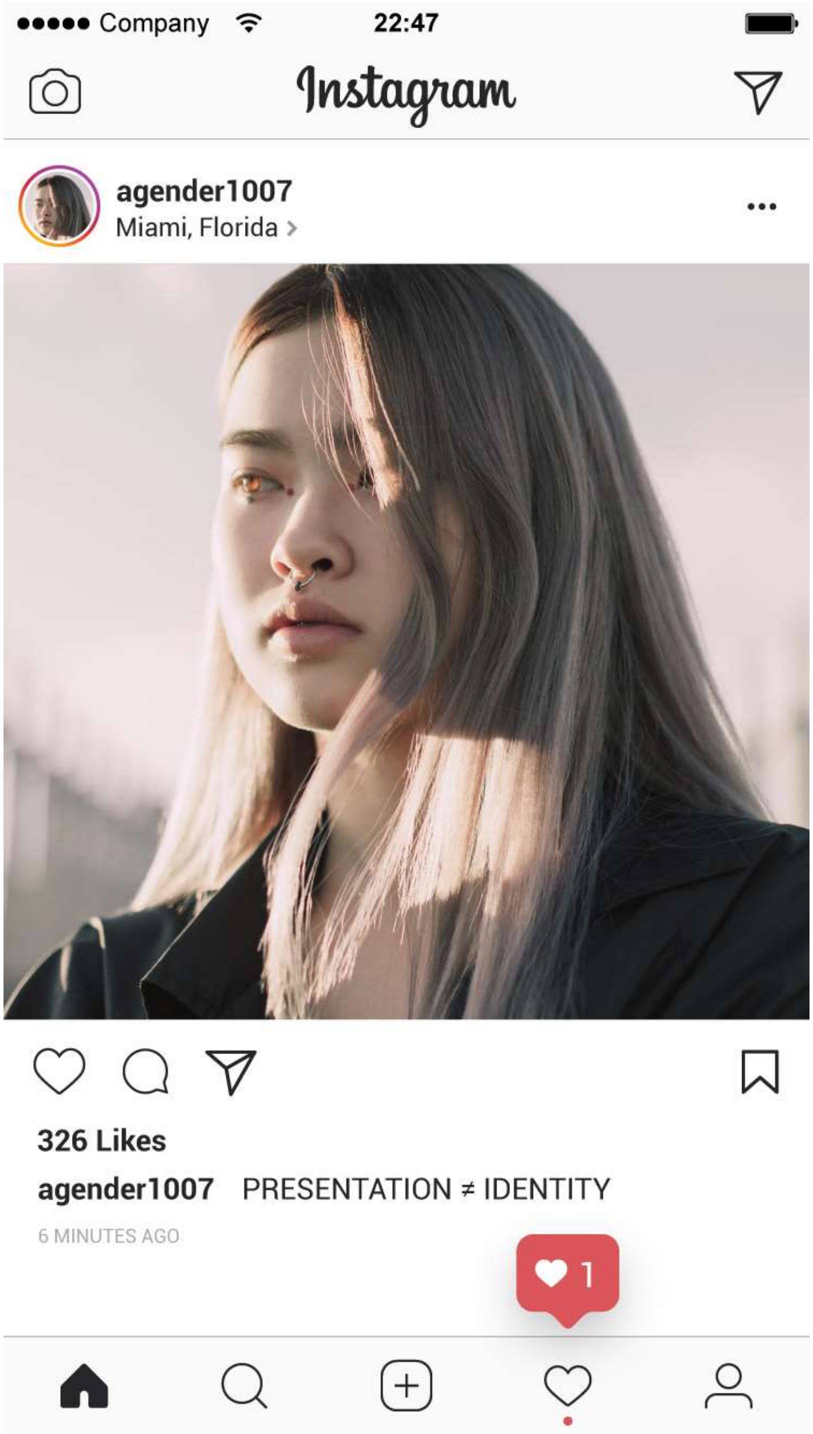
All services, save for Google's, included gender and age classification in their facial analysis. I found that services from large tech companies in the United States (such as Amazon Rekognition, Google Cloud Vision, and IBM Watson Visual Recognition) omitted ethnicity and race. However, smaller, independent companies (such as Clarifai and Kairos) and non-US companies (like Chinese-based Face++ and German-based Beta Face) included ethnicity and race.

Finally, some form of "safe search" classification was common across FA services. These classifiers included ratings for attributes like "raciness" or "nsfw." IBM's Watson, for example, includes classification results for *explicit*. Microsoft Azure has two classifiers for *adult* and *racy* content.

When returning classifier results, some services also included a probability score for the result. This was variously referred to as a "score," "confidence score," "accuracy," and so on, but represented the likelihood of the returned value being accurate.

Returning to gender classification, the range of values I observed is important. Gender was defined as a binary—and never a spectrum—with only two categories. Despite often being called "gender," the categories always used the biologically essentialist[11] terms "male" and "female" (as opposed to actual gender identities like "man" and "woman"). One interesting exception was Clarifai, whose gender classifier is specifically termed "gender appearance" and returns the values "masculine" and "feminine." These terms insinuate a potential shift in gender classification from a biologically associated category that someone is assigned to a perceived quality someone could define.

[11] Focusing on specific, fixed biological features to differentiate men from women.



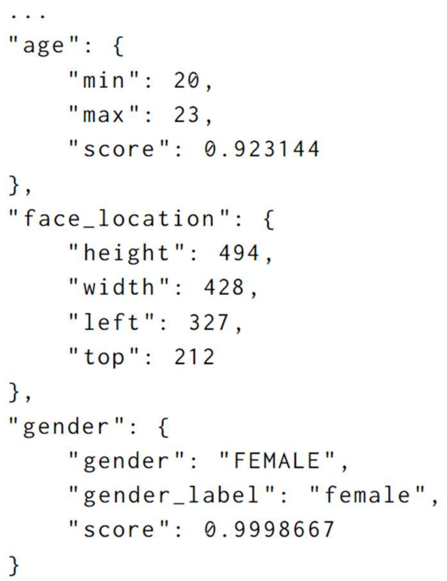
agender1007
Miami, Florida

326 Likes

agender1007 PRESENTATION ≠ IDENTITY

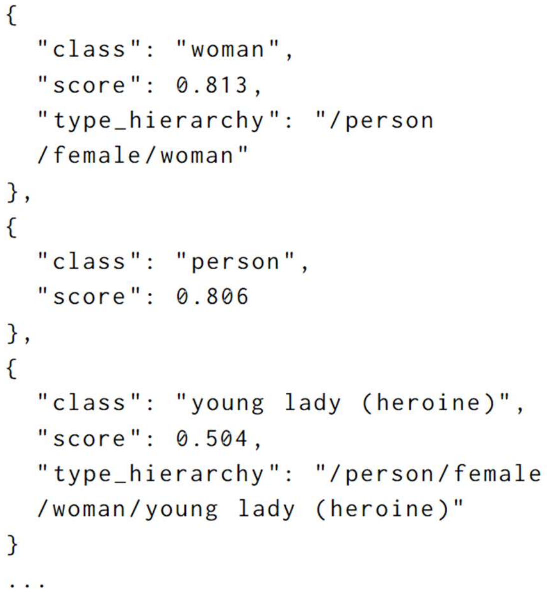
6 MINUTES AGO

Figure 2. An example of an #agender Instagram post.



```
...
"age": {
    "min": 20,
    "max": 23,
    "score": 0.923144
},
"face_location": {
    "height": 494,
    "width": 428,
    "left": 327,
    "top": 212
},
"gender": {
    "gender": "FEMALE",
    "gender_label": "female",
    "score": 0.9998667
}
```

Figure 3. An example of the gender classification portion of the result from IBM Watson's facial analysis service.



```
{
  "class": "woman",
  "score": 0.813,
  "type_hierarchy": "/person
  /female/woman"
},
{
  "class": "person",
  "score": 0.806
},
{
  "class": "young lady (heroine)",
  "score": 0.504,
  "type_hierarchy": "/person/female
  /woman/young lady (heroine)"
}
...
```

Figure 4. An example of the image labeling portion of the result from IBM Watson's computer vision service.

During analysis I also noted that probability scores (when included) never fell below 0.5 for the classified gender. Some services, like Kairos, provided two probability scores that total to 1.0 (e.g., "femaleConfidence": 0.00001, "maleConfidence": 0.99999, type: "M"), clearly exposing a binary classifier for which male and female are opposites.

In my initial exploration, I found that the results of gender classification were inconsistent across platforms. However, it was often difficult to determine why. In one instance, for example, a photo of a man dressed in drag from the dataset was classified as female (Watson) and male (Azure). These inconsistencies demonstrate the differences in how gender is operationalized across these services. However, it is unclear whether this is a result of differences in training data and the creation of models, or if there are more fundamental things at play. I return to this concern in section Phase II: Evaluating Five Facial Analysis and Image Labeling Services with a more rigorous evaluation.

The uniform simplification of gender across most services was surprising, but also prompted me to consider other places in which gender might exist but be less structured. With this in mind, I also analyzed the more open-ended image labeling features.

Labeling

In contrast with the consistent set of classifier data accompanying FA results, the data returned for image labeling is far more open-ended. As previously mentioned, labeling requests produced a list of the objects detected—but the range of labels provided is extensive. While I could not find exact numbers, Google and IBM both claim that they have classification for "thousands" of "classes." As with facial analysis, some services provided probability scores for their results, while others did not.

The absence of an explicit gender classifier, however, does not mean that gender was absent from labeling results. In fact, gender was evident throughout labels, including terms like *"woman," "man," "boy,"* and "*aunt*," to name a few. Moreover, unlike the binary gender classification, labels are not mutually exclusive. As a result, I frequently saw images labeled with multiple and seemingly contradictory terms. For example, a set of labels including *"person," "boy," "daughter,"* and "*son*" was not unusual.

Independence in Classification Tasks

Given FA and labeling were offered by the same provider, I expected there to be consistencies across gender classification in both FA and the gendered labels. If anything, I found the opposite. As evidenced by the prevalence of multiple gender labels being assigned to a single image, gendered labels were decoupled from the gender classifications in FA. Probability scores for FA gender classifications and label classifications would often differ, sometimes greatly. For example, Amazon assigned the label "*female*" (.612) to an image, yet the probability score for this label was much lower than the probability for its female gender classification (.992). This suggests that the gender classifiers in FA services are decoupled from the classifiers used to label images.

These inconsistencies provide insights into how I might better understand probability scores. Clarifai, in particular, provided a unique glimpse into how it discretely classifies gender into two categories by showing the probability scores for both male and female classification. For some of the images colleagues and I tested, the probability scores were close to one another on both sides of the binary, showcasing *a lack* of confidence in its gender classification. For example, an image of a young woman standing next to a statue was viewed as only .500002 likely to be female, and .49997 likely to be male, tipping the scales towards female just slightly. However, this image was labeled with a series of seemingly contrasting binaries: "*child*" (.986) and "*adult*" (.893), "*boy*" (.91) and "*girl*" (.904). As evidenced by the independent probability scores for these labels, classification is based on the detection of the label or not (e.g., "woman" or "not woman") rather than the either-or binaries (e.g., "female" or "male") I observed with gender classification in facial analysis.

Takeaways from Phase I

During my initial analysis, I found inconsistencies across services in how gender was classified. Despite these inconsistencies, however, both facial analysis and image labeling used binary language to describe gender. Moreover, the inconsistencies between how gender and labels were classified, even within the same service, suggests that classifiers are developed independently from each other and, subsequently, gender is being operationalized in a piecemeal fashion.

Examining the APIs associated with these services, as if I was a third-party developer, impressed the important role of data schemas. The structure of the data returned by gender classifiers to third-party developers presents gender as a property that can be easily incorporated into the design of their applications. In the process, however, much of the context around gender, or even a classifier's certainty, can be lost. Established literature on algorithmic fairness has noted how biased training data can result in biased classification and recommendations. What I see here is how these systems can also produce bias in how they framed their results—in this case, through the schemas around which services and their APIs are constructed.

What my initial study did not tell me is how image classification performed on a diverse set of images. Moreover, my analysis of image labels was not exhaustive. In the next phase of this research, I sought to quantify the performance of gender classification across a diverse set of genders and analyze a larger and more diverse set of image labels.

Phase II: Evaluating Five Facial Analysis and Image Labeling Services

Having detailed how FA and image labeling services function, and the affordances that they provide to third-party developers, I analyzed how a subset of these services performed with a dataset of gender diverse images. Specifically, I:

1. Performed a system analysis to understand the affordances of these systems.
2. Manually determined the performance of the gender classifiers found in facial analysis services using a dataset of diverse gender images.
3. Conducted a qualitative error analysis to understand how the gendered language of the system compares with the gender expressions of the Instagram dataset authors.

I start by describing how I narrowed the services studied. I then outline the construction of the dataset, including how I selected the genders studied and the inclusion criteria used for images. For both of these, I discuss how ethical considerations impacted the methodological choices made.

Facial Analysis Services

To conduct a more in-depth analysis, I selected five services based on two criteria: (1) a diversity of classification and labeling affordances (in other words, the types of information returned about images) and (2) their presence in the market and size of their user base. Specifically, Amazon (*Amazon Rekognition – Video and Image - AWS*, 2019), Google (*Vision API - Image Content Analysis | Cloud Vision API | Google Cloud*, 2019), IBM (*Watson Visual Recognition*, 2019), and Microsoft (*Face API - Facial Recognition Software | Microsoft Azure*, 2019) have been notably active this past year in their investments and involvements in the state of facial recognition technology (e.g. (Makena Kelly, 2019; R. Metz, 2019; Natalia Drozdiak, 2019)). I also included Clarifai (*Clarifai*, 2019), a small startup company known in AI and facial recognition markets for its involvement in government contracts (C. Metz, 2019). These five services each included gender information in either their facial analysis feature, their labeling feature, or both.

Ethical Considerations for Service Selection

Following my technical analysis of ten services (see Phase I: Technical Analysis of Facial Analysis and Image Labeling), I also decided to eliminate several services from the dataset evaluation on ethical grounds. It is common for service providers to make use of the data provided to them for product development. However, such data use and retention policies present ethical concerns in the context of this study. Face++, for example, retains the right to use analyzed photos for internal research and to improve their products. I decided against using these services because I was unsure what unethical or potentially harmful use cases these services might then use the data for. I reviewed the terms of service (TOS) for each service to ensure that the service I used did not store images for any purpose or use the images to further train models, or alternately, allowed me to opt out of data storage and training.

	Folksonomies Turned Hashtag		
	Instagram #		Instagram #
AFAB	28,191	***Man***	36,466,751
Agender	1,864,879	*Neutrois*	28,060
AMAB	20,332	***Non-Binary***	2,780,477
Androgyne	223,125	*Pangender*	156,596
Bigender	855,370	*Polygender*	103,620
Cisgender	97,971	*Third Gender*	12,592
Demiboy	597,320	*Trans*	5,933,800
Demigirl	592,703	*Trans Feminine*	26,060
Female	6,379,367	***Trans Man***	843,139
Femme	3,132,240	*Trans Masculine*	132,380
Gender Nonconforming	84,780	***Trans Woman***	452,743
Genderless	236,082	*Transgender*	7,849,435
Genderqueer	1,990,117	*Trigender*	171,539
Male	6,884,437	***Woman***	41,269,789

Table 6. The folksonomies generated by the seven author contacts. The Instagram Posts column indicates

the number of posts under the associated hashtag. The bolded gender labels indicate which genders I chose to use for the final dataset.

Dataset

Next, colleagues and I constructed a dataset of gender diverse images. After sampling and cleaning the data, the dataset comprised of 2450 photos that included a face, were posted publicly on Instagram, and were labeled with a gender hashtag by their author. The dataset contained seven different genders, with 350 images for each.

To identify a diverse set of gender hashtags, I crowd-sourced gender labels from seven author contacts, all of whom were queer, trans, and/or non-binary individuals. This method was especially useful for generating a list of genders beyond cisgender and binary ones, as trans communities often develop and employ folksonomies to self-identify (Dame, 2016).

Contacts provided a total of 24 unique gender folksonomies. From these, I selected seven diverse genders, weighing what was most commonly used on Instagram, while also excluding folksonomies that could have multiple meanings (e.g., androgynous, queer, transgender). The final folksonomies-turned-hashtags were #man, #woman, #nonbinary, #genderqueer, #transman, #transwoman, and #agender (see **Table 6**).

Having selected a set of gender hashtags, colleagues and I then used an open-source Python tool[12] to collect a sample of photos and their associated metadata for each hashtag. To ensure a diversity of images, colleagues and I collected images from Instagram's Recent feed rather than Instagram's Top feed. (The Top feed is algorithmically curated to showcase popular content and is biased towards celebrities, influencers, and other popular accounts.) Examining the photos returned, I discovered that a number of the images were irrelevant for analysis, such as memes or illustrations, or not suitable for simple facial analysis (e.g., images which are pixelated, low-quality, distorted by filters, etc.). As such, I adopted the following three inclusion criteria for photos:

[12] https://github.com/rarcega/instagram-scraper

1. A single human face must be present. Images without a face, or with multiple faces, were excluded. Including images with multiple faces would make it difficult to differentiate which data is associated with which face in the image.
2. 75% or more of the face must be visible. I eliminated faces that were cropped out of the photo or hidden behind an object, hand, or hair.
3. The image must be clear and not visibly altered or filtered. I eliminated photos where an individual's face was unclear or heavily distorted with filters; I also eliminated images that were low-quality or pixelated.

Knowing that facial analysis technologies are not 100% accurate (Grother et al., 2018), removing these photos allowed me to focus on gender presentation in ideal circumstances and better isolate how these services classify and label gender in each individual image.

The final dataset consisted of 2450 total photos associated with the seven hashtags, with 350 images each (a breakdown can be seen in **Table 6**). After finalizing the dataset, colleagues and I processed all images through the five selected services. I used the results in three ways. First, I performed a quantitative evaluation of the gender classification returned by face analysis. Next, I qualitatively analyzed results from labeling requests. Finally, to understand how self-held gender aligns with computer vision classifications, I conducted a content analysis of a subset of Instagram captions to compare with the face classification and labels.

Gender Hashtag Nuances

The fluid and imbricated nature of gender makes it difficult to divvy it up into neatly divided hashtags. In fact, many of the posts I collected used two or more of these hashtags. For this reason, explaining the nuanced meaning behind these labels and the way I employ them for this study is necessary.

Foremost, even the gender binary is not easily split into simple cisgender or transgender categories. For example, #man and #transman could easily represent the same person: like cis men, many trans men identify with the label "man" without the trans prefix. The same is true with trans women. Thus, I cannot assume that the men in #man and the women in #woman are cisgender. In the same way, I cannot assume that the trans men in #transman and the trans women in #transwoman identify solely with the binary. For example, some trans women may identify with the label trans women, but not women. Non-binary, meanwhile, is often used as an umbrella term. #agender and #genderqueer may fall under #nonbinary; individuals who tag with #transwoman and #transman may also be non-binary.

For the quantitative analysis (see Phase II: Evaluating Five Facial Analysis and Image Labeling Services), I collapsed gender into binary categories for the purpose of assessing performance of binary gender classifiers. #man and #transman became ground truth for “male” and #woman and #transwoman became ground truth for “female.” This allowed me to understand how binary gender classification performed on binary genders, whether cisgender or trans.

While I acknowledge the issues with collapsing genders in this way, doing so allowed me to assess and compare the true positive rates for each gender category. However, for the qualitative analysis, I examine the nuance of gender in comparison to the binary outputs of facial analysis services. I engage with how users self-describe their own genders in their Instagram photos and compare these with how facial analysis and image labeling services classify gender. This surfaces how classification binaries fail to capture the full range of gender.

Ethical Considerations for Data Collection and Dataset Construction

I recognize the sensitive nature of collecting public user data for research purposes. Thus, I considered what the benefits and the risks were to choosing this method (Fiesler & Proferes, 2018). I feel that this work is important in highlighting the current limitations, and potentially negative implications, that FA and image labeling technologies have for individuals of diverse genders. Due to the lack of available ground truth image data of trans individuals, I felt it was important to work with a ground truth dataset that did not contradict users’ self-held gender autonomy. I did not collect or store Instagram usernames. Furthermore, I destroyed all of the files containing the images and posts after the completion of analysis. The dataset constructed will not be published, to both protect user identity and to ensure user images are not appropriated for unethical or harmful research.

To further protect the identities of the Instagram users in the dataset, the images I include throughout this paper are not part of the dataset and serve only as exemplars. Exemplars are images from Unsplash, a stock website that provides license for unlimited image use for commercial and noncommercial purposes—they do not represent true ground-truth data. I also paraphrased or created composites of user quotes, rather than directly quoting users, so that the identities of users cannot be identified through search (Markham, 2012). I believe that the steps colleagues and I have taken mitigate the possibility of harm to users to such a degree that the benefits outweigh the risks.

Performance of Facial Analysis Services on a Diverse Gender Dataset

After completing the system analysis (see Phase II: Evaluating Five Facial Analysis and Image Labeling Services), I sought to understand how FA and image labeling services performed on an image dataset of people with diverse self-identified genders. In this section, I specifically focus on the gender classification provided in the results of facial classification requests.

Using the gender hashtag provided by individuals in their Instagram post as ground-truth data, I calculated the accuracy of gender classification results from four services across 2450 images. For this analysis, I examined results from Amazon, Clarifai, IBM, and Microsoft. I excluded Google as its Vision service does not provide gender classification.

I calculated the True Positive Rate (TPR) (also called recall) for each gender hashtag across each service. For the purposes of this study, I refer to this as “accuracy”—the accuracy at which the classification correctly identified the ground truth gender of the person in the image. Finally, it is important to note that I analytically calculated the accuracy rate for #agender, #genderqueer, and #nonbinary as 0%. As noted in Phase I: Technical Analysis of Facial Analysis and Image Labeling, FA services with gender classification only return binary gender labels. Given that these three genders do not fit into binary gender labels, it is not possible for any of the services I evaluated to return a correct classification.

TPR Performance Per Gender Hashtag

Hashtag	***Amazon***			***Clarifai***			***IBM***			***Microsoft***			***All***
	T	*F*	*TPR*	*T*	*F*	*TPR*	*T*	*F*	*TPR*	*T*	*F*	*TPR*	*Avg*
#woman	348	2	99.4%	333	17	95.1%	345	5	98.6%	100	0	100.0%	98.3%
#man	334	16	95.4%	344	6	98.35	341	9	97.4%	348	2	99.4%	97.6%
#transwoman	317	33	90.6%	271	79	77.4%	330	20	94.3%	305	45	87.1%	87.3%
#transman	216	134	61.7%	266	84	76.0%	250	100	71.4%	255	95	72.8%	70.5%
#agender, #genderqueer, #nonbinary	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 7. The True Positive Rate (TPR) for each gender across the face calls of each of the facial analysis services analyzed.

My analysis reveals differences across both genders and FA services (see **Table 7**). Differences in true positive accuracy likely indicate how these models are trained to recognize contrasting “female” and “male” features (e.g., lips and cheekbones (Muthukumar et al., 2018)). Due to the stark differences in accuracy between cisnormative[13] images (#man and #woman) and trans images (#transwoman and #transman), it is likely that the training data used to train FA services does not include transgender individuals—at least those who do not perform gender in a cisnormative manner. Differences between services make it evident that each service not only employs different training data to classify gender, but potentially different requirements for the underlying infrastructure driving the task of gender classification. These differences result in subjective notions about what male and female actually are to the respective system.

As seen in **Table 7**, #woman images had the highest TPR rate across all services, with the exception of Clarifai, which classified men more accurately than women. #woman was classified correctly, on average, 98.3% of the time, with Microsoft providing the highest TPR rate and Clarifai the lowest. #man had the second highest TPR rate, also with the exception of Clarifai. The average correct classification rate for #man was 97.6%. Microsoft, again, correctly classified the greatest number of images and Amazon correctly classified the least number of images. These high rates of true positive accuracy suggest that the training data used to train “male” and “female” classification aligned with cisnormative gender presentation. By extension, it may be the case that service providers considered cisnormative images best suited to the task of gender recognition when creating training datasets, technologically reproducing gender binaries in these systems.

When comparing #transwoman and #transman with female and male classification, respectively, true positive accuracy decreased—particularly for #transman. Images from the #transwoman dataset had the second lowest TPR rates across all services, averaging at 87.3%. IBM had the highest accuracy of #transwoman images, while Clarifai had the lowest accuracy; the difference between the two was 16.9%. These differences in classification accuracy suggest that these services are using different training data to classify what “female” looks like. The #transman dataset had the lowest true positive accuracy,

[13] The assumption that all individuals are cisgender, and thus cisgender is the expected norm.

averaging at 70.5%. Microsoft had the highest accuracy (72.8%), while Amazon had the lowest (61.7%). The difference between these two services was 11.1%. While, in general, accuracy rates for #transman were poor, the difference in accuracy across services similarly suggests differences in the range of images being used to train classifiers as to what constitutes "male." The training data selected likely excluded non-normative gender presentation to a higher degree than that found in "female." In other words, cisnormative masculine presentation was likely less varied and diverse in the training data. TPR differences—for men and trans men, and for women and trans women—suggest a subjective view, on the part of computer vision services, as to what "male" and "female" is.

In comparing the two trans datasets, I illustrate the differences across binary trans accuracy rates. #transwoman images had a higher accuracy rate across all services than #transman images. The most stark difference between #transwoman and #transman rates was within Amazon, with a difference of 28.9%. The lower TPR for #transman across all services suggests that variance of images for "male" is smaller, resulting in services that understand "male" as something more specific and bounded than "female." The two genders with the greatest TPR rate difference were #woman and #transman. The high TPR of #woman further suggests either a greater range of "female" gender presentations in the training data used by these providers or a more normative gender presentation on the part of the trans women in the dataset. The greatest difference here was again found on Amazon, which classified #woman accurately 37.7% more frequently than #transman. This also suggests that models trained to recognize "female" images are recognizing images of trans men as female as well.

To understand the different range of images classified as male or female, I qualitatively examined misclassified images. As suggested by the differences in classification rates, the specific images that services misclassified varied. Clarifai misclassified the greatest number of #woman images. For instance, it was the only service to misclassify an image of a woman with a ponytail wearing a leather jacket, yet it was highly confident in its classification (.951). Discrepancies between what images were misclassified across services again suggest differences between datasets used to train these systems. However, underlying all of these systems is a design choice about how to operationalize gender in the first place, and thus what data should be used to train these classifiers.

These findings led to more questions about how gender is metaphorically "seen" and classified by computer vision services, prompting me to conduct a qualitative analysis of the *labeling* assigned across diverse genders.

How Gender is Labeled, How Labels are Gendered

While most computer vision services include gender classification in their facial analysis results, many also include image labeling. While standard practices in computer vision and machine learning suggest that gender classification and image labeling are algorithmically distinct, I felt that studying the labels was important in presenting a holistic view of how gender was understood within these services. I qualitatively analyzed the labels assigned to Instagram posts across the five services. I examined the associated computer vision labels using a subset of 100 Instagram posts per gender hashtag from the full dataset. Colleagues and I hand-coded these labels for gender-specific concepts, including explicit gender labels (e.g., "*woman*") and implicit gendered concepts (e.g., clothing types, aesthetic qualities).

My analysis cannot identify how these labels were derived. However, typical industry practices involve creating lists of thousands of concepts that these systems can detect in a fairly haphazard fashion. When approached as a technical problem, designers may choose to focus on the ability to accurately detect the given object without considering the social meaning of such objects or the potentially harmful implications.

In this section, I outline how images were labeled *within* and *across* services. I demonstrate the relationships between these outputs and the Instagram posts by describing the people in the images. I also present the probability scores associated with each label in the presentation of these findings, when services made them available.

The "Cultural" Language of Labels: What is Feminine and What is Masculine

Many images were unanimously associated with explicitly gendered labels. For example, a photo of a blonde woman in a gown was labeled "*woman*," "*girl*," and "*lady*" by Microsoft. Labels were often redundant in this way, assigning many different variations of "*woman*" to single images. It was also common for labels to be feminized versions of otherwise gender-neutral words, like "*actress*" and "*starlet*."

This was rare for male-classified photos. The only examples I found were from IBM: "*muscle man*" and "*male person.*"

Gender was often implicitly manifest in the labels as well. For example, a portrait of a woman with long dark hair and heavy makeup was labeled by Microsoft as "*beautiful*" and "*pretty*," concepts often associated with feminine women. Like Microsoft, Clarifai also labeled this image with the traditionally feminine label "*beautiful*" (.937). In contrast, labels like "*beauty*" and "*pretty*" were rarely assigned to male-classified images. Likewise, it was uncommon for traditionally masculine labels to be assigned to women. For example, none of the labels contained concepts like "*handsome*" or "*rugged*."

Gendered labels extended to specific physical features and clothing as well. Services recognized traditionally masculine facial features, like beards, mustaches, and stubble. In my analysis, I did not see traditionally female equivalents. However, feminine gender was assigned to garments quite often. For example, IBM returned labels that explicitly included femininity: "*halter (women's top)*," "*women's shorts*" and "*decolletage (of women's dress)*," for example.

The implicit and explicit gender in the labels I observed showcases how the services I studied conceptualize binary genders—what is female and what is male, what is feminine and what is masculine. Masculinity was often portrayed as the "neutral" position—it was rarely used as a modifier. Femininity was, however, used to further describe otherwise neutral terms. Labels like "*actress*," "*heroine*," "*starlet*," and "*women's apparel*" invoked an explicit femininity not present in masculine labels.

A Black Box of Gender Labeling

My analysis is unable to determine why services assign the labels they do to any given image. However, when labels are gendered, it can be assumed that those annotations are based on cultural gender norms, as found in previous facial analysis literature (Muthukumar et al., 2018; P. J. Phillips et al., 2011). For example, that women wear makeup and men do not. Connections like this may be responsible for the labels assigned to an image of a #transman with long wavy hair, winged eyeliner, and red lipstick: Microsoft labeled this image with "*woman*" and Google with "*lady*" (.864).

On the other hand, gender classification did not always seem tied to binary standards of gender performance. In one case, a thin blonde "#femme" trans woman with fuchsia lipstick was labeled "*boy*" by

Microsoft. This seems to contradict the notion that cultural constructs of gender performance, such as makeup being feminine leads to "*female*" labeling. These examples demonstrate that it is actually impossible for human beings to determine how these services are making labeling decisions and what specific objects in an image these labels are tied to. While the cause may be self-evident with a label like "*toy*", it is less clear with labels such as "*pretty.*"

Increasing the complexity, image labeling does not necessarily consider the person in isolation. Instead, myriad labels are typically produced based on whatever is detected in the image. As a result, there were often seemingly divergent gender labels assigned to the same photograph. As one example, Microsoft labeled an image of a woman with long dark hair and a low-cut red dress with "*woman," "girl," "lady*," as well as "*man*." Likewise, after accurately classifying the gender of a muscular trans man with glasses, Microsoft provided both "*man*" and "*woman*" as labels. While these divergent labels still reinforce a gender binary, they also suggest a lack of clarity on what about these images results in the rather general labels of "*man*" and "*woman.*" Typical industry practices involve creating lists of thousands of concepts that these systems can detect in a fairly haphazard fashion. Given the breadth and variety of concepts that may be included on such a list, designers may focus on accuracy of detection and not the social meaning of such objects.

Self-Identified Gender versus Computer-Classified Gender

After analyzing computer vision services, their facial classification, and their labeling, I wanted to understand how gender was presented by people in the larger context of their image captions. I performed a content analysis of these captions, specifically focused on how users discussed gender in their original Instagram posts. While #man and #woman users typically did not comment extensively on gender, when analyzing trans hashtags, I identified three categories: personal narratives, gender declarations, and critical commentaries of the relationship between self-presentation and gender. While the results of classification and labeling may be presented to end users in different ways in third-party applications, and some of these results may be entirely invisible to them, I present each of these in more detail as a way of illustrating the potential impact gender classification infrastructure could have on real people. By doing this, I explicitly build on prior work examining the potential harms FA might cause in real-world scenarios (see my own prior work (Hamidi et al., 2018) or (Keyes, 2018)).

Personal Narratives Highlight Potentials for Affirmation and for Harm

It is possible that facial analysis technologies could be affirming to trans users with binary genders, as has also been discussed by participants in my prior work (Hamidi et al., 2018). Many users expressed struggling with feelings of gender dysphoria, the emotional distress associated with an individual's experience with their gendered body or social experiences. In another photo, a #transwoman user posted a smiling selfie. This user also wrote that she was having a hard day because "*dysphoria is driving [her] nuts."*[14] The services assigned this image female-centric labels, like Microsoft's "*lady*" and Clarifai's "*girl*" (.981). In these cases, classifiers categorizing her as female might be affirming.

Of course, while labeling might be affirming to some binary trans individuals experiencing dysphoria, the high rates of misclassification also presents the potential for *increased* harm. The impact of these misclassifications can also be connoted when comparing them with user statements about gender dysphoria and misgendering. For example, Microsoft, IBM, and Clarifai misgendered a #transwoman who lamented on her post: "*I spent an hour getting ready just to be addressed as 'sir' at the store.*" In another example, a #transman expressing "*severe dysphoria*" was misclassified as female by all of the services. Effectively, system misclassification could have the same negative impact as human misclassification, or even compound everyday experiences of misgendering.

Multiple gender labels could also be problematic when classifying binary trans individuals. A trans woman who wrote "*I've felt dysphoric the past few days*," was also gendered female by all of the services, but was labeled "*woman," "girl,*" and "*man*" (Microsoft). The presence of a "male" gender label, which cannot be associated with specific characteristics, might also hold negative connotations for those dealing with gender dysphoria.

[14] As described in Ethical Considerations for Data Collection and Dataset Construction, this quote is paraphrased to protect user identity.

Classifications Can Never See Non-Binary Genders

The impact of incorrect classification for non-binary people is evident in many of the captions I examined. For example, a #genderqueer user wrote in all caps: "*THIS IS A NON-BINARY ZONE. DO NOT USE GIRL/SHE/HER/HERS.*" This user was then classified as female by every service. As described in Phase I: Technical Analysis of Facial Analysis and Image Labeling and Phase II: Evaluating Five Facial Analysis and Image Labeling Services, there were no classifications outside of male or female. While gender-neutral labels were provided for some images (e.g., "*person," "human*"), gender was always present in the facial analysis services that used gender classification (every service besides Google)—and also when other explicitly gendered labels were provided.

These binaries also reinforced concerns about gender presentation in non-binary individuals. For example, a #nonbinary user posted a photo of themselves wearing heavy winged eyeliner but discussed in their post the "*dysphoria*" and "*inner turmoil*" they experience "*when wearing makeup as a non-binary person*" due to makeup's association with femininity. All of the services classified this person as female. Descriptions of "inner turmoil" at being associated with the incorrect gender offer insight into the potential emotional and systemic harms facial analysis and image labeling systems might have when classifying nonbinary genders in real-time.

Declarations of Gender that Cannot be Seen by Computer Vision

These systems lacked the ability to contextualize implicit and explicit visual markers of gender identity, particularly in the context of trans images. For example, Microsoft misclassified an image of a bearded #transman holding up a syringe, presumably for testosterone injection based on the Instagram caption which read: "*I'm back on T (testosterone) after months so hopefully I'll be back to myself.*" This is considered a definitive marker of hormone replacement therapy and trans identity and includes "insider" markers contextual to trans communities. Another photo was of a shirtless #transman with top surgery[15] scars, a visual marker of his trans identity. This image was classified as female by Amazon (though male by every other service).

[15] A term used to describe a gender confirmation procedure resulting in the removal of breasts (S. C. Wilson et al., 2018)

As evidenced through these services' abilities to only recognize "male" and "female," they also have the inability to recognize whether someone is transgender, binary or not. So, while some users expressed pride in their identities, the systems are unable to affirm this. One example was of an #agender person wearing a t-shirt that read "*not cis.*" The underlying infrastructure would not have the ability to recognize trans from this declaration. Instead, this image was classified as male by Clarifai, IBM, and Microsoft, but female by Amazon. While gender labels that align with user gender expressions could be affirming to their journeys, these services are still unable to recognize their identities as trans.

User Commentary Critiquing Gender Binary

Many users also critiqued the notion of gender being tied to performance or external appearance. When examined alongside the classification infrastructures highlighted in previous sections, these critiques could be viewed as a point of contention aimed at the premise of technologies like facial analysis and gendered image labeling. For example, a #genderqueer user expressed frustration with the normalization and authority of cisnormative binary gender, writing a series of statements that read: "*Down with cissexism. Down with cisnormativity. Down with cis privilege.*"

Another agender user critiqued the historical representation of gender as solely binary in Western cultures. They wrote that *"some transphobics say that people are inventing new genders ... but they aren't new."* They wrote a commentary in their caption outlining the history of two-spirit and transgender roles in Native American life (*cf.*, (Towle, 2005)). As presented in the previous section, the classification and labeling schemas, as well as the higher performance rate on binary genders, privileges the cisnormativity these users are critiquing.

Discussion

What is gender? A simple question with no single answer. Gender can be understood through a multitude of perspectives: a subjectively held self-identity (Stryker et al., 2006), a self-presentation to others (Goffman, 1956), a social construct defined and maintained through performative acts (Butler, 1988), and a demographic imposed by society (Rubin, 2013; Valentine, 2016). In the context of computer vision, I have shown how the design and use of facial analysis and image labeling systems collapse these perspectives into a singular worldview: presentation equals gender.

Forms of self-presentation are encoded into computational models used to classify these presentations. When classifying gender, designers of the systems I studied chose to use only two predefined demographic gender categories: male and female. As a result, these presentations are recorded, measured, classified, labeled, and databased for future iterations of binary gender classification. These gender classification models are then bundled up for commercial use, often in the form of cloud-based services, providing an infrastructure that third parties can use to create or augment their own services. In the process, these services propagate a reductionist view of gender provided by the underlying infrastructure. Self-identity is not used by computer vision systems. After all, it cannot be seen.

In order to illustrate how gender is used by computer vision services and experienced by gender diverse individuals, I synthesize the findings through an engagement with Butler's notions of gender performativity (Butler, 1988) and Bowker and Star's notions of residuality and torque (Bowker & Star, 2000)). I map my discussion to Edwards et al.'s problem of infrastructure (Edwards et al., 2010), discussing how the perspectives of identity outlined above interact across three layers: the infrastructure, the third-party applications that make use of that infrastructure, and people. Through this discussion, I highlight the layers of translation gender is sifted through as it moves from human being to infrastructure, and then back again.

Classifying Human Gender Performativity through Infrastructure

Literature on gender classification has highlighted numerous methods for identifying gender (e.g. periocular regions (Mahalingam & Ricanek, 2013); facial morphology (Ramey & Salichs, 2014); lips, eyes, and cheeks (Muthukumar et al., 2018)), but how, when, and by whom gender is embedded into the pipeline of data, labels, and models is opaque to outsiders (e.g., Kemper & Kolkman, 2018; Raji & Buolamwini, 2019). However, in examining the commercially available affordances and infrastructure, the findings shed light on how the visible presentation of an individual and their performative expressions of gender, through grooming and style, are used by computer vision systems in two ways.

First, through gender classification in facial analysis services, I see binary gender categories applied to individuals. Second, through image labeling, specific aspects of an image are detected and assigned a descriptive label (e.g., “beard”). The services I studied adopt a particular cultural view of gender that privileges self-presentation and gender performance. However, the manner in which this cultural view relies on presentation can be seen as archaic and normative, adopting systematic demographic gender categories that embrace the binary. I expound on this perspective by unpacking the infrastructure underlying both facial analysis and labeling.

Facial analysis makes use of the most rigid gender categorizations within commercial computer vision services. Even when the self-expression defies the binary mold, FA employs binary gender classification in a way that collapses diverse expressions and reinforces what gender should look like. Even if a model is trained to recognize diversity in images of men and women, it can only apply those learned standards in a constrained classification environment when classifying images of trans people. A diversity of training data will not address this bias. My analysis of FA infrastructure suggests that (with one exception, Google Vision) designers of these services decided to first, include gender in their products, and second, define it as “male” *or* “female.” The bias in gender classification cannot be attributed to algorithms alone. Its root sits with how designers conceptualized the problem in the first place.

However, it is important to note that rigid approaches to gender classification are not inherent to all of image classification. I found that, in comparison with face classification features, labeling features had the ability to assess images of people in ways that were gender neutral (e.g., the label “*person*”) or ambiguous (e.g., including multiple gender labels). Examining how labels manifest in images across all genders suggests traditional performative markers of binary genders might not be as inherent to facial classification decision-making as one might expect. Not only can “*man*” and “*woman*” exist within one image, labels can represent concepts independent of gender identity: men can wear makeup, women can have beards. Label classification is decoupled from facial classification; they do not impact the results of the gender classifiers I analyzed in facial analysis services. Perhaps this is beneficial, because binary gender classification is not determined by labels for concepts like makeup or beards.

Despite the potential occurrence of multiple labels, gendered labels themselves still typically conformed to binary notions of gender. I saw labels for man, male, boy—but not trans man. Moreover, while labels associated with men were often gender neutral, I found that women were often positioned as an outlier. Many of the labels for feminine presentation used terms that were explicitly gendered synonyms of otherwise gender-neutral concepts (e.g., military woman, gown (of women)). Specifically with labels, it is important to emphasize the subtleties in how concepts were gendered—in many cases, even when unnecessary. The abundance of gendered labels I observed points to the importance of considering gender beyond the training of classifiers, but also in the seemingly mundane human work of creating labels that will be associated with these classifiers.

Finally, labeling—because it is focused on discrete object detection—does not consider the other objects detected as contextual factors. While in many cases this may contribute to the plurality of concepts identified in images, this also presents an interesting challenge when classifying self-identity. Many of the posts by trans users included context clues intended to communicate details about an individual's gender to their viewers. Details like wearing a tee shirt with "not cis," wearing makeup as a non-binary person, or writing in a social media profile that you inhabit a "non-binary zone"—is lost when using simplistic object-based classification systems.

There is a bias encoded into systems that render only specific gender performances and specific genders visible. The consequence of current computer vision infrastructure is the erasure of residual categories of gender—categories which cannot exist in a system that is trained to recognize only traditional notions of male and female. As Bowker and Star explain, there is value in exposing residual categories: "[T]hey can signal uncertainty at the level of data collection or interpretation" especially in situations where "more precise designation[s] could give a false impression" of the data (Bowker & Star, 2000).

For those who fall into the residual categories of computer vision systems—whose gender cannot be seen by gender classification schemas—the likelihood of experiencing torque is high. An alternative to prescriptive binary gender classification might lie in embracing a polyphony of performative features, embedding labels into the infrastructure with the intention of supporting gender fluidity. Instead of collapsing gender identity to a single category (as occurs with current gender classifiers), computer vision services could embrace the fluidity of gender performativity by providing more comprehensive and inclusive gender concepts in their labeling schemas.

However, supporting a larger number of gender identities and broader definitions of any given gender is not without its limits. Bias goes beyond models and training sets, or even a new approach that might consider assigning multiple genders to an image. FA is limited by the premise that gender can be seen. Trans scholar Viviane Namaste's critique of gender and queer studies is instructive here as well: "[O]ur bodies are made up of more than gender and mere performance" (Namaste, 2000) Performances are what these computer vision systems understand, and "gender" is a prominent structure by which they have been designed to see. The premise of computer vision elides the perspective that gender is subjective and internally held, and that gender performance is not always an indicator of gender. As one #agender user wrote: "*PRESENTATION ≠ IDENTITY.*"

The Bias Propagated Through Third-Party Applications

As evidenced by the many differing computer vision services available, this technology is often designed in silo—their models, their data, and their labeling practices are generally proprietary blackboxes. However, even though this study focused on computer vision services, I cannot overlook their role as infrastructure and how the design of these services propagates into the applications that make use of them. In the previous section, I posit that the facial analysis and image labeling services reiterate archaic language about gender repackaged as neutral and technologically advanced. These services, designed to serve as infrastructure, have the potential to cascade into endless domains. In the hands of third parties—where the neutral presentation of this worldview as a technological service might not be questioned—the notion of external gendered appearances as an indicator of a binary gender classification becomes calcified. It becomes embedded in numerous other infrastructures representing numerous other use cases. Here we can see many of the now familiar critiques of big data. However, in the context of these cloud-based services, there is a shift from data that is analyzed to *affordances* that are *used*.

As I have already discussed, labels present a potential alternative to rigid gender classifications and opportunity to embrace a more diverse worldview. Yet, the diversity of the data returned by labeling services presents a challenge for third-party developers. In contrast to the data standard that gender classifiers offer to third-parties, the dynamic set of labels provided to developers provides a small, but not inconsequential, technical challenge. Labeling features only provide a list of what was detected, not what wasn't, and it can be difficult to discern *why* specific labels manifest and others do not.

My focus on how third parties make use of these services is critical as that is where it is most likely that the choices embedded into cloud-based infrastructures will cause harm. Gender identification, particularly mandated through the state, has already been used to police trans identities (e.g., by barring trans individuals from accessing healthcare (L. Khan, 2011)). Social and physical harm could be perpetrated using computer vision technology to trans individuals, who already face high levels of harassment and violence (James et al., 2016; O. Wilson, 2013). Binary gender classification in facial analysis could be used to intentionally obstruct access to social spaces (e.g. bathrooms (Bender-Baird, 2015; Herman, 2013)), restrict movement (e.g. the U.S. Transportation Security Administration (TSA) (Currah & Mulqueen, 2011)), and even enact systemic and targeted violence if adopted by virulently anti-trans governments (e.g. (Hicks, 2019)).

Even if harming trans individuals is not an intentional outcome of a third-party system, interacting with tools that use the FA and image labeling infrastructure I studied could result in torque. It is not hard to imagine how the proliferation of large scale services like the ones I have studied could also scale experiences of misgendering documented by others (Julia Kapusta, 2016; McLemore, 2015). For example, the high rate of gender misclassification I observed, particularly for trans men, results in their identities as men being erased and twisted to fit into "female" classifications. Trans women, too, were frequently erased by misclassification, compounding the archaic and dangerous conflation of trans women as men in disguise (Bettcher, 2007; Wodda & Panfil, 2015). Given the rate at which trans individuals, even in my small dataset, discussed the emotional toil of dysphoria, designers should attend to how insensitive FA classification could exacerbate the torque associated with both misgendering and dysphoria. The emotional harm—caused by misgendering and resulting in dysphoria—caused by gender misclassification can compound the torque already experienced by trans individuals on a daily basis.

Furthermore, for those who fell between these binary classifications altogether—existing only in residual categories that are not captured within the classification schema—the potential for torque is high. Binary classification forces non-binary users to conform to cisnormative expectations of gender performance. Non-binary genders were, metaphorically, molded to fit into two buckets of demographic gender (male versus female). Non-binary genders present a challenge for gender classification, but also highlight the challenges of designing human-centered systems built on computer vision infrastructure. In the eyes of these services (as they currently exist), human beings can only exist on a male/masculine versus female/feminine spectrum and that spectrum exists on a measurable, numerical probabilistic scale. As one #genderqueer person from my dataset posted in their caption: "*There is no right way to do gender, as long as you do it your way. Why settle for someone else's gender label when you can define your own?*" Yet, these services effectively assign "someone else's gender" to individuals. The opacity of these blackbox systems, and the limited understanding of how gender classifications are being made, might also intensify torque. Individuals may not understand how their gender is being classified by the system, potentially resulting in increased self-doubt and negative affect about self-presentation.

It is difficult to predict how third parties might use these commercial services, both in the present and in the future. Any number of use cases, intentionally harmful or not, could exist without the knowledge of the providers of this infrastructure. We have already seen instances of this in the alleged use of Microsoft Azure's by a Chinese company, SenseNets, to track Muslim minorities in Xinjiang (Doffman, 2019). But even beyond scenarios covered by popular press, it is likely that FA infrastructure is being used in countless smaller instances that may seem benign, but collectively reproduce a particular view of gender into our sociotechnical fabric. Political and social agendas, Bowker and Star remind us, are often first presented as purely technical interventions: *"As layers of classification system become enfolded into a working infrastructure, the original political intervention becomes more and more firmly entrenched... It becomes taken for granted"* (Bowker & Star, 2000). With this in mind, it is critical that designers, researchers, and policymakers think through designing the future of gender in facial analysis and image labeling services.

Design and Policy Considerations

Understanding how gender is represented in facial analysis and image labeling infrastructure and how those representations might impact, in particular, trans individuals who come into contact with applications that use this infrastructure leads me to contemplate two key places to intervene: design and policy. In this section, I present implications for the design of computer vision models and datasets, as well as considerations for computer vision policies and standards.

Design of Facial Analysis and Image Labeling Services and its Applications

Use gender in classification carefully.

The prevalence of gender classification across services may be an indicator that this is a feature that is important to and used by third-party clients. However, as the varied exclusion of race and ethnicity from services suggests, the creators of these services should consider why gender classification is being used in the first place. While this is perhaps obvious, I feel it is important to posit that designers carefully think what benefits gender brings to their system and consider *abandoning* gender classification in facial analysis technology. Before embedding gender classification into a facial analysis service or incorporating gender into image labeling, it is important to consider what purpose gender is serving. Furthermore, it is important to consider how gender will be defined, and whether that perspective is unnecessarily exclusionary (e.g., binary). Binary gender should never be an unquestioned default. I propose that stakeholders involved in the development of facial analysis services and image datasets think through the potentially negative and harmful consequences their service might be used for—including emotional, social, physical, and systematic (state or governmental) harms.

Embrace gender ambiguity instead of the gender binary.

When gender classification synthesizes gender performance into simplistic binary categories, the potential for gender fluidity and self-held gender identity is reduced. Labels like “person,” “people,” and “human” already provide inclusive information about the presence of human beings in a photograph. Rather than relying on static, binary gender in a face classification infrastructure, designers of applications should consider embracing, and demanding improvements, to feature-based labeling. Labels based on neutral performative markers (e.g., beard, makeup, dress) could replace gender classification in the facial analysis model, allowing third parties and individuals who come into contact with facial analysis applications to embrace their own interpretations of those features. This might actually be more precise, for the purposes of third-party applications. For example, performative markers like makeup would actually be more relevant to beauty product advertisers than gender classification, because they could then capture all genders who wear makeup. The multiplicity of labels does come with some technical overhead. Parsing and designing around a dynamic set of labels will always be more complex than simply checking for one of two values from a gender classifier. I acknowledge this, but also suggest that gender should not be simple.

Focus on contextualizing labeling.

As I explicated in this chapter's discussion, computer vision services are currently unable to piece together contextual markers of identity. Rather than focusing on improving methods of gender classification, app designers could use labeling alongside other qualitative data, like the Instagram captions, to formulate more precise notions about user identity.

If gender must be used, consider the context of its application.

If gender is something that is found to be useful to a system, designers should carefully consider the context the system will be used in. For example, while gender may be relevant to mitigating gender bias (Zliobaite & Custers, 2016), consider what kinds of bias are being privileged and what kinds of bias are being made invisible. Furthermore, consider the potential implications of gendered data being leaked, hacked, or misappropriated. For example, if attempting to mitigate bias against trans individuals, consider whether attempting to explicitly embed trans gender recognition into a model could do more harm than good. This presents a tension. On one hand, gender classification could be used for the benefit of mitigating gender bias—through recognizing performative markers of underrepresented genders. On the other hand, the same system could be adopted in a way that undermines the benefits and results in harm. Service providers should consider how to provide gender classification functionality to third-party developers in a way that enables more scrutiny and oversight.

Design of Image Datasets

Some have appealed for gender inclusive datasets, including images of trans people with diverse genders (e.g. (Schrupp, 2019); others are concerned with the implications of training facial analysis services to recognize trans identities (e.g. (J. Rose, 2019; J. Vincent, 2017)). I urge designers to consider the risks of training computer vision to identify trans individuals in an attempt to be more gender inclusive. The current work highlights the challenges involved in creating an inclusive dataset—and, in fact, argues that a truly, universally inclusive dataset is not possible. With that in mind, I recommend three approaches to consider towards developing more inclusive image training datasets. In all cases, designers should make explicit exactly what the data is being used for and ensure not to sell that data to other parties who might use it for harm.

Use self-identified gender in datasets.

When gender classification is appropriate, it is important for it to be accurate. Like I found in my analysis, the same image might be classified differently across computer vision services. Primarily, gender labeling practices currently require labeling to be done based off of subjective interpretation of external appearance. For something as complex and personal as gender, relying on datasets where human labelers have inferred gender leads to inaccuracies. However, the caveat to ignoring gender in datasets is potentially reifying gender bias (e.g. (Corbett-Davies & Goel, 2018; Lambrecht & Tucker, 2016)). If the computer vision application requires gender, creating a dataset of self-identified gender could mitigate some bias inherent in subjective labeling. To do this, designers should seek explicit consent from individuals to use their images and label data through a continuous informed consent process. However, given that computer vision is limited to what can be seen, designers might find it necessary to build systems that rely on more than just computer vision and make use of other forms of data. Given that computer vision is limited to what can be seen, designers might find it necessary to build services that rely on more than just computer vision and make use of other forms of data.

Consider the tensions of gender classification annotations in datasets.

Whether gender is necessary to include in the infrastructure of a computer vision model should determine how gender should be built into data labeling practices at all. In use cases where the classifiers' purpose is to mitigate bias, gender labeled data would be necessary for the model to function. For example, a classifier built for the purposes of trying to improve gender parity in hiring women. Trans women are also women, so should their images be labeled as "woman"? However, trans individuals are also underrepresented in hiring (James et al., 2016) and should be accounted for. This would require explicit trans labeling, which could be an issue of consent (in which trans women do not wish to be outed as trans) and open the doors for potential misuse of the system (intentionally not hiring trans women).

Focus on including a diverse set of performative gender labels.

One method to consider when building a diverse and inclusive dataset is constructing a heuristic for diverse gender markers across a range of skin tones. In doing this, designers should consider creating datasets that allow multiple performative values to exist. For example, working to develop a range of gender markers that could overlap (e.g., beards, long hair, makeup, clothing style) using participatory design methods with gender diverse individuals (including cisgender and trans individuals).

Policies and Regulations

There has already been an increasing call for policy regulation for how facial analysis technologies are built and used (Knight, 2018; Makena Kelly, 2019); some governments have already moved towards enacting such policies (Brandom, 2019; Fussell, 2019; e.g., O'Sullivan, 2019). Considering long-standing state policies around gender identification are recently changing (with the advent of legally permissible 'X' gender markers (e.g., Dance, 2019; Schmelzer, 2018; Sopelsa, 2018), current gender representations in commercial computer vision services are already obsolete in the United States, where these companies are based. I recommend future-looking policy considerations for gender classification in facial analysis and image labeling.

Create policies for inclusive standards for how gender is used in computer vision systems.

As I found in Functionality of Facial Analysis and Image Labeling Services, gender is used in some services but not all. When it is present, it is not consistent across all services. I recommend that relevant stakeholders—including designers, policymakers, engineers, and researchers of diverse genders—work to establish principled guidelines towards gender inclusivity for computer vision infrastructures. Not only would such a policy promote equity in gender representation, it would make bias auditing and harm mitigation for facial analysis and image labeling services easier. It would also benefit third-party developers who need to move between services. As the concept of gender is shifting, both socially and legally, I’d also recommend reassessing policies and guidelines regulating how gender is used by computer vision systems regularly.

Establish policies to hold companies accountable for how services are used.

Currently service providers are often not held accountable for how their services are used, but we are starting to see a shift in public expectations (e.g. (O’Brien, 2018)). However, the current architecture of these services may limit the services ability to understand the ultimate purpose or use of the third-party system (e.g. (Doffman, 2019)). I acknowledge the significant challenges here; however, it is important to develop policies that establish layers of accountability and transparency for how FA and image labeling services are used by third-party applications to ensure that identity classification in computer vision services is not used to perpetrate harm.

Treat trans identity as a protected class.

Much like how the Fair Housing Act extends its anti-discrimination policies to online advertisers (McKinnon & Horwitz, 2019), policymakers can consider how to expand legally “protected classes” to encompass facial analysis technologies. Establishing policies for how biometric data and face and body images are collected and used may be the most effective way of mitigating harm to trans people—and also people of marginalized races, ethnicities, and sexualities. Policies that prevent discriminatory and non-consensual gender representations could prevent gender misrepresentation from being incorporated into FA systems in both the data and infrastructure by regulating the use of gender as a category in algorithmic systems. For example, by banning the use of gender from FA-powered advertising and marketing.

Limitations and Future Work

At many points in this study, I reached the limitations of my methods—we cannot see inside these black boxes. However, an inside perspective is crucial for future scholarship. Conducting research on how gender is embedded into these services throughout their development pipeline, particularly by talking with designers and practitioners who are developing current systems, is necessary for a deeper understanding of why, when, and how gender is conceptualized for computer vision services. Future work should focus on uncovering the motivations and rationale behind the development of gender classification in commercial facial analysis and labeling services, and the points of translation through which gender moves from a complex social concept to a data point amenable to computation.

Furthermore, while I sought a diverse representation of binary and non-binary genders (including genderless “genders,” i.e., agender), there is boundless opportunity to include other genders in computer vision research. I also briefly discussed in Gender Hashtag Nuances that assessing FA services required assumptions to be made about gender identity and pronouns. I recognize this as a limitation, still rooted in binary conceptions of gender that assume trans men use he/him and would be situated in “male” classification categories. Also, while I did not evaluate the impact of skin tone or ethnicity, knowing that skin tone impacts classification performance of gender classification in facial analysis software (Buolamwini & Gebru, 2018), future work would benefit from analyzing gender diversity alongside skin tone and ethnicity. Certainly, more diverse genders and skin tones should be included in imagining ethical solutions to categorization schemas. Future work might also explore different measurements of accuracy and performance on diverse gender datasets. While sufficient for the purposes of this study, I also recognize that a dataset of 2450 images, sub-divided into seven 350 gender datasets, is rather small in the world of machine learning. I believe that larger datasets with more diverse genders and skin tones would provide new and interesting insights to this research domain.

Conclusion

Current research on gender classification in computer vision services, specifically with facial analysis technologies, has unearthed crucial issues of racial bias Buolamwini & Gebru, 2018) and trans representations in current automatic gender recognition approaches (Hamidi et al., 2018; Keyes, 2018). This study builds on the inroads these researchers have already paved by providing empirical evidence to support their findings about how gender is handled in gender recognition systems. I directly examined how (1) commercial computer vision services classify and label images of different genders, including non-binary genders, as well as how (2) labeling constructs a cultural reality of gender within computer vision infrastructure.

To do this, colleagues and I constructed a dataset of photos including diverse genders to demonstrate how these services see—and are unable to see—both binary and non-binary genders. Through a systems analysis of these services, quantitative evaluation of gender classification, and a qualitative analysis of images labeling, I provide new insights into how computer vision services operationalize gender. I found that binary gender classification provided by computer vision services performed worse on binary trans images than cis ones and were unable to correctly classify non-binary genders. While image labeling differed by providing labels that allowed for gender neutrality (e.g., "person") or multiplicity (e.g., "man" and "woman"), they still made use of a binary notion of gender performance.

I discussed how different perspectives are encoded in cloud-based infrastructure that propagate into software developed by third parties, potentially resulting in harm to the individuals who interact with technology that uses this infrastructure. Throughout I have highlighted the importance of considering how gender classification becomes mediated across technological layers—from infrastructure, to third-party developers, to end users. I conclude with recommendations for designing infrastructure and datasets, and outline implications for policy that would improve inclusivity and mitigate potential harm.

5
HOW CREATOR VALUES ARE COMMUNICATED THROUGH ARTIFACTS

As demonstrated by the work presented on facial analysis datasets in Chapter 3, modern computer vision research relies heavily on datasets of images and/or videos that are used to develop and evaluate computer vision algorithms. Given the centrality of datasets to computer vision practices, many researchers have proposed new data reporting practices to increase data transparency. For instance, multiple scholars have all put forth data reporting frameworks for improving dataset documentation (Bender & Friedman, 2018; Gebru et al., 2021; Holland et al., 2018). This work is forward-looking, intended to improve future dataset practices. However, these guidelines do not generally offer guidance in incorporating specific *values* into the dataset curation process, or even in articulating the values shaping dataset development. My own critiques of the values of computer vision datasets have been constrained to specific identity categories, such as race and gender (see Chapter 3), and their potential downstream impacts on model bias.

In this chapter, I focus on examining the broader politics historically and presently incorporated into computer vision dataset development. I seek to address the gap in understanding about exactly what values are present, or absent, in computer vision data practices. Unlike in social computing and technology ethics work, computer vision—and machine learning more broadly—does not have a documented culture of reflexivity about values in their work (e.g., as seen in my prior work (Raji et al., 2021) and (Jo & Gebru, 2020)).

In failing to articulate the values that shape dataset development, dataset authors aid in rendering the value-laden components of the dataset invisible, whether intentionally or not. Moreover, when values inherent in dataset development—through task formulation, collection of data instances, structuring of data via annotation processes, and more—are left unaccounted for, dataset creators signal that the myriad of decisions were not important, or even consciously made. Consequently, the resulting dataset is more likely to be viewed as a natural reflection of the world, rather than a constructed and situated reflection of a particular worldview. The manner in which dataset developers choose to present and describe their work has tangible consequences for how their dataset is adopted and, more generally, impacts cultures of dataset development and use within the field. The lack of the aforementioned documentation practices is not solely on the shoulders of individual authors; it reflects larger institutional values within the field of computer vision, leading to an unchanging culture within conferences, journals, and education that encourages better documentation. A naturalized and objective practice contributes to a culture of uncritical and unquestioning dataset use by many computer vision practitioners.

In this work, I leverage the texts associated with computer vision datasets to examine the values operative in dataset development. How dataset creators choose to describe their datasets and the processes that went into their development signals what the creators value. These texts provide unique grounds for analyzing the values underlying dataset development in the field. In this work, I focus on answering the following research questions:

1. What do authors document about the dataset curation process?
2. How do authors document the dataset curation process? What language do they employ in describing the process, the dataset itself, and its value to their audience?
3. What do answers to questions 1 and 2 communicate about the values of computer vision datasets? What values are *not* communicated?

Specifically, colleagues and I[16] analyzed documentation from 113 different computer vision datasets (114 publications) across a variety of vision-related tasks—face-based, body-based, and non-corporeal tasks like object recognition. I built an extensive codebook to capture different segments of the dataset curation pipeline, from data collection to annotation to dissemination. I employed both structured content analysis and qualitative thematic analysis. My structured analysis was focused on *what* authors

[16] Colleagues in this chapter include Emily Denton and Alex Hanna.

explicitly document in their dataset curation, in terms of what they feel is valuable to communicate to readers; my qualitative analysis was focused on excavating *how* the process was communicated, in terms of what language and statements are used to communicate the dataset process and its contribution. Synthesizing both structured and thematic analysis allowed me to identify specific, overarching values in the computer vision dataset curation process.

I present findings in three themes, focused on the different levels of the dataset process. First, I present findings on the disciplinary-specific practices of dataset authors. Second, I present findings on data instances, and what makes certain data sought after for computer vision datasets. Third, I present findings on human actors involved in the data process—the annotators and data subjects—and how dataset authors discuss their roles. Through my discussion, I synthesize these three themes into larger values around the computer vision dataset curation process.

I identify and discuss four values of computer vision datasets: efficiency, universality, impartiality, and model work. For each present value, I identify a contrasting *silenced* value—values that are overlooked or implicitly devalued in favor of the embraced values. Efficiency is valued over care, a slow and more thoughtful approach to dataset curation. Universality is valued over contextuality, a focus on more specific tasks, locations, or audiences. Impartiality is valued over positionality, an embracing of the social and political influences on understanding the world. And model work is valued over data work, with most authors focusing little on explicating data practices in favor of detailing the proposed machine learning method or model. For each silence, I recommend steps towards actively valuing them in dataset curation. In highlighting what values are currently embraced in computer vision data practices, and what values are systematically overlooked or devalued, I see opportunities for intervention throughout the dataset curation pipeline. I argue that embracing the silenced values has the potential to change the process of curating computer vision data to be more trustworthy, ethical, and human-centered.

Methods

Researcher Positionality

All three authors involved in this chapter within the computing space, particularly on issues of fairness and equity in machine learning. I have a background in human-computer interaction, as well as media studies and gender studies; the second author has a background in computer science, specifically machine learning and computer vision; and the third author has a background in sociology. All three authors conducted their work while at Google.

My colleagues and I realize that our own perspectives color how we are interpreting these values, and that computer vision researchers may not characterize their values in the same ways. This may be seen as a weakness of the study—that because researchers from the computer sciences do not interpret their values as such, our results may lack external validity. However, I don't believe that this is a weakness of this study, but a strength. Because colleagues and I explicitly acknowledge our own positionality in this analysis, I hope to remain reflexive in acknowledging our own subject position as researchers concerned with social computing and the workings of AI as sociotechnical systems. Moreover, I highlight a perspective that is oriented more towards science as both a technical *and* social practice.

Manuscript Corpus and Keyword List

I identified relevant computer vision datasets by examining the datasets used for a wide variety of computer vision tasks in academic papers. To obtain a corpus of manuscripts, I used IEEE proceedings, which is the publisher of most of the premier computer vision conferences (e.g., Computer Vision and Pattern Recognition (CVPR)). I identified computer vision proceedings by (1) selecting proceedings related to computer vision using the IEEE proceedings list; and (2) broadly searching "computer vision" in IEEE Xplore, the organization's digital library. By employing both these search methods, I was able to triangulate on potentially missed manuscripts. This method yielded a corpus of 50,694 computer vision manuscripts with metadata.

Before examining the manuscripts for datasets, I first needed to define a set of computer vision tasks that I could narrow the search to. By task, I mean the broad classification of a problem that a computer vision is meant to solve (e.g., facial recognition, object detection, pedestrian detection, etc.), which may be associated with a specific type of data (e.g., face images). Given the vast diversity of computer vision literature, selecting datasets by task ensured a diverse variety of datasets meant for different purposes. To define a list of tasks, I parsed the keyword metadata in the corpus. I used both standardized keywords (IEEE keywords) and personalized keywords (author keywords) to reduce the risk of gaps in the task list. This gave me 247,126 keywords with which to work. To make the keyword list more manageable for analysis, I narrowed to keywords used 50 or more times in the corpus. This gave me a list of 1,622 keywords, representing about 57% of all keywords in the list. I then manually removed keywords that were too high-level and not task-specific (e.g., computer vision) or were too vague to be tied to any tasks (e.g., video, engines). This left me with a manageable list of 345 keywords. During analysis, I also organically derived 10 keywords from abstracts while checking for understanding of some vague keywords (e.g., gunshot detection). The finalized keyword list was 355 keywords.

These finalized keywords represented both tasks and data types. Colleagues and I thematically clustered the keywords from the list into 21 conceptually related topics (e.g., medical tasks, low-level image processing tasks, image generation tasks, etc.). Colleagues and I then reviewed these 21 concepts as a team to determine which clusters were broadly scoped to “any type of data” (e.g., image processing, image indexing, image enhancement, etc.). We removed 12 clusters of keywords during this phase. This resulted in nine conceptual clusters of 130 keywords that we then grouped into three broad categories: Face-Based, Body-Based, and Non-Corporeal (as in, not related to human bodies). We grouped keywords into related categories of tasks under each of the three clusters (e.g., the category "gender classification" contained the keywords "gender classification," "gender estimation," "gender prediction," and "gender recognition").

Identifying Computer Vision Datasets

I went back to the original manuscript corpus and randomly sampled five papers for each task category (e.g., for gender classification). I used the keywords in each category to sample by keyword in the corpus. For example, I randomly sampled five papers that matched the keyword "body detection," and so on. If I could not identify five papers using a single keyword, or the papers did not contain dataset references, I cycled through keywords in each task category and resampled. For each random sample, colleagues and I read the paper to determine which dataset(s) were being used or cited. I listed all datasets found within each paper under that task keyword. Colleagues and I gathered 331 unique datasets from this process.

After sampling each keyword in the task categories, colleagues and I then found each original paper associated with the dataset. While doing this, we also snowballed new datasets we found in two ways: (1) we found a new dataset reference in the original dataset publication; or (2) we found a new dataset from the associated authors or organization. Colleagues and I snowballed 355 additional datasets. We did not snowball new datasets from online lists, forums, or other sources where we could not verify the methodology. We did, however, supplement face datasets from the open source list published in the associated paper of Chapter 4 (Scheuerman, Paul, et al., 2019) which provided an additional 66 datasets.[17] This provided a total of 487 datasets.[18]

Sampling Datasets for Analysis

For conducting the analysis, colleagues and I decided to sample a more manageable number of datasets from the full corpus of 487. We sampled in two ways. First, given we felt it was important to understand the most popular datasets, we sampled datasets with over 4,000 citations on Google Scholar. This yielded 13 datasets that were a mix of Face-Based, Body-Based, and Non-Corporeal. Given the size and vast documentation of ImageNet, as well as its influence on the field and large number of citations per paper, we decided to code each major ImageNet paper separately. Second, we sampled 100 additional datasets stratified by each category (face, body, and non-corporeal). We decided to sample each

[17] https://zenodo.org/record/3735400\#.YAHWk5NKjUJ

[18] In the process of conducting our coding, we also identified an additional 271 datasets, which we added to our dataset list, but were not part of the original population from which our sample was drawn, giving us a total of 753 datasets. All datasets in the corpus are provided at https://zenodo.org/record/4613146\#.YJwdwKhKiF5

category proportionally to the number in the overall corpus, rather than equally across all categories. We chose this approach to reflect the implicit popularity of certain types of datasets in the computer vision community. The final sample included 113 datasets: 47 face-based, 25 body-based, and 41 non-corporeal.

Corpus Breakdown		
Property	***Population***	***Sample***
Face-based	205	47
Non-Corporeal	174	41
Body-Based	2015	25
Citation mean / median [a,b]	421.17 / 165	390.75 / 166
Year mean / median [b] (rounded)	2011 / 2012	2011 / 2011

Table 8. A table showing the breakdown of the corpus into categories and sampling statistics.

[a]Minus top 14 papers. [b]Only databases which have Google Scholar citation information.

Table 8 displays descriptive statistics for the sample and the population corpus. Each broad category is proportionally represented in the sample. The mean year (rounded) of the population and sample are both 2011, and the median as 2012 and 2011, respectively. The range of years in the sample ranged from 1994 to 2020. The sample appears to be a good representation of the types of papers which are common in the population corpus.

Codebook Development

My focus was on the disciplinary practices of documenting the creation and maintenance of computer vision datasets. Therefore, colleagues and I developed a codebook for comprehensively capturing different stages of the dataset curation process, from motivations to annotations to the availability of the data. The documentation we coded included research publications, the datasets themselves, websites, and auxiliary materials like slide decks. Colleagues and I developed the codebook iteratively during the structured analysis phase of coding, often discovering through reading documentation or discussion among the research team new variables to capture. We met weekly to discuss disagreements and edge cases, which resulted in new category creation (e.g., new questions arose on whether the dataset used synthetic humans rather than real ones, as evidenced from my encounter with data created from generative machine learning algorithms). In the end, we coded for 95 different variables. The codebook is part of a larger research project aimed at understanding the genealogy of datasets (Denton et al., 2020).

For the purposes of this work, I focus on presenting findings which explicitly or implicitly spoke to the values imbued in datasets. Details on access to the full codebook and coded dataset can be found in the Access to Research Materials section.

Analysis

Structured Coding

Colleagues and I took both structured and thematic coding approaches for the analysis of the sample. The structured content analysis focused on identifying common themes from the literature around machine learning datasets, including descriptions of data annotators (Gray & Siddharth, 2019; L. C. Irani & Silberman, 2013), data availability for research purposes (Borgman, 2017; Pasquetto et al., 2017), properties of data subjects and categories (see Chapter 4), and descriptions of decisions more generally which typically are obviated or obscured in the data collection process (Bender & Friedman, 2018; Gebru et al., 2021; Stuart Geiger et al., 2020). Colleagues and I found that using both structured and thematic coding of the datasets allowed us to abductively reason about the different dimensions, motivations, and silences around dataset creation that either method alone would not robustly allow.

Instead of approaching the structured content analysis through a process of independently coding data (Krippendorff, 2018), the authors split up the responsibility of coding each of the datasets, with the majority of the coding being performed by the first author. Colleagues and I decided against using a formal inter-rater reliability metric because coding articles for variables involved understanding the process of developing computer vision documents, and thereby could be disputed when looking at individual instances, which is one of the cases highlighted as a reason not to seek out an inter-rater reliability metric (McDonald et al., 2019). Instead, after coding data independently, colleagues and I performed cross-check coding on each author’s codings. Every collaborator coded a random five datasets from one another and then met to discuss and resolve disagreements. I also went through the dataset entirely at the end of the project and validated the structured content by calculating cross tabulations in Python for variables which depended on each other for consistency. For instance, structured variables which relied on containing human subjects should have only been coded as Yes or No if the "Contains Human Subjects" variable had been set to Yes for the variable.

Thematic Coding

In addition to the structured content analysis, colleagues and I also thematically coded language that communicated underlying values. Dataset documentation signals what the authors found necessary to communicate to the readers and, more broadly and in aggregate, what the computer vision community finds valuable to communicate about data. Similarly, what was not communicated about the dataset in documentation, even if present in the data itself, signaled what is valued and not valued. The structured coding phase of analysis informed how I defined values, as every author became deeply familiar with the data and patterns in how datasets authors document their processes. Specifically, I define value language as statements the dataset authors imbued, often implicitly, with perspectives on importance and moral judgements. For instance, in the example of one the papers developed around ImageNet, the authors use the word *accurate* to denote desirable properties of the dataset:

> *"ILSVRC makes extensive use of Amazon Mechanical Turk to obtain accurate annotations … To collect a highly accurate dataset, we rely on humans to verify each candidate image collected in the previous step for a given synset." —ImageNet (Paper 2)*

Colleagues and I split out all value statements in each dataset and then performed an initial line-by-line open coding phase. After completing the open coding phase, we then performed a second more focused coding on each of the statements, grouping them into higher-level themes (e.g., the variety of open codes that could be grouped under the theme of "unbiased data"). Colleagues and I also developed categorical relationships between these higher-level themes (e.g., unbiased data and realistic data were what we defined as "desirable data properties"). Colleagues and I concluded thematic coding phase by writing memos on each of the higher-level themes and their categorical relationships, and discussing and refining those memos as a team (Charmaz, 2006).

Access to Research Materials

Given that computer vision datasets and their associated publications are largely available to the public, I felt it ethically responsible and appropriate to release the corpus of datasets. Colleagues and I have created an open access repository for: (1) the codebook of 114 coded datasets; (2) the population corpus of computer vision datasets; (3) the documentation on the coding procedures; and (4) the Python analysis code. For each of the datasets, I listed the original publication, venue it was published at, year published, and Google Scholar citation count at the time of collection. I encourage future researchers to use these resources for additional research or to build on this work. The data can be found at https://zenodo.org/record/4613146#.YJwdwKhKiF5.

Findings

I summarize the fourteen focused codes constructed from my qualitative content analysis and the broad themes under which they fall: *dataset authors and their disciplinary practices, data and its data properties* and *human actors as annotators and data subject*s. Codes are not mutually exclusive; some of them are highly related.

Dataset Authors and their Disciplinary Practices

Data as Essential for Scientific Progress

Several of the datasets that colleagues and I sampled (18 of 114; 15.8%) describe data as being crucial to particular subfields progressing beyond their current state-of-the-art. Although most of the papers I evaluated discussed some kind of new method or algorithm in concert with the release of the dataset (15 papers were dedicated entirely to the documentation of the dataset), the data typically worked in service of making progress to the field. This is because data can present new challenges, define a new task, or build a common research agenda for improving performance on a task. For instance, the authors of the Lippmann2000 dataset write that creating benchmark datasets allows standardization and mitigates the need for practitioners to create their own datasets:

> *"These images have* **accelerated advancement in the field** by, first, allowing scientists and engineers practicing at or near the state-of-the-art to carry out their work **without the additional burden of needing to become experts in generating quality images**; and, second, creating a small level of **ad hoc standardization** such that processed images are more quickly evaluated due to general familiarity with the original input." — Lippmann2000

Similarly, many present the lack of relevant data to be one of the limiting factors in driving research in their subfield, and thus present their dataset as contributing to improving research in that subfield. The act of creating a dataset was both a barrier to entry for new work on a given task, but also an attempt to introduce a standard around the under researched task. Data is difficult to attain, and its attainment is a necessity of the machine learning process, including standards-making. The authors of Leeds Sports Pose introduce data to improve pose estimation:

> *"As noted, current methods have been* **limited by the lack of available training data** – to overcome this we introduce a new annotated dataset of 2,000 diverse and challenging consumer images which will be made publicly available." —Leeds Sports Pose

However, some of these overt claims around the necessity of data to progress are not reflected in the actual practices of data management and curation used by researchers. The dataset itself was the main explicit contribution for a slim majority of papers (59 of 114 papers; 51.8%), while a method or algorithm was the main contribution of the dataset in 53 of 114 papers (46.5%) or an empirical study in 2 of 114 papers (1.8%). Still, the dataset is rarely unaccompanied by a methodological innovation: papers typically contain some kind of new algorithm for the computer vision task under question (97 of 114 papers; 85.1%). Lastly, colleagues and I looked at how much of the paper was dedicated to describing the dataset. We calculated this value by counting the number of paragraphs dedicated to describing the dataset over the total number of paragraphs in the paper. **Figure 5** shows a bimodal distribution of how much of the paper is dedicated to describing the corpus, with a mean at 0.41 and a standard deviation of 0.33. This seems to indicate that papers are either wholly dedicated to the description of the dataset, or they provide scant information on the dataset, opting to discuss methodological innovations instead. Given the majority of papers were archival publications (105 of 114; 92%), the lack of documentation in a paper about a dataset occurred even in archival publications—publications in professional venues like conferences and journals.

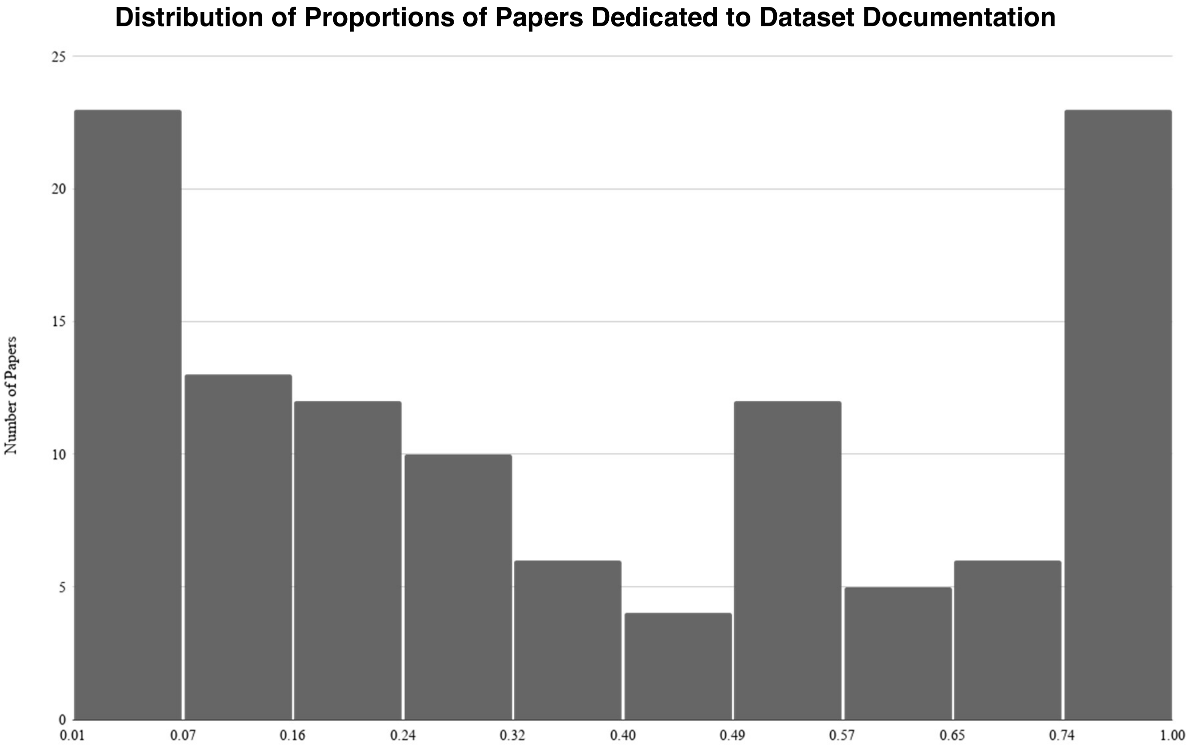


Figure 5. A histogram showing the distribution of the proportion that papers in the corpus dedicated to documenting only the dataset. Proportion was calculated by hand-counting paragraphs in each paper. The histogram is bimodal, with the majority of papers having either (a) near-0% of the paper about the dataset or (b) near-100% of the paper about the dataset.

Standardization for Evaluation and Reproducibility

Some dataset authors (BigHand2.2M; KAIST; UMass FDDB; 300-W; IUPR) noted that they needed a common set of evaluation benchmarks, because the evaluation criteria which had been used in their subfield had too many different types of quantitative evaluations and needed a standardized benchmark. For example, the authors of 300-W were motivated to create their dataset to establish a standardized benchmark:

> *"The main goal of this challenge is to compare the performance of different methods on a new-collected dataset using the same evaluation protocol and the same mark-up and hence to develop the first* **standardized benchmark** for facial landmark localization." — 300 Faces in the Wild

"Standardized" takes on a specific meaning for dataset authors. Generally, that meaning is one of quantitative measures which are viewed as reliable or objective. The authors of the IUPR dataset describe the impetus of benchmark datasets as being more "objective" as a ground for machine learning datasets:

> *"Ground-truth datasets are crucial for* **objectively measuring the performance of algorithms** in many fields of computer science. The availability of such datasets for use in research and development lays the basis for comparative evaluation of algorithms." —IUPR

Given the importance of data to standardizing algorithmic performance, dataset creators often claim to release datasets for the purposes of reproducibility and replicability, noting that this has been a failing of methods reporting results in the past. The authors of UG^2 highlight that research would benefit from reproducible standards in datasets:

> *"New video benchmark dataset represen*ting both ideal conditions and common aerial image artifacts, which we make available to facilitate new research and to simplify the **reproducibility of experimentation**." —UG^2 Challenge Dataset

Dataset authors also devalued qualitative or heuristic assessments of data quality and classifier results. Qualitative assessments were seen as less standardized and therefore less reliable. For example, the authors of HumanEva write that qualitative assessments decrease certainty and make it difficult to rigorously compare methods:

> *"Despite clear advances in the field, evaluation of these methods remains mostly* **heuristic and qualitative**. As a result, it is difficult to evaluate the current state of the art **with any certainty or even to compare different methods with any rigor**." — HumanEva

Overall, dataset authors valued standardization, in terms of quantitative measurements, because they viewed those standardizations as objective, reliable, and reproducible.

Open Source Data

The majority of the datasets which were sampled provided a URL or some web identifier for obtaining the dataset. 69 of 114 (60.5%) of the datasets provide a URL in the paper; in addition, I was able to discover 28 additional websites via a web search of the publication or the dataset name, bringing the total number of sites were able find to 97 of 114 (85%). Several authors stated this explicitly in their documentation (Abstract Paintings / Artistic Photographs; KinFaceW; IterNet RGB-D; NWPU-RESISC45; SFEW; Urban Stereo Scene; UvA-NEMO Smile; UG^2). For instance, the authors of the UvA-NEMO Smile Database state that:

> *"The database, its evaluation protocols and annotations are* ***made available to the research community****." —UvA-NEMO Smile Database*

Other authors (ASLLRP SignStream; CUAVE; SFEW) directly mention the research medium through which their data will be distributed, highlighting its benefit to the research community:

> *"Finally, one of the main purposes of all speech corpora is to allow the comparison of methods and results in order to stimulate research and fuel advances in speech processing. This is a main consideration of the CUAVE database,* **easily distributable on one DVD**." —CUAVE

While most of the datasets report an URL (or have a findable URL through web search), many of the datasets did not have any institutional mechanism for the stability of the datasets. Despite the expressed value of open-source datasets to the research community, most of the datasets are not maintained in a stable repository. Only 3 of the 114 (2.6%) of the datasets attached a Digital Object Identifier (DOI) to them, while only one (0.9%) was posted on an institutional repository such as Dataverse or Zenodo—The Child Affective Facial Expression (CAFE), hosted on NYU's Databrary (https://databrary.org/). Stability of these identifiers and hosting of datasets matter: of the 69 datasets which had a URL in the paper, only 46 (66.7%) are still available. Of the 80 datasets in which data was openly accessible (without signing a user agreement, agreeing to a Terms of Service, or downloading a software package), 59 (73.8%) were still downloadable. **Table 9** summarizes full details about data hosting and availability.

Dataset Availability via Publication Documentation

Dataset Property	**k**	**N**	**%**
URL in Paper	69	114	60.5
Any Website (In Paper + Discovered through Search)	97	114	85
Website in Paper Still Available	46	69	66.7
Data Still Downloadable	59	80[a]	73.8
DOI	3	114	2.6
Hosted on Personal/Lab Website	102	114	89.5
Hosted on Institutional Repository	1	114	0.9

Table 9. A table showing the breakdown of dataset availability in the corpus.

[a] The number of datasets which did not require registration to download.

Technical Documentation

In many cases, the amount of documentation for the technical setup of data collection and algorithm design far outstripped the amount of documentation available for the actual data collection work and annotation and labeling procedures (see **Figure 5**). The authors of the KAIST Multi-Spectral Day/Night Data Set, created for autonomous vehicle research, provide a good example of this kind of technical detail in the configuration of lens capture devices placed on top of data collection cars:

> *"In our case, we select a long focal lens to observe remote objects. If wanting to a wider field of view, a short focal lens may suffice. Recently, Tesla and Mobileye devised tri*focal camera system (HoV–20°, –50°, –150°) as a new hardware configuration..." — KAIST Multi-Spectral Day/Night Data Set for Autonomous and Assisted Driving

The authors of the CMU-MultiPIE dataset, used for facial recognition research, report on data collection hardware, illumination configuration, and requisite computing hardware:

> *"To systematically capture images with varying poses and illuminations during data acquisition we used a system of 15 cameras and 18 flashes connected to a set of Linux PCs. An add*itional computer was used as master to communicate with the independent recording clients running in parallel on the data capture PCs." —CMU-MultiPIE

Reporting the technical details of hardware is ostensibly done to report improvements on data collection procedures, to resolve disagreements about instrumentation for other researchers working in this space, or to report on more precise conditions needed for future reproducibility of the experiment and data generation.

Data and its Desirable Properties

Diverse and Varied Data

Common dimensions of diversity referenced by dataset creators include the diversity of scenes and object categories (ImageNet (Paper 1); KAIST; UG^2; Scene Geometry Layout, GHIM-10k); diversity of conditions of image capture, such as sensor quality, camera angle, or lighting conditions (TUM GAID; IUPR; Leeds Sports Pose; Exact Street2Shop); diversity of object poses or background clutter (IUPR; NWPU-RESISC45; ImageNet (Paper 1); Animals on the Web; INRIA Pedestrian; BSDS300); and diversity of the expressions or poses of individuals in the images (Exact Street2Shop). The diversity of camera quality was especially important for datasets that were motivated by real-world applications, where low-quality sensors would likely be common—most commonly, in intended surveillance applications (SCFACE; ND-QO-Flip; 300-W).

Dataset creators emphasized the importance of diverse training data for ensuring model robustness to variation that is present in real world settings (Nis Web-Collected; SYNTHIA; SCFACE; COCO-Text) and diverse testing data for providing better estimates of real-world performance (300-W; PPB; MORPH; IIITD). For example, the authors of Nis Web-Collected Database emphasized that diversity of age data could lend to universal age estimation:

> *“*[T]he derived human age estimator is **universal owing to the diversity and richness** of Internet images and thus has good generalization capability.” —Nis Web-Collected Database

> *"In practice, even the best visual descriptors, class models, feature encoding m*ethods and discriminative machine learning techniques **are not sufficient to produce reliable classifiers if properly annotated datasets with sufficient diversity are not available**." —SYNTHIA

Diversity was also motivated as crucial for the effectiveness of deep learning methods, a class of machine learning which methods have a very large number of parameters relative to traditional computer vision methods (NWPU-RESISC45; SYNTHIA). For example:

> *"[W]e can generate a* **broad variety** of urban scenarios and situations, which we believe is very **useful to help modern classifiers based on deep learning**." —SYNTHIA

> *"In addition, almost all existing datasets have a number of limitations, including the small scale of scene classes and the image numbers, the* **lack of image variations and diversity**, and the saturation of accuracy. These limitations **severely limit the development of new approaches especially deep learning-based methods**." —NWPU-RESISC45

I found notions of diversity or variety to be closely coupled with the concept of "natural" data. For example, creators of 300-W created a dataset of "*naturalistic, unconstrained face images*" that included variations such as "*unseen subjects, pose, expression, illumination, background, occlusion, and image quality."* This connection was also visible in the way dataset creators described diversity as a property that emerges naturally from "unconstrained" data collection methods:

> *"The images in this collection display* **large variation in pose, lighting, background and appearance**. Some of these variations in face appearance are due to factors such as motion, occlusions, and facial expressions, which **are characteristic of the unconstrained setting** for image acquisition." —UMass FDDB

Generally speaking, I found dataset diversity to be closely coupled with notions of realistic data and challenging data. Diverse datasets were ones that more closely mimicked the real world, and in doing so presented new challenges.

Unbiased Data

Dataset creators frequently motivated unbiased data as desirable[19], often connecting dataset bias to issues of generalization. For example, some attribute a model's failure to generalize—from one dataset to another, or from a dataset to the real world—to the existence of dataset bias (KinFaceW; KITTI). In a similar vein to discussions of diversity, unbiased benchmark datasets were motivated through their role in comparing and standardizing algorithms (ImageNet (Paper 2)). Discussions of dataset biases were generally divorced from a broader examination of social or cultural bias or the impacts of dataset bias on individuals from different sociodemographic groups. The only dataset in the corpus to explicitly connect socio-demographically biased data with discriminatory outcomes was PPB:

> *"It has recently be*en shown that algorithms trained with biased data have resulted in algorithmic discrimination..." —PPB

Otherwise, biased data was implicitly understood to be a negative property, and one that might impact classification performance at a more general level. Bias was generally discussed in relation to the process of data collection or properties of images themselves. For example, selection bias, photographer bias, recency bias, and biases in object/person pose or illumination were all topics of discussion (VAIS; PASCAL; COCO-Text). Despite frequent references to unbiased data, I observed that many claims of unbiased or less-biased data remained largely unqualified. For example, some dataset creators describe data that has been created without the particular machine learning task in mind as less biased with no further justification or explanation (PASCAL; COCO-Text):

> *"The use of personal photos which were not taken by, or selected by, vision/machine learning researchers* **results in a very 'unbiased' dataset**, in the sense that the photos are not taken with a particular purpose in mind i.e. object recognition research." —PASCAL

[19] One notable exception to this is the Animals on the Web dataset that described intentionally inducing a bias between the training and test distributions so that the test set was biased towards more "difficult" examples.

Mentions of bias frequently overlap with discussions of diversity or variety, with the assumption being that sufficient variation across a particular aspect of the dataset minimizes the potential for bias. For example, the creators of INRIA Pedestrian state:

> *"Many are bystanders taken from the image* backgrounds, so **there is no particular bias on their pose**." —INRIA Pedestrian

The desire for unbiased data was intricately tied to controlling for human bias in data collection and annotation. I discuss how dataset authors work to manage human bias in Human Bias in Data Collection and Annotation.

High-Quality Data

Dataset creators frequently described their datasets as high quality. The dataset creators that specified what they viewed to be "high quality" data tended to focus on two aspects of datasets: the images themselves and/or the annotations associated with the images. High quality images were generally described as those captured by high quality sensors and as having high resolution (UG^2; NOAA Fisheries; Urban Stereo Scene). For example, the authors of the Urban Stereo Scene Labeling Benchmark Dataset declared the higher resolution of their images in comparison to previous datasets:

> *"To our knowledge, the only comparable urban segmentation dataset with stereo vision data has been proposed by [16], which* **our dataset exceeds in terms of ... image resolution** (1024 × 440 px vs. 360 × 288 px), which is an essential factor for appearance-based segmentation." —Urban Stereo Scene Labeling Benchmark Dataset

Many of the datasets were also described as having high-quality labels (BigHand2.2M; Stanford Region Labeling; IterNet RGB-D; PASCAL; SUN RGB-D; MS-COCO; ImageNet (Paper 1); Animals on the Web; Middlebury Stereo; BSDS300). High quality labels were often defined in terms of label accuracy relative to pre-specified "gold standard" labels or label consistency across different annotators (Stanford Region Labeling; ModaNet; ImageNet (Paper 1)). Other dataset creators used the performance of models trained on the dataset to validate the accuracy of annotations (BigHand2.2M). The authors of the BSDS300 dataset describe how they obtained high quality labels through their annotation interface:

> *"In addition to simply splitting segments, the user can transfer pixels between any two existing segments. This provides a tremendous amount of flexibility in the way in which users create and define segm*ents. The interface is simple, yet accommodates a wide range of segmentation styles. In less than 5 minutes, one can create a **high-quality, pixel-accurate segmentation with 10–20 segments using a standard PC.**" —BSDS300

Datasets containing exclusively high quality images, in terms of clarity and pixel values, were often viewed as oppositional to realistic or challenging data, which I highlight in Diverse and Varied Data and Challenging Data. For this reason, some dataset creators emphasized the importance of including images with varying levels of quality in order to better reflect the real world:

> *"There are* **large variations in the quality of the contributed photographs**, lighting, indoor vs outdoor environments, body shapes and sizes of the people wearing the clothing, depicted pose, camera viewing angle, and a huge amount of occlusion due to layering of items in outfits ... These characteristics reflect the **extreme challenges and variations that we expect to find for clothing retrieval in real-world applications**."
> —Exact Street2Shop

However, high-quality annotations were nearly universally sought after. An overall high-quality dataset was one that was useful and accurately annotated, even if image quality was purposefully varied.

Realistic Data

Realistic data was described as data that was reflective of more "natural" conditions, in terms of not controlling for lighting, pose, expression, angle, occlusion, or other image factors. "Realistic" was often used to describe data captured in an uncontrolled or unposed settings (UG^2; ND-TWINS-2009-2010; KinFaceW; SFEW; SCFACE); collected from the internet (KinFaceW); collected in a public spaces (SCFACE); or data that is varied along several dimensions such as sensor quality (SCFACE). Some dataset creators described the realism of scene arrangements, for example, by describing a computer mouse on the floor as "unrealistic" (SUN RGB-D). Finally, datasets that consisted of synthetic imagery described photorealism as a desirable property that they strived to achieve (IterNet RGB-D; GTA5; SYNTHIA; Middlebury Stereo).

Dataset creators describe realistic data as being critical for measuring and comparing model performance (George Mason University Kitchen; CUAVE; SCFACE); for developing models that generalize to real-world settings (BiosecurID; Street2Shop; Subway; Middlebury Stereo); and, more generally, advancing research (SFEW). For example, the authors of BiosecurID discussed how data generated through laboratory experiments is of limited utility when estimating model performance in the real world:

> *"In order to overcome the difference in performance between laboratory experiments and practical* implementations, there is an **urgent need for the collection of realistic**

> **multimodal biometric data** which permits to infer valid results from controlled experimental conditions to the final application." —BiosecurID

I observed close ties between the notions of realism and variety, both of which are understood to give rise to more challenging datasets. I discuss these properties in more depth in the next section, including how realistic data is challenging.

Challenging Data

Dataset difficulty is another characteristic often touted by dataset creators. Dataset creators frequently motivate the creation of a new dataset with reference to saturated model performance on previous, easier datasets. For example, the creators of INRIA Pedestrian state that their new algorithm gives *"essentially perfect results on the MIT pedestrian test set, so we have created a more challenging set"* (INRIA Pedestrian). Dataset creators also describe the importance of new and more challenging datasets for advancing progress in the field (SCFACE; DUT-OMRON; ImageNet (Paper 1)) and mitigating the potential for methods to overfit (LFW). For example:

> *"As computer vision research advances, larger and* **more challenging datasets are needed** for the next generation of algorithms." —ImageNet (Paper 1)

The characteristics that constitute challenging data tend to relate to variability along a variety of dimensions including, but not limited to, sensor quality, image scale, illumination, object or person pose, camera viewpoints, and locations of data collection (INRIA Pedestrian; Stanford Region Labeling; ND-QO-Flip; OSU Thermal; Exact Street2Shop; Leeds Sports Pose; H3D; VAIS). For example, the authors of the OSU Thermal Pedestrian Database wrote that their data was challenging because of varying backgrounds and thermal intensities, and they want to collect more challenging data for future work:

> *"The approach was demonstrated with a* **difficult dataset of thermal imagery with widely-varying background and person intensities**. In future work, we plan on **extending the dataset to include additional situations involving many more distractors** moving through the scene." —OSU Thermal Pedestrian Database

As touched on in Realistic Data, challenging data was often characterized as data that mimics real-world (unconstrained) conditions and real-world variation:

> *"Similar to conventional face recognition, when the targets are unconstrained faces in the wild [12] (i.e., variation in pose, illumination, expression, and scene) the* **difficulty level further increases**, and the same being true for kinship recognition. These are, unfortunately, challenges that need to be overcome. Thus, FIW **poses realistic**

> **challenges needed to be addressed before deploying to real-world applications**."
> —FIW

In short, challenging data was compatible with realistic and diverse data, but often incompatible with high-quality data, in terms of resolution quality. Low resolution images proved more challenging than high resolution images.

Comprehensive and Large-Scale Data

Dataset creators often described their datasets as large-scale, indicating their datasets contain a large number of data instances and, in some cases, a large number of categories. The importance of large-scale datasets, or "big data," are varied. Some creators emphasized the importance of large datasets for improving generalization performance and reducing risks of overfitting (BigHand2.2M; LFW; SUN RGB-D). Relatedly, some creators identified large datasets as key to the development of reliable models and reliable evaluation methods (OU-ISIR Gait). But a number of dataset creators describe large scale datasets as being broadly essential to advancing computer vision research (KAIST; ImageNet (Paper 1 & 2); BigHand2.2M). Related to the above category of standardization and benchmarking, dataset creators specifically called out the importance of large-scale benchmark datasets:

> *"Because data-driven AI-based methods have enabled breakthroughs in both academia and industry,* **large-scale benchmarks have become one of the most important factors** to advance this technology." —KAIST

Several authors specifically identified the importance of big data for advancing deep learning methods in particular (NWPU-RESISC45; SYNTHIA; BigHand2.2M; One-Million Hands; GTA5):

> *"However, the* **lack of publicly available `big data' of remote sensing images severely limits** the development of new approaches especially deep learning based methods." —NWPU-RESISC45

Many dataset creators included claims of dataset completeness or comprehensiveness, often referencing the dataset size or high variability of data instances (FIW; TUM GAID; CUAVE; BigHand2.2M). Others describe the categories or annotations that structure the dataset to be comprehensive or complete (GTA5; VOC; SUN; ImageNet (Paper 2)). The ImageNet creators claim their dataset provides the "*most comprehensive and diverse coverage of the image world*" (ImageNet (Paper 1)). Similarly, the authors of SUN wrote that their dataset of scenes contained all discursively important images:

> *“First, we seek to quasi-exhaustively determine the number of different scene categories with different functionalities. Rather than collect all scenes that humans experience - many of which are accidental views such as the corner of an office or edge of a* door - **we identify all the scenes and places that are important enough to have unique identities in discourse, and build the most complete dataset of scene image categories to date**.” —SUN

Large-scale and comprehensive data was compatible with notions of challenging and realistic data. As described further in Annotation Labor and Time Costs, high-quality annotations become increasingly challenging and costly as datasets grow in size and scope.

Human Actors as Annotators and Data Subjects

Authors often discussed the presence of human actors in the dataset construction and documentation process, ranging from the authors themselves to annotators and data subjects. By and large, human considerations were often distinctly aimed at making the data collection and annotation processes objective. By objective, I mean that there was an attempt to mitigate or remove human subjectivity from all data processes. This was particularly salient when discussing the role of data collection and annotation. Human subjectivity was seen as a detriment to consistent, "clean," comprehensive, and accurate data and annotations—the properties of desirable data. Less commonly, dataset authors also discussed the diversity of human data subjects, from the perspective of the technical benefit of diverse human subjects for developing models. In this section, I discuss the four different human considerations I found: *annotation labor and time costs; human properties as a barrier; humans as diverse data;* and *human bias in data collection and annotation*.

Annotation Labor and Time Costs

A number of datasets utilized human annotation (63 of 114; 55%). Within papers, there is a variable amount of information available on the identity of human annotators and the cost of their labor. In cases in which human annotation is used, human annotators are described in 40 of 63 (63.4%) cases. Of these, 23 of 63 (36.5%) were reported as third-party workers (mostly from Amazon Mechanical Turk); 9 of were the authors; 6 were students; and 5 were some mixture of these. Only 5 of 63 (7.8%) papers report annotator demographics, and only 4 of 63 (6.3%) papers report if annotators were compensated.

A major focus in discussing human annotation was the time and monetary cost of annotation, particularly as a barrier to annotating large-scale datasets (BigHand2.2M; Leeds Sports Pose; UNBC-McMaster; Stanford Region Labeling; ModaNet; SYNTHIA; MS-COCO; PASCAL; ImageNet (Paper 1); LFW; GTA5), which were otherwise highly valued by dataset authors, as highlighted in Comprehensive and Large-Scale Data. High costs negatively result in "*slowing down the development of new large-scale collections like ImageNet*" (SYNTHIA). Yet, manual annotation was also seen as desirable for "high-quality" data. For example, the authors of SYNTHIA wrote of the necessity but difficulty of having large amounts of annotated data:

> *"Having a sufficient amount of diverse images with class annotations is needed. These annotations are obtained* **via cumbersome, human labour** which is particularly challenging for semantic segmentation since pixel-level annotations are required." —SYNTHIA

Here, we can see the SYNTHIA authors express the need for human labor. This need stemmed from the view that manual labeling was more accurate than automated labeling, and ground truth accuracy of labels is a desirable data property (see High-Quality Data). However, while manual labeling was often prized for its accuracy, authors tried to minimize the amount of time and money spent on human annotation labor (BSDS300; FIW; ImageNet (Paper 1); MS-COCO; Stanford Region Labeling). Often, Amazon Mechanical Turk and other crowdworking platforms were viewed as an extremely valuable tool for the computer vision community, particularly due to the demand for a great deal of labor at a low cost (SUN RGB-D; SUN; MS-COCO; ImageNet (Paper 1); H3D; Stanford Region Labeling; CORE). The following examples showcase authors touting minimal labor and low costs:

> *"We now discuss the pro*cedure followed to collect, organize, and label 11,193 family photos of 1,000 families with minimal manual labor." —Families in the Wild (FIW)

> *"We constructed using Amazon Mechanical Turk (AMT), at a total cost of less than $250." —Stanford Region Labeling* Dataset

The use of human annotators and explication of human annotations as valuable to computer vision signals the importance of human beings in the modeling pipeline. Yet, there is also the goal of minimizing human labor costs, suggesting a devaluing of labor that is otherwise valuable to the process of dataset curation. Further, there is also an underlying expectation that human annotators are flawed, in that they make mistakes or introduce subjective beliefs that do not align with the expectations of the authors.

Human Properties as a Technical Barrier

Much like the subjectivity of human annotation can risk the desired objectivity of labeling, human behaviors may be seen as technically problematic to either building a dataset or the resulting accuracy of a model. The complexity of human properties was generally perceived as a barrier to human data collection and task specification. Human characteristics were portrayed as difficult to control ranged from the diversity of human appearance (UMass FDDB; SHEFFIELD; ND-QO-Flip; MORPH; BP4D-Spontaneous; CMU-MultiPIE; ModaNet; Paper Doll; INRIA Pedestrian; TUM GAID; Leeds Sports Pose; OU-ISIR Gait; Exact Street2Shop) to the autonomy of human decision making (M3; IIITD). For example, the authors of Leeds Sports Pose described the difficulty of pose estimation given the way range of human appearance and natural imaging conditions:

> *"The task is particularly challenging* **because of the wide variation in human appearance present in natural images due to pose, clothing and imaging conditions**." —Leeds Sports Pose

Notably, no authors discuss ethical considerations in their use of the data. Human autonomy posed issues to accessing data deemed necessary for the desired task. For instance, the authors of M3 discussed how it was difficult to collect biometric data, like fingerprints, due to subjects revoking their consent due to privacy concerns. Human desires for privacy makes some data difficult to come by:

> *"We found that collecting fingerprint data is especially dif*ficult because some recruited subjects later **decided that they are reluctant to provide their fingerprint data due to privacy concerns**." —M3

Similarly, the authors of IIITD Plastic Surgery Face Database described having difficulty building a dataset of before-and-after plastic surgery images. They discussed that people did not wish to share these images with them or online due to privacy concerns, which they instead scraped from the web as a result:

> *"Due to the sensitive nature of the process and the pr*ivacy issues involved, it is extremely difficult to prepare a face database that contains images before and after surgery." —IIITD Plastic Surgery Face Database

I noted that even datasets with especially sensitive human data, such as the nude detection dataset by Lopes et al., where nude images were collected from the web, had no discussion of ethics, privacy, or even an ethics review process. Only 5 of the 100 datasets (5%) containing human subjects mentioned having an IRB or international equivalent (Forensic Facial Examiner Study; HumanEva; MORPH; ND-TWINS-2009-2010; CAFE). Five mentioned privacy considerations in any capacity (Beauty 799; BiosecurID; FACEBOOK100; KITTI; SCFACE). For example, the authors of SCFACE outlined limiting access to specific subject images due to privacy concerns:

> *"For legal reasons and for the privacy of the database participants, images that can appear in reports, papers, and other documents published or released are those with subjectID: 001, 002, 045 or 102 in the SCface datab*ase." —SCFACE

When performance failures occurred, such failures were sometimes attributed to issues of human diversity or behaviors. Such failures might occur when training data is too controlled to reflect the real world. For example, the authors of OU-ISIR Gait Database wrote of their model's gait classification failures:

> *"These failures mainly originated from the unique walking style (e.g., some subjects raise their arms higher than generic subjects) or special clothing (e.g., a long dress or coat), whi*ch cause a large difference between this test sample and the generic training samples." —OU-ISIR Gait Database, Large Population Dataset with Age

Given the barriers presented by human-based data, whether face or body, some authors would describe how they mitigated or bypassed such barriers. For example, authors might trade-off real-world diversity of appearance by attempting to implement controlled conditions (inside or outside of studio settings) during the data collection process (e.g., M3). In the case of privacy or consent concerns, authors have instead scraped images from the web to surpass participant autonomy (e.g., IIITD).

Humans as Diverse Data

As previously highlighted in Diverse and Varied Data, diversity took on a very specific meaning: describing diverse instances of data. This most commonly included such instances as a diversity of object types, a diversity of lighting conditions, and a diversity of angles. The term "diversity" was not commonly used in terms of human conditions, such as race, gender, or ability. Even in cases where the task was tied to human conditions, like age, race, or gender, authors did not necessarily discuss diversity in terms of representation. Of the datasets containing images of humans, 41 of 100 (41%) provided information about the sociodemographic diversity of data subjects.

When used to describe humans, diversity was most often attributed to diversity of ages (CASIA NIR-VIS 2.0; FIW; Nis Web-Collected; OU-ISIR Gait; RAFD; UvA-NEMO Smile; MORPH; SFEW). Other occurrences included diversity of race or ethnicity (PPB; CAFE; MORPH; Ethnic DB; FIW). Diversity of gender was rarely discussed in the sample; gender was entirely binary, as I discuss in Chapter 4, and few datasets discussed the distribution of men versus women (PPB). Diversity statements were often written like those in CAFE and The CASIA NIR-VIS Database, describing that diversity in the data as a feature:

> *"It* is also **racially and ethnically diverse**, featuring Caucasian, African American, Asian, Latino (Hispanic), and South Asian (Indian/Bangladeshi/Pakistani) children." —The Child Affective Facial Expression (CAFE)
>
> *"In the new database,* **the age distribution of the subjects are broader**, spanning from children to old people." —The CASIA NIR-VIS 2.0 Face Database

Much like claims about diverse data instances in Comprehensive and Large-Scale Data, there were attempts to claim universality of the human diversity captured in the data. The authors of Nis Web-Collected Database claimed a universality in age estimation due to the diversity of their data:

> *"The derived human age estimator is* **universal owing to the diversity and richness of Internet images** and thus has good generalization capability." —Nis Web-Collected Database

Generally, diversity was posited as a technical benefit. That is, a diversity of human characteristics benefits the technical accuracy of the model proposed. A few dataset authors discussed benefits beyond the technical; that is, given the potential deployment of the system itself, how it could socially benefit people to use diverse human data. For example, the authors of PPB discuss how groups underrepresented in data would suffer from the resulting lower accuracy rates but nonetheless suffer social consequences:

> *"In other contexts, a demographic group that is* **underrepresented in benchmark datasets can nonetheless be subjected to frequent targeting**." —PPB

However, diversity of the humans involved in the overall creation of the dataset was attributed primarily to data subjects. Most authors did not discuss the diversity of annotators or those involved in the data collection process. There were a few exceptions. Five datasets included demographic information on annotators (Beauty 799; Forensic Facial Examiner Study; ModaNet; RAFD; CAFE). Of these 5 datasets, most focused on gender and age distribution, but 2 also provided ethnicity information (CAFE; Beauty 799). For example, the authors of CAFE described the demographic distribution of their annotators:

> *"One hundred undergraduate students (half male, half female) from the R*utgers University-Newark campus participated (M = 21.2 years) … The sample was 17% African American, 27% Asian, 30% White, and 17% Latino (the remaining 9% chose ‘Other’ or did not indicate their race/ethnicity)." —The Child Affective Facial Expression (CAFE)

Diversity in human data stood to ensure data users of the utility of the data, in terms of being unbiased and ideally more accurate for all groups. Though not explicitly described, the few instances of described diversity in annotator demographics insinuated to data users that representation in annotation would result in less biased or skewed labels.

Human Bias in Data Collection and Annotation

As demonstrated throughout this section, much discussion about the role of human beings in constructing and annotating datasets is around controlling human behaviors. Another aspect of controlling human behaviors is mitigating human bias. When discussing the roles of researchers and annotators, it was often to ensure readers that human bias was accounted for and mitigated in the collection (Multi-Spectral Pedestrian; Abstract Paintings / Artistic Photographs; COCO-Text; UG^2; VAIS; Animals on the Web; PASCAL) and/or annotation process (BigHand2.2M; PETS04; SYNTHIA; ImageNet (Paper 1)). For example, the authors of the PASCAL VOC dataset wrote that their data collection process resulted in unbiased images:

> *"The use of personal photos which were not taken by, or selected by, vision/machine learning researchers* **results in a very `unbiased' dataset, in the sense that the photos are not taken with a particular purpose in mind** i.e. object recognition research." —PASCAL VOC

Some authors would justify their choice of collecting certain types of data due to potential biases. For example, the authors of the VAIS dataset decided to collect images to avoid the "photographer bias" they associate with web images. The authors of COCO-Text discussed how the MS COCO dataset’s collection method makes it a good source for unbiased text images:

> "**Images from the web often suffer from photographer bias**, in that images with more aesthetic appeal tend to be uploaded." —Maritime Imagery in the Visible and Infrared Spectrums (VAIS)

There was also concern that, although manual annotation was viewed as valuable (One-Million Hands; Texas 3D; CUAVE; UAVDT; USF; PETS04), and sometimes are superior to automated annotation (ModaNet; SUN), that human subjectivity would result in inaccuracies (BigHand2.2M; PETS04; SYNTHIA; ImageNet (Paper 2)). For example, the authors of ImageNet (Paper 2) explained two sources of bias that they attempted to control for during the annotation process:

> *"While users are instructed to make accurate judgment,* **we need to set up a quality control system to ensure this accuracy**. There are two issues to consider. First, human users make mistakes and not all users follow the instructions. Second, users do not always agree with each other, especially for more subtle or confusing synsets, typically at the deeper levels of the tree." —ImageNet (Paper 2)

Given the risk of human subjectivity and its associated errors, many authors employed checks to mitigate human subjectivity (PASCAL; 300-W; ImageNet (Paper 2)). Therefore, they could benefit from the robustness and accuracy of human visual assessments in their annotation, while ensuring those annotations met their expectations. For example, the authors of the PASCAL VOC challenge and its associated dataset discuss how the organizers of the challenge employed manual checks on annotations to ensure they were correct, in terms of what was expected:

> *"Following the annotation party,* **the accuracy of each annotation was checked by one of the organisers**, including checking for omitted objects to ensure exhaustive labelling." —PASCAL VOC

Through these statements, dataset creators meant to ensure potential users of their creations that author and annotator contributions would be unbiased.

Discussion

The increasing application of computer vision technologies in public life presents profound stakes for how human beings interact with the world. The tasks that computer vision models are designed to do are ancillary to the data used to train, test, and validate those models. Data is crucial to the process of model design, and therefore to the advancement of the field of computer vision. Thus, it is a key point of the computer vision pipeline for examining how values become embedded into the technical artifacts designed by researchers and practitioners in the field.

In this chapter, I conducted a large-scale analysis, focusing on a range of disciplinary practices in dataset collection, curation, annotation, and release, across both human and non-human related tasks. Analysis centered around common themes previously identified in prior literature on machine learning datasets, including data annotation, data availability, data categories, and data collection processes. Through a structured and qualitative content analysis, I uncovered both explicit statements and silences about data, and what those statements and silences implied about the values of datasets in computer vision. Overall, I found that computer vision dataset authors valued *efficiency, universality, impartiality,* and *model work*.

I characterize the values of computer vision datasets by discussing them as trade-offs to otherwise silenced values: *efficiency versus care, universality versus contextuality, impartiality versus positionality,* and *model work versus data work*. The values explicit in computer vision datasets often ignored, or implicitly critiqued, values of human-centered computing approaches. For example, social computing has embraced value-centered and value-sensitive design (Friedman, 1996) that includes considering context (Chancellor et al., 2019; Muller et al., 2020; Taylor et al., 2017), reflexivity and positionality (e.g., Garcia & Cifor, 2019; Kaeser-Chen et al., 2020), and situated expertise (e.g., Easley et al., 2018; Kempe-Cook et al., 2019; MacKay, 1999). For each silence, I offer recommendations for dataset authors to begin to address values in the design process, drawing on concepts from prior work in social computing and algorithmic fairness.

Many recommendations are aimed at actions that individual dataset authors can take, but I acknowledge the role of larger institutional incentive structures that may prevent individuals from effectively implementing change. Thus, I would similarly encourage larger institutions—conference venues, journals, and academic departments—to engage with my recommendations at an institutional level. For instance, NeurIPS has recently developed a "Datasets and Benchmarks Track"[20] to incentivize work on machine learning datasets and as *"an incubator to bootstrap publication on data and benchmarks."* Further, while I focus specifically on computer vision given the empirical focus of this

[20] https://neuripsconf.medium.com/announcing-the-neurips-2021-datasets-and-benchmarks-track-644e27c1e66c

chapter, the highlighted values and silences, and subsequent recommendations, can be abstracted to other machine learning domains—though I would still advocate specific values analyses for each domain.

Efficiency over Care

Dataset authors valued efficiency, both in terms of the time spent and the associated costs, monetarily and computationally, of gathering and annotating data. Authors sought desirable properties, in terms of objective, unbiased, neutral, and comprehensive data, that are easily available, quickly and cheaply classifiable, and able to be quickly but accurately annotated. The value of efficiency was clear in a number of practices in the findings. In terms of higher-level disciplinary practices, a focus on efficiency may have led many dataset authors to document only the barest technical details of the dataset creation process. Few authors wrote more than a few paragraphs documenting their dataset practices, insinuating not only a focus on model work over data work (see Model Work over Data Work), but an efficient and highly condensed means of conveying the technical details viewed as most important to the dataset. In seeking desirable data, many authors employed the practice of scraping publicly available data from websites, seen as both an easy and cheap method of amassing large amounts of data in a short period of time. Human concerns about privacy and data ownership were sometimes posited as barriers to collecting data, while the reasonings for those concerns were otherwise ignored.

Some dataset creators invested time, labor, and money into dataset debiasing and quality control measures in the service of high quality data and annotations. These investments were often justified with respect to the potential gains in algorithmic advances that stem from large-scale datasets with quality annotations. These investments were often framed as a necessary cost, but still a cost to be minimized as much as possible. Many authors employed crowdworkers for annotating that data, explicitly due to the low cost of paying crowdworkers for annotation work. Further, crowdsourcing platforms like Amazon Mechanical Turk allowed authors to hire many cheap laborers at once, leading to faster annotations. Untrained annotators were often framed as desirable, with some dataset developers investing significant efforts in developing their own annotation interfaces that would allow untrained annotators to be utilized en masse (e.g., ScanNet).

The lack of attention paid to the perspectives, labor, and rights of human actors in the dataset curation process points to the lack of value in the field in explicating or thinking through the human role in technical practices. As discussed by Chancellor et al., the discursive practices in defining the human subject in machine learning is both dehumanizing and a threat to scientific rigor (Chancellor et al., 2019). This is particularly salient in datasets with highly sensitive data, like the referenced nudity detection dataset by Lopes et al., which proposed nudity detection as a form of object detection but did not engage with social issues of privacy or gender-based bias in their documentation.

Valuing efficiency was at the cost of care, valuing slow and thoughtful decision-making and data processes, considering more ethical ways to collect data and treat annotators, and seeking fairer compensation—or even reporting compensation—for data labor. Generally, compensation for either data or annotation labor was not reported. Further, authors did not discuss the costs of efficiency to ethical scientific practice or potential harmful social implications, such as publishing sensitive data (e.g., Lopes et al.'s nudity dataset) or contributing to the documented class divisions between technology experts in industry and academia and the gig economy associated with crowdworkers (Fort et al., 2011; Pittman & Sheehan, 2016; Williamson, 2016). In general, there was little to no discussion about ethics when conducting work with annotators or with human subjects as data instances. Even commonly accepted forms of ethics accountability were not regularly reported by dataset authors, such as IRB or international equivalents of institutional ethics checks.

A counterexample within the corpus which prioritized care within the process was the Child Affective Facial Expression (CAFE). The authors of CAFE sought explicit consent from parents to collect face data from their children. The dataset collection went through IRB review, who deemed the children a vulnerable population and required the authors to outline potential harms and expected benefits to participants. While data collected in a studio setting and required participants to explicitly consent may result in a far slower and more laborious data collection process, it also shows mindfulness towards the privacy and autonomy of human data subjects. To further incorporate a value of care into their processes, the authors of CAFE might have included ethical considerations of data use and decided to compensate both data subjects and annotators.

Recommendations for Incorporating the Value of Care

Valuing care might lead to less efficient, but more reflexive data practices (Miceli et al., 2021). While not an exhaustive list of potential approaches, I highlight several areas for dataset authors to embrace care over efficiency:

- Collect data with special attention to the privacy and ownership rights of those data. This may mean going through the laborious step of obtaining permission from copyright holders (for example, when scraping Flickr for object recognition tasks) or data subjects themselves (for example, when collecting images of human faces). Given the significant overhead of studio data collection processes, and its associated deficiencies in terms of realism and diversity, authors should consider reaching out to copyright holders on social media sites for explicit permission of data use, instead of otherwise scraping those photos without permission. Asking for explicit use could allow a balance between efficiency and care, and still allow dataset authors to collect properties of desirable data. Initiatives around data licenses with differing data use permissions (e.g., Contractor et al., 2022) are useful resources for streamlining collecting data more ethically, and initiatives around labels offer appropriate labels for specific groups (e.g., for indigenous groups (J. Anderson & Christen, 2013).
- Compensate data subjects for their data and annotators for their labor. Many institutions, in both academia and industry, have fair contribution guidelines for human subjects research. Litman et al. found that high-quality responses on Amazon Mechanical Turk necessitated payments above the minimum wage (Litman et al., 2015). Dataset authors should consider the context of compensation, in terms of what is most valued by annotators; money or gift cards may not be the most appropriate compensation for all contexts. For example, Hodge et al. found that money was not useful to their participants who live in care homes (Hodge et al., 2020). Further, if the authors expect to make a profit from their dataset, they should consider what fair compensation would be to their subjects and authors given that profit necessitates their data and labor.
- Institutional review board (IRB) or ethics reviews, while certainly imperfect, are not the norm in machine learning; often, because they are not required for scraping data, researchers do not submit an IRB at all (Metcalf & Crawford, 2016). Normalizing IRB and ethics review processes for

machine learning is an ongoing conversation within machine learning communities.[21] However, IRBs and other institutional ethics review processes could provide checks to minimize harm to data subjects or annotators during the dataset curation process. Given the nature of machine learning data to have downstream impacts on the model, researchers should also report predicted uses of the data and how it may implicate the privacy of human data subjects.

- Store and safeguard datasets with proper data stewardship protocols in place, such as gatekeeping access to data, terms of service and potential licensing agreements. Institutional repositories such as the ICPSR and university libraries can aid in this process and implement best practices for data curation and preservation (*Guide to Social Science Data Preparation and Archiving: 6th Ed.*, 2012). Dataset authors can work with librarians and data stewards to provide open access to other researchers with proper agreements which set conditions and terms of use. Dataset authors might look to initiatives such as IBM's data privacy passports to protect sensitive data across multiple cloud infrastructures (*IBM Data Privacy Passports*, n.d.). An added benefit is that data repositories can record histories of data use and access, which can open up the ability of external researchers to independently audit these histories.

Universality over Contextuality

I found computer vision dataset creators valued large-scale, diverse, and realistic data that lent to a belief in inherently comprehensive or complete categorical classifications of real-world phenomena. I observed the widespread valuation of these properties was rooted, in large part, in an assumption that larger and more varied datasets provide better approximations of the real world and thus afford the development of high capacity models that are able to generalize well to varied real-world settings. Implicit in this belief is the value of universality, insinuating a world that is able to be neatly captured and classified, often for the purposes of state and economic management (Johnson & Scott, 1999; Koopman, 2019; Murphy, 2017). I also observed annotation practices were frequently portrayed as objective; annotation quality checks were aimed at ensuring annotators' worldviews matched the dataset author’s. Universal data properties

[21] For example, a recent workshop on "Navigating the Broader Impacts of AI Research" was hosted as part of the top-tier machine learning conference NeurIPS (https://nbiair.com/)

were viewed as valuable not only to real-world application, but also to standardization and reproducibility. It would be difficult if not impossible to standardize classifications about the world. Framing quality around a presumed objectivity of labels suggests a universal ordering.

Despite diversity posing more difficulty to fully capturing the world, dataset authors implicitly acknowledged that the world is full of diverse data. Capturing that diverse data in the dataset meant that the model's observation of the world would be more complete. This included not only a diversity of different types of objects for object recognition tasks, but a diverse array of human beings. For example, the importance of the inclusion of varying lighting conditions and poses was a frequent point of discussion. Authors of human-centered (face- and body-based) datasets also highlighted diverse ages, ethnic categories, and gender distributions, as these were seen as important categories of diversity for human subjects.

Universality was embraced at the expense of *contextuality*, how circumstances such as time, location, or use shape the world and thus the data in a dataset. For example, the geographic origins of images within object recognition datasets were rarely discussed. The language employed to classify objects or people in datasets was not attended to. Why specific identity markers were chosen for representing diversity was absent. Further, how important technical components of explaining the diversity of data, such as lighting, differently affected different groups—such as those of different ethnic origins—was never discussed. Beyond categorizing the data itself, authors also rarely discussed the potential impacts of the dataset, and resulting computer vision systems developed from it, on members of different social groups. Datasets were often posited as generalizably useful to broad tasks, such as general object recognition or human detection. A notable exception is the PPB dataset, which was motivated by differential classification scores in computer vision for women with darker skin tones.

As discussed by prior work, the natural world is actually difficult to capture and classify, and an attempt to reduce socially-shaped categories down into data is difficult (De Vries, 2012; Mager et al., 2018). To embrace the value of contextuality would be to embrace this difficulty and instead focus on the circumstances of use. For example, where data will be used, who should be included in the dataset based on its intended use, and how factors like culture, language, and location should shape the data collected. Further, contextual decisions would need to be motivated by which data subjects might be impacted by data curation decisions and how. LFW provides a good example of contextuality by stating explicitly that there is no naturally accepted distribution of faces for all possible domains. This move acknowledges different worldviews on human diversity. Similarly, NOAA Fisheries was constrained to a very specific context-of-use—classification to aid in live fish recognition within fisheries—which informed what kind of data would be collected and how it should be classified.

Recommendations for Incorporating the Value of Contextuality

Context can be scoped in multiple ways—from the context the dataset is meant to be used in (e.g., a specific discipline, work practice or area of application) to the context the dataset is meant to capture (e.g., the diversity of plants in North America). Designing for specific contexts, in terms of specific workplaces, communities, and cultures, has a rich history in HCI and user-centered design that dataset authors could use to guide their work (Clemmensen & Roese, 2010; Grudin, 1994; Hayes, 2011).

- Design datasets for specific temporal, cultural, geographic, or community contexts, rather than for generalized, universal use. Datasets scoped to specific contexts would be more rich and robust for those contexts (e.g., having a dataset of only fish increased the opportunity to robustly capture and classify fish in a useful way than a dataset trying to capture every animal on earth). Similarly, datasets scoped to certain cultures should reflect cultural language and expectations, a method for reducing cultural bias that stems from universalism.
- Conduct empirical studies to understand the context of intended use *before* designing the dataset (or associated methods or models). Employing empirical methods, such as surveys, interviews, or ethnography, can ground dataset curation decisions and make datasets more contextually useful for intended stakeholders.

Impartiality over Positionality

Although universality was highly valued in the findings, I found that the universe of data was often highly constrained to a specific, impartial worldview. Dataset authors strive for *impartiality* on behalf of the data collectors (often the authors themselves) and the annotators. They seek a debiasing of human subjectivity, so that data is “unbiased” and, implicitly, more trustworthy. Bias in this case takes on a statistical and cognitive definition. In particular, selection biases or observer biases were of major concern to authors. They strived to ensure potential data users that there were no selection biases driving them to select specific types of data (e.g., only images which held aesthetic appeal may have been uploaded to the web for Maritime Imagery in the Visible and Infrared Spectrums (VAIS)). They also tried to mitigate potential observer biases in annotation by focusing on how the subjective perceptions that annotators have would detrimentally lead to inaccurate labeling.

While there are explicitly stated concerns about introducing human bias, in both the process of data collection and the process of annotation, there is little discussion about how bias, in terms of hermeneutic perspectives or as a matter of interpretation, is unavoidable in vision-based tasks. Dataset authors did not report on their own *positionality*, such as how one's social and professional position can give rise to differential resources and knowledge gaps. For example, industry might afford higher budgets for recruiting data subjects, but individuals may lack the domain-specific information of professionals working in, for example, marine biology (relevant to fishery datasets); nor did dataset authors report how identity characteristics might impact the perspectives of annotators, such as how local or regional culture might influence perspectives on beauty. Instead, they assumed that there are inherently neutral practices to strive for, disregarding the rich scholarly history discussing how all human decisions are inherently value-laden (Teo, 2014; Wallace, 2019; Zhang et al., 2020).

Others have argued that this false objectivity lends to a decrease in the utility of computer vision data (see Chapter 4). While dataset authors discuss their work from the perspective of increasing trust and reliability in its objectiveness and utility, they unknowingly decrease that trust and reliability by refusing to describe its most relevant stakeholders. The tasks that datasets are meant to assist both computer vision researchers and domain experts with are inherently human-centered, but datasets lack a human-centered approach to their construction and documentation.

Despite the critical role annotators play in determining the contents of a dataset, dataset creators tended to omit critical details regarding who was performing the annotation tasks. For example, only five publications provided any demographic information regarding who annotated the data instances. As Elizabeth Anderson argues, objections to value-laden inquiry misunderstand the goals and methods of science (E. Anderson, 1995). Values are inherent in all science and adopting specific subject positions do not diminish scientific practices of reliability and empirics. To better incorporate values of positionality into datasets, authors would need to acknowledge professional and personal identity characteristics of both authors and annotators, and embrace a position that there is no “view from nowhere” (Haraway, 1988).

A positive example from analysis includes the Pilot Parliaments Benchmark (PPB) dataset. The authors include rationales for data selection, as well as acknowledgments of the identity-based limitations of their work, specifically around binary gender categories. They also report that their ground truth labels come from a board-certified surgical dermatologist, showcasing how an annotator’s experience and professional training can provide trust, rather than distrust, in the data process.

Recommendations for Incorporating the Value of Positionality

Attia and Edge provide guidance towards becoming a reflexive researcher, building trustworthiness in a research approach through the expression of personal values and professional skills (Attia & Edge, 2017). I suggest data authors embrace movements towards reflexivity, rather than attempting to remain impartial and rejecting their own subjectivity or the subjectivity of those with whom they work, such as annotators, collaborators, and other stakeholders.

- Positionality statements are one method dataset authors might consider when incorporating reflexive thinking into their work (G. Rose, 1997). These statements have become accepted practice, not only in feminist scholarship where the concept proliferated, but increasingly in social computing (e.g. (Dym et al., 2019; Roldan et al., 2020; Simpson & Semaan, 2021)). Positionality statements include researcher reflections on how their own perspectives and values shape the work. For example, these statements can include how one’s disciplinary training lends expertise in certain ways of approaching a research problem, while acknowledging that training's own limitations in other approaches. They also can include how one’s nationality, race, gender, class, or other socio-historical identity impact the context and outcome of a project. Dataset authors should be careful not to incorporate an inherent distrust of the humans helping to shape datasets solely because they have an identity and a perspective. At the same time, I acknowledge that authors should not feel the need to disclose sensitive attributes about their identities that might otherwise endanger them.

- Reporting author and annotator demographics may be one useful tool for making transparent how decisions may have been made during the data process. There are numerous guides for collecting and reporting on identity data in ethical and sensitive ways (e.g., my work on gender guidelines (Scheuerman, Spiel, et al., 2019) for gender, (*Race Reporting Guide*, 2015) for race). Authors might even consider targeted recruitment of annotators with specific expertise or identity experiences who would be best suited to annotating the data. For instance, Patton et al. find that annotators who are familiar with gang-related activity have significantly different annotations of Twitter activity than those who are not (Patton et al., 2019).
- Writing ethical considerations is becoming increasingly required for machine learning venues. Authors are being tasked with actively engaging with the potential social and political implications of machine learning research. For example, NeurIPS began requiring "impact statements" in every submission to the conference in 2020; impact statements encourage authors to elucidate how their contribution might result in societal consequences, both positive and negative. Many of the publications I analyzed put forth some broad societal justification for the computer vision task they were contributing to but did little to engage with the real-world implications for that work. While several authors have put forth guidelines for writing impact statements (e.g., Ashurst et al., 2020), I also encourage authors to incorporate their values and ethical considerations early on in the dataset construction process, before data collection, and into every step of the pipeline thereafter. Similarly, authors might outline refusals—data which they refused to collect, uses they refuse to condone, or opportunities for allowing data subjects or annotators to opt out—in their ethical statements (Garcia et al., 2020).

Model Work over Data Work

Despite the fact that data is posited as crucial, many pieces of the data collection and curation process are missing from documentation, including, often, the data itself. Many datasets were not available at the time of writing, with URLs that go nowhere enshrined in archival papers. Reporting on the details of algorithms—or the model work—takes priority in the publications associated with these datasets. As discussed in Dataset Authors and their Disciplinary Practices, the vast majority of these datasets do not get published unless they report some kind of algorithmic improvement to a machine learning task. Publications that report solely on datasets are typically not published. If they are published without a corresponding model or technical development, they are typically relegated to a non-archival technical report, rather than published in a top-tier venue. For this matter, reporting and evaluation of the model work is what is typically incentivized, rather than the careful, slow data work. The findings here are in line with recent characterizations of machine learning dataset work as undervalued and deprioritized (Hutchinson et al., 2021; Sambasivan et al., 2021).

Dataset documentation is itself undervalued in computer vision. The fact that the majority of papers in the corpus were focused on detailing the design of algorithms showcased that, while data is necessary to model development, it is auxiliary to the proposed model or method itself in the publishing process. Moreover, maintenance and stability of the data itself is devalued, with much of the data unavailable only a few years after being published.

Even though most of the dataset creators report that they are interested in sharing the dataset as part of a larger research community, most are silent about the infrastructure necessary to maintain the upkeep of datasets. As noted above, while most of the datasets report an URL or have a find-able URL, many of the datasets do not have any kind of manner or mechanisms for data persistence. Practices such as providing each dataset with DOI, storing them at an institutional repository, or putting the dataset under version control are not followed or even mentioned within the majority of paper texts or on the websites of the datasets. As such, some datasets were no longer available, with sources in publications leading to expired links, making the data unavailable for use or scrutiny—even while it may be still in use by others who were lucky enough to download it when it was still available. The impermanency of datasets has several downstream effects, including increases in technical debt, lack of maintenance, and the inability to replicate and reproduce scientific results.

Information scholars note that data reuse itself is embedded in scientific practice and that *"investments in data sharing… may have long-term consequences for the policies and practices of science"* (Pasquetto et al., 2017). The amount of work that is put into data sharing has significant effects for the scientific work of communities of practice. Computer scientists have similarly written on the high cost of so-called "technical debt" in machine learning systems (Sculley et al., 2015), including "data dependencies", which cost more than dependencies in machine learning code. Moreover, it's more appealing to create a new dataset than to maintain its upkeep, especially as datasets are created as the demand for new and more novel models emerge with increasing frequency. However, this leaves old, publicly available datasets in the lurch, to be deprecated through neglect and disuse.

The crisis of reproducibility in psychology and other empirical fields has become more pitched as large-scale empirical work has been adopted throughout computational sciences (Open Science Collaboration, 2015). Stodden et al. found that for several issues of *Computational Physics*, *"Some artifacts [were] made available"* in only 5.6% of 306 cases. Moreover, *"computer code, input and output data, with some reasonable level of documentation*" was available for only 4% of the cases, and that only 18% of authors provided necessary artifacts for replication upon request (Stodden, Krafczyk, et al., 2018). Similarly, Stodden et al. found that, of 204 articles in the journal *Science*, only 24 had the requisite information to obtain requisite code and data. After requesting code and data from authors, only 26% of results in all articles were able to be replicated (Stodden, Seiler, et al., 2018).

One of the datasets which meets the bar for doing the requisite data work—in terms of data reuse, maintenance, and reproducibility—is the CAFE dataset. CAFE was the only dataset hosted at an institutional repository (NYU Databrary) in the sample. In addition, being deposited in the Databrary also assigns a DOI to the dataset,[22] which allows persistent access to the dataset.

Recommendations for Incorporating the Value of Data Work

I have several recommendations, in accordance with work on data curation and stewardship practices.

- Assigning datasets a stable DOI and storage location would increase transparency, usability, and reproducibility. This may be in the form of a DOI, which will allow for a permanent reference to the dataset which is not bound to a particular URL, and will therefore be easier to find or link to from scientific papers and other persistent artifacts. It may also mean storing datasets in institutional repositories—such as Databrary, Harvard Dataverse, or Zenodo—such that data can be maintained at a persistent third-party location, rather than on a lab or personal website. Many of these services allow for version control of data, which is an added benefit, because as data

[22] http://doi.org/10.17910/B7301K

changes and updates are made to datasets, new versions can be named and accessed, without erasing previous versions of the dataset. Scientific results can name a particular version of the data, rather than leaving would-be replicators to make assumptions about versioning.

- Create a data maintenance plan in order for data to remain relevant and useful. Dataset authors can maintain their data by ensuring it is still accessible. As Hutchinson et al. state: *"If you can't afford to maintain a dataset, and you also can't afford the risks of not maintaining it, then you can't afford to create it"* (Hutchinson et al., 2021)*.* Any updates should be clearly and transparently documented so that users understand how changes may impact their use. This would also allow researchers to understand that older versions of data may still be in use, and what those older versions look like.
- Publishing rigorous and detailed dataset documentation alongside the dataset makes datasets more trustworthy, transparent, and reproducible. Recognizing data work as a specialty area will increase incentives for documenting data more rigorously. Following Jo and Gebru, dataset development should be understood as a specialty area of computer vision (Jo & Gebru, 2020). It should be given intentional space in textbooks and curriculum, and publications focused exclusively on dataset developments should be recognized as meaningful contributions. I also suggest research tracks at top-tier computer science conferences that allow for data work to be incentivized as a publication in and of itself. On the level of machine learning more broadly, the inclusion of dataset documentation could be incentivized through journal policies and publication requirements. While publication policies alone are not sufficient to increase documentation sufficient for replication (Stodden, Seiler, et al., 2018), multiple frameworks exist for the reporting of data (e.g., Gebru et al., 2021; Gil et al., 2016; Holland et al., 2018). The open-source codebook for this study provides a framework for dataset curators to build and document datasets.

Conclusion

Technical artifacts are imbued with politics. Previous scholars have examined the underlying politics and values of a variety of artifacts, many of which center data practices—from how values are shaped by the production of data (Vertesi & Dourish, 2011) to how it then shapes our scientific practices more broadly (Bowker, 2006). Researchers are increasingly interrogating the values of machine learning data, specifically, including how gender is represented in datasets for facial analysis tasks (e.g., see Chapter 4), how racial bias in face data leads to biased outcomes (e.g., Buolamwini & Gebru, 2018), and how images in object recognition datasets skew towards Western countries (e.g., De Vries, 2012; Shankar et al., 2017). I built on this work by broadly examining what the documentation and reporting practices of computer vision datasets say about the values of the discipline. I analyzed what aspects of dataset development authors gave space in the publications accompanying their datasets and how authors documented and described different components of the dataset. I also analyzed what went unsaid in these publications, in order to better understand what dataset developers valued. I found that, broadly, computer vision dataset practices value *efficiency over care, universality over contextuality, impartiality over positionality,* and *model work over data work*. For each of the silences I identified, I recommended potential steps dataset authors could take to attend to them. I hope that this move—acknowledging the values implicated in data creation and annotation, and taking steps to develop a careful, contextual, position-aware data practice—can lead to more replicable, accountable, and ethical research in computer vision, and machine learning more broadly.

PART TWO:

DEVELOPMENT

The work presented in this section is focused on the role of human workers in developing computer vision artifacts, like datasets and models. Before delving into the studies presented in Chapters 7, 8, and 9, I describe the ethnographically informed methods underlying each of them in Chapter 6. Then, I present three studies:

1. In Chapter 7, I show how traditional tech workers approach identity in developing computer vision. I describe how traditional workers reference their own positionalities in developing computer vision. I show how traditional workers navigate contextual constraints within their companies and negotiate their own positional perspectives with their colleagues' positional perspectives. I present a model of how different contexts and different actors influence the positions traditional workers take in their work.
2. In Chapter 8, I show how data workers approach identity in collection and annotation work for computer vision. I describe how the positionalities data workers occupy influence their interpretations of identity categories when conducting data work. Beyond showcasing the role of data workers' positionalities in conducting data work, I discuss the failures of current bias mitigation approaches. I instead propose "positional (il)legibility"—attending to certain perspectives in data work as either legible or illegible to workers.
3. Finally, in Chapter 9, I examine the relationship between traditional worker positionalities and data worker positionalities in the process of development. I identify how these two types of workers have different levels of positional power in implementing identity in computer vision. I describe positional power as the ability to implement subjective decisions about identity during the development process. Traditional worker positionalities dominate the development process, giving them agency to negotiate decisions with their colleagues and make final decisions during each step of development. Meanwhile, data workers are expected to put their own positionalities aside and instead enact the positional perspectives of traditional workers. I provide potential alternatives to the current structure of power present in computer vision.

6
FIELD SITES AND METHODS

In Part One of this dissertation, I demonstrated how identity has historically been embedded in computer vision artifacts and how those artifacts implicitly reflect the values of their creators. In Part Two, I will now focus on the creators—the humans responsible for developing computer vision artifacts. To understand how human actors approach identity concepts in their work on computer vision, I conducted a multi-site ethnography focused on three field sites: (1) technology companies; (2) EnVision Data, a specific data outsourcing company; and (3) Upwork. This section provides the grounding for the work presented in Chapters 7, 8, and 9.

Historically, anthropologists have treated ethnography as bound to a singular physical space—the ethnographer would deeply engage with the physical environment that they are studying. I do not engage with only one field site, nor do I engage with their physical spaces. Marcus pioneered the idea of using multiple field sites to do ethnographic work, to build a richer picture by comparing cultures and practices (Marcus, 1995). This was especially transformative for research in organizational contexts, where my own research is situated. Similarly, Boellstorff demonstrated that ethnography need not be bound to physical spaces (Boellstorff, 2008). As our lives, including our work, move increasingly online and is increasingly distributed across the globe, situating ethnography in a digital context is reflective of the reality of work. The workers in my study worked with colleagues across the globe, sometimes never stepping foot in a physical office. Thus, I not only study multiple fields of computer vision development, I study them digitally, reflective of how the work in the computer vision space is done.

In my work, I employed observations of work, observations of communications between workers, document analyses, and interviews. Data collection for this study began in 2019 and concluded in 2022, a span of four years. The participants (see **Table 10**) in these chapters are largely divided into two groups:

data workers, who provide data services for computer vision, and traditional workers, who work full-time on computer vision products at technology companies.

Conducting ethnographically informed work allowed me to engage deeply with workers as they developed computer vision. Not only was I able to get workers' perceptions about their work through interviews, I was able to analyze the projects they were working on. I was able to compare approaches across multiple workers and projects at the same field site and across field sites. I was able to understand the broader context of the companies they were embedded in, the histories of the projects they were working on, and how their personal lives intersected with their work.

In the rest of this chapter, I detail my approach to the ethnographic work in Part Two of this dissertation. I begin by describing how my own positionality contributed to specific interactions and ways of viewing the field sites and participants I engaged with. I then describe the language I decided to use to describe workers and field sites. I then dedicate a large section of this chapter to describing each field site, including some of the methods and barriers I hit when working with them. I also describe the data projects I observed. Finally, I conclude with a description of the methods used in this chapter, describing how I approached interviews with workers and how I analyzed the data.

			Participants		
	Company (Size)	***Alias***	***Role***	***Country***	***Subregion***
DATA WORKERS	EnVision Data (Small)	Yasmin	Annotator, project supervisor	Bulgaria	Southern Europe
		Ghaliyah	Annotator, trainer	Bulgaria	Southern Europe
		Dinorah	Project supervisor	Bulgaria	Southern Europe
		Aakrama	Annotator	Bulgaria	Southern Europe
		Abyar	Annotator	Bulgaria	Southern Europe
		Wares	Annotator	Afghanistan	Central Asia
		Sumbul	Annotator	Afghanistan	Central Asia
		Shokouh	Annotator	Afghanistan	Central Asia
		Sadham	Annotator	Lebanon	Western Asia
		Raiha	Annotator	Lebanon	Western Asia
		Makaarim	Annotator	Lebanon	Western Asia
		Hijrat	Annotator	Lebanon	Western Asia
		Baksish	Annotator	Lebanon	Western Asia
		Azyan	Annotator	Lebanon	Western Asia
	Upwork (Medium)	Jaako	Collector (EnVision Data)	Kenya	East Africa
		Rebecca	Collector (EnVision Data)	Philippines	Southeast Asia
		Thanh	Collector (EnVision Data)	Vietnam	Southeast Asia
		Manjola	Collector (EnVision Data)	Albania	Southern Europe
		Lyonis	Annotator, collector, trainer	Uganda	East Africa
		Pelumi	Annotator, collector	Uganda	East Africa
		Malik	Annotator, collector, supervisor	United States	Northern America
		Sadhil	Annotator	India	South Asia
		Nedeljko	Annotator	Serbia	Eastern Europe
		Gemma	Annotator, collector	Kenya	East Africa
		Raines	Annotator	Russia	Eastern Europe
		Bernardita	Annotator, trainer	El Salvador	Central America
		Lucano	Annotator	Venezuela	South America
TRADITIONAL TECH WORKERS	Aqueous (Large)	Jeremy	Software engineer	United States	North America
		Coleman	Principal data scientist	United States	North America
		Kaleigh	Program manager	United States	North America
		Vasuda	Project manager	United States	North America
		Ethan	Senior principal research manager	United States	North America
		Callia	Principal research manager	United States	North America
	EnVision Data (Small)	Irina	CEO	Bulgaria	Southern Europe
		Zephyr	Chief impact officer	Bulgaria	Southern Europe
		Thalia	Chief operations manager	Bulgaria	Southern Europe
		Samuel	Chief commercial officer	Bulgaria	Southern Europe
	Maelstom (Large)	Jacqueline	Lead UX researcher	United States	Northern America
		Elliot	Research scientist	United States	Northern America
		Madison	Lead research scientist	United States	Northern America
	MultiplAI (Small)	Lynn	Head of data operations	United States	Northern America
		Kenny	Vice president of business development	United States	Northern America
	Resoom (Small)	Nicholas	Chief IO psychologist	United States	Northern America
		Lydia	Head of data science	United States	Northern America
	Exodia (Large)	Macy	UX researcher	United States	Northern America
	Inoculus (Medium)	Nitesh	Data engineer	United States	Northern America
	Phrenx (Small)	Kelly	Developer advocate	United States	Northern America
	SensEyes (Small)	Solange	AI product manager	France	Western Europe
	Sybil (Small)	Siddharta	Computer vision scientist	United States	Northern America
	Verus (Small)	Aishwarya	Computer vision research intern	United States	Northern America
	Zeta (Large)	Beiwen	Machine learning research intern	United States	Northern America

Table 10. 51 participants total. 27 participants were data workers; 24 participants were traditional workers. Small companies indicate 500 or fewer employees. Large companies indicate 10,000 or more employees. Medium companies had between 501 and 9,999 employees. Participant aliases were created using the same cultural origins as the participants' real names. Company aliases were randomly generated.

My Positional Relationship to Work in This Section

My positionality undeniably shaped my engagement with participants and field sites during this work. As I have previously described in positionality sections in previous chapters (see Chapter 4 and 5), feminist practices of reflexivity posit that research is mutually shaped by researcher and researched.

In the studies at hand, that is especially pertinent, as I was working with globally distributed field sites with very different approaches to work. From my own sociocultural and epistemic perspective (Giere, 2006), I interpret how participants express their identities, both implicitly and explicitly. Participants' expressions of their perspectives are shaped by the context of research—for example, by the language of interviews (English), the remote format and their location at the time of the interview (where family may have been present in the home), their perspective of the interviews as part of their work, and their perception of me as a person. In trying to understand how the positionalities of workers influence the outcomes of computer vision artifacts, it is crucial that I also actively engage with which aspects of my own positionality I believe shaped my engagement with participants and the interpretations I made about the role of their positionalities.

How my positionality influenced this work differed between traditional tech workers and data workers. When it came to traditional tech worker participants, the trust that I was able to build through my relationships and reputation with certain technology companies awarded me opportunities to interview individuals who would otherwise be inaccessible. In order to build relationships and reputation with people at large tech companies in the United States, I had significant positional advantages. Namely, being a United States citizen who speaks English and is obtaining a doctorate degree from a highly regarded department at a reputable institution aided me in gaining access to these companies. Beyond my prior publications, my title and affiliation helped to instill trust in my credentials as a researcher. Even while I have potentially differential experiences than my participants because I come from a lower middle-class background and am a first-generation college student, given the majority of my participants are based in the United States and have higher degrees, our experiences were accessible and legible to one another, making communications easier. As I have ascended both academic and class ladders, I have learned over the years how to negotiate and assimilate into spaces associated with higher class and academic lineages.

While my position as a U.S. based academic certainly helped me access traditional tech worker participants, it also potentially affected my relationships and perceptions of data worker participants. As a researcher born and based in the United States, I have a Western-centric perspective based on the culture, context, and communities I was raised in and continue to be part of. Given I am also white, I have had a very particular experience in the U.S. and as a researcher—one which has granted me great power and privilege. Most of my data worker participants were both non-U.S. and non-Western citizens; they were also primarily non-white. These dichotomies undoubtedly shaped my interactions with data worker participants, often in ways invisible to me. However, some interactions were visible to me. For example, because I am English speaking, all interviews were conducted in English, shifting power to me as a researcher and necessitating many of my participants communicate with me in a second language. This likely resulted in some miscommunications and loss of nuance and context that I might have been able to account for could I communicate with participants in their first languages. While I come from a lower-class economic background, given my position as a U.S. researcher and my experience with traditional white collar tech work, I was in a position of economic privilege when recruiting participants from the Global South who held lower paying jobs than myself.

Beyond building trust and rapport with participants, my positionality influenced my engagement with analysis. I acknowledge that my own identity as queer and trans-identifying not only motivated me to pursue this line of research on identity in AI, but shapes how I interpreted and engaged with gender and sexuality characteristics in this study. My position as a white, U.S.-based, English-speaking academic with access to resources awarded the privilege to engage in this work, while my experience as a sexual and gender minority and a lower-class first-generation college degree-holder awarded me a specific perspective in which to develop and conduct this research.

I am also committed to continuous (trans)feminist and anti-racist learning, and thus intentionally challenged how my own positionality might shape my assumptions. Even while trying to consistently consider my own power and privilege, I undoubtedly have gaps in my ability to understand certain perspectives due to the positionalities listed above. Engagement with the results and implications of this work should keep in mind the reflections above.

Language used in this Section

In conducting this work, I was highly conscious of the language I was using to describe workers and the contexts they were embedded in. After all, language is power (Dervin, 2015). That perspective is evident in the work I just presented in Chapters 3, 4, and 5. Given my previous work on computer vision artifacts (see Chapters 3, 4, and 5), I was aware there was a major division between the types of work that goes into it: model work and data work. Part of understanding how positionality shaped identity in computer vision was understanding the development of these two types of artifacts. Therefore, I sought to understand who was conducting data work and who was conducting model work. I carefully mulled over how to differentiate these two workers and settlers on “data workers” and “traditional tech workers.” I describe why I chose these two terms in the next section. I also describe the use of the terms “clients” and “customers” for clarity’s sake.

Further, I was highly conscious of how best to describe the cultural contexts of a globally situated workforce. I thus describe my use of U.N. geoscheme subregions, rather than more colloquial terms, at the end of this section.

Data Workers vs. Traditional Tech Workers

In this work, I distinguish between two types of workers: (1) traditional tech workers, who are employed directly by a tech company and are generally given benefits by that company; and (2) data workers, who are often contingent and contracted for short-term projects by tech companies. There are certainly many overlaps between these two groups of workers: they both work white collar desk jobs, they perform information work, and they both contribute core artifacts for computer vision. While these distinctions are not always clear cut, I have decided to refer to them as such in this work for clarity.

These two groups have historically been separated along class and geopolitical lines. Traditional tech workers are highly valued in many societies, and they are often viewed with high intellectual regard (Binder et al., 2016) and are paid high salaries (Liu, 2023; Miller, 2022). Traditional tech workers and their employers also drive computer vision products, in terms of vision, requirements, and development. They are also largely the targets of interventions or discussions in improving fairness in machine learning, broadly (e.g., Gebru et al., 2021; Holstein et al., 2019; Mitchell et al., 2018). Many tech companies are headquartered in the United States, or the Global North more generally (as reflected in **Table 10**).

Data workers, on the other hand, are consistently undervalued; their work is traditionally perceived as low skill (Musa, 2019) and is generally low paid (Perrigo, 2023; Ramnani, 2022; Yuan, 2018). They are otherwise referred to as "gig workers" and "ghost workers," given their short-term roles and invisibility to the development of AI. Data workers are hired to do tasks viewed as too burdensome, menial, and time-consuming for traditional tech workers to be doing, and may be framed as "mechanical" workers whose work can potentially be automated (Gray & Siddharth, 2019). They are given tasks to complete but are not expected to contribute to any creative direction to the projects they are hired to do. Computer vision researchers are often concerned with the bias data workers may introduce to a dataset, though they rarely view their own positionality as a source of such bias (see Chapters 5 and 9). In contrast to tech companies and their workers, many data workers are outsourced from the Global South for lower wages (M. Graham et al., 2017; Kak, 2020; Tubaro et al., 2020).

I included both data workers and traditional workers in this study because, despite their inequitable treatment and privileges, both contribute directly to the development of computer vision. More specifically, both types of workers imbue their own positionalities into their respective work when implementing identity characteristics, whether through tasks like project scoping or through labeling. Given both workers are central to the development of identity in computer vision and embed their own positional perspectives into their work, the purpose of including both perspectives is to understand how certain perspectives are privileged and thus impact the outcome of identity work in computer vision.

In discussing worker roles in this section, I also discuss another relevant actor: clients. Data workers often provide data services to traditional tech workers, who are using the data to directly develop their own models. On the other hand, traditional tech workers provided modeling services. I use the term “customer” to specifically refer to traditional workers’ clients. I use the term “clients” to refer broadly to anyone seeking data or modeling services. Though data workers also occasionally provide services to customers who are not building the models themselves but also outsourcing modeling work to traditional tech workers, I have chosen to represent the relationship between these three actors as linear for simplicity’s sake. The relationship between these can be seen in **Figure 6**.

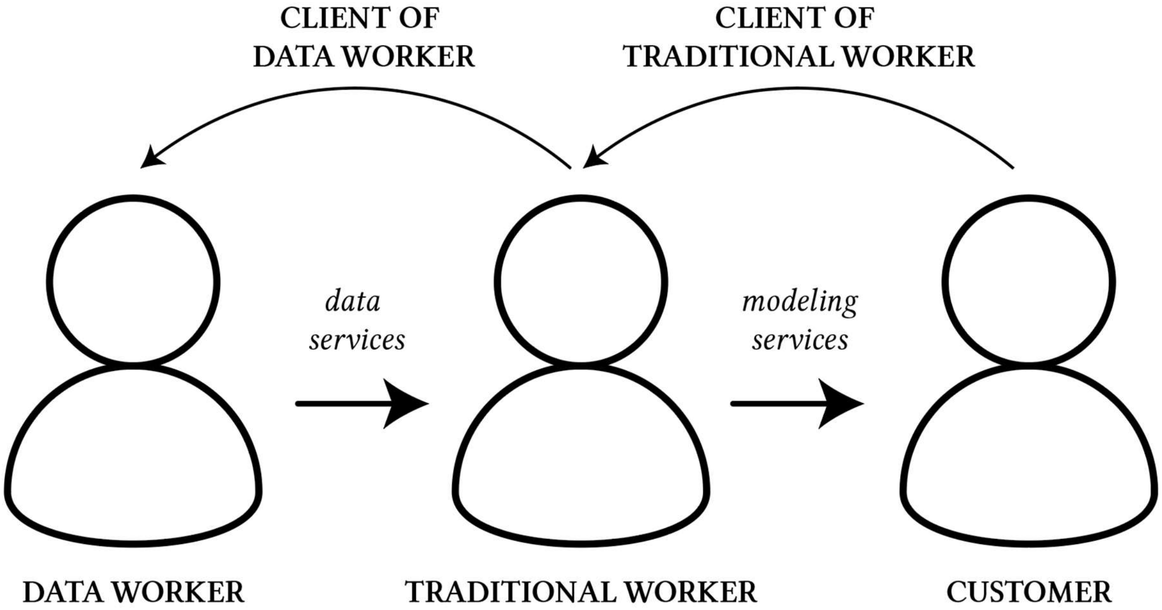


Figure 6. A diagram showcasing how different workers provided services to one another. Arrows point from the worker providing the service to the worker receiving the service. Data workers provided services to traditional workers; traditional workers provided model services to customers.

Geoscheme Regions

In describing the geographic locales of specific employees, I have chosen to use the United Nations geoscheme subregions. Subregions are more specific than broad regions such as "Europe'" or "Asia." I chose to use geoscheme subregions as an attempt to avoid Eurocentric and colonialist terminologies (Hanafi, 1998) and shifting geopolitical conditions present in colloquial terms like "Middle East." To respect the origins of each participant's true name, participant pseudonyms were intentionally chosen using the culture of origin of each participant's real name.

Field Sites and Recruitment

I conducted research at three different field sites. I conducted deep ethnographic observations with EnVision Data and interviews at all three field sites. The following sections provide details about each field site, methods used, and participants, but I will first briefly describe them:

1. The first field site was at different technology companies. I recruited traditional tech worker participants from a variety of technology companies, ranging from small companies to large companies.

2. The second field side was EnVision Data, a data outsourcing company where I conducted both interviews and ethnographic observations with their traditional tech workers (e.g., CEO) and their data workers.
3. The third and final field site was Upwork, a freelancing platform. I specifically interviewed freelance data workers who conducted work on Upwork.

Below, I describe each of these field sites, followed by a section describing the data projects I observed and engaged with during this project.

Technology Companies

I chose to focus on technology companies as my first field site given their central role in developing computer vision products which are actually deployed and impacting people in the world. As such, I sought to talk with traditional tech workers at companies ranging from small startup companies to large tech giants, like Google and Microsoft. Each type of company would have its own approaches, goals, and constraints. Small startups might be more agile but have less capital to work with. Meanwhile, large tech companies might be more bureaucratic, but have both brand name recognition and huge amounts of capital to work with. I wanted to understand computer vision development in a range of different contexts, particularly to understand how individual workers operated within these contexts.

I conducted interviews with 19 traditional workers employed at a variety of technology companies in various roles. Interviews were conducted slowly over the course of four years. Interviews first began in 2019, but recruitment was especially difficult, and the project was put on hold for much of 2020 and 2021. Recruitment methods and difficulties are further described in the Recruitment Methods and Barriers and Participation Concerns sections. Interviews were iteratively conducted as I gained access to tech industry spaces and began to build rapport and trust with members of certain companies. More direct recruitment of unfamiliar individuals was done in 2022.

Participants all worked on computer vision as part of their role; for many participants it was their primary focus, but others also worked on other technologies (e.g., NLP tools). I aimed to interview participants in both "technical" roles (focused on technical implementations, like software engineering and data science) and "non-technical" roles (focused on ideas, management, and research). While the boundaries between "technical" and "non-technical" are not clearly delineated and often overlap, the distinction helped me to focus recruitment efforts to be more balanced between role types. Interviewing participants in a variety of roles provided a more diverse perspective of identity implementation.

In the next section, I describe my recruitment approach. I also describe how difficult it was to recruit participants, particularly those located in larger tech companies.

Recruitment Methods and Barriers

Recruitment of traditional tech workers was difficult, particularly in comparison to recruitment of data workers. Participants were identified through an *ad hoc* sampling approach (Taherdoost, 2018). *Ad hoc* sampling was chosen for several reasons. First, because computer vision is a relatively narrow subfield of machine learning, identifying potential participants to recruit was difficult. I located computer vision companies largely through search tools on Google and LinkedIn, but identifying employees of those companies and whether they had direct engagement with computer vision products was opaque. In particular, those in more technical roles, like data scientists and software engineers, had less web presence than those in research or C-level roles. Second, those in technical roles were particularly difficult to identify and were even more unlikely to respond than those in non-technical roles. While I sought those in technical roles, for their specific expertise on implementation and testing, I was also open to any participants I could recruit from other roles. Further, those in other roles (e.g., research, business, project management, etc.) offered unique perspectives on identity implementation in computer vision. Finally, even after identifying employees as computer vision companies, many who responded did not work on human-centric computer vision. Many instead worked on products less central to my research questions, such as document analysis or robotic industrial applications.

Participants were recruited through a variety of mechanisms, some of which were more successful than others. One method for recruitment was directly contacting potential participants. I directly contacted potential participants in two ways: through email and through LinkedIn. Approximately 15% of email recruitments were successful. Approximately 17% of LinkedIn recruitments were successful, though I could not measure deleted connection requests. I recruited one participant through Twitter; two attempts at Twitter contact did not respond. I had tried recruitment by posting on computer vision Reddit boards and having industry friends post a call on Blind, neither of which were successful.

Beyond direct recruitment, I recruited via *snowball* sampling. In some cases, people (participants or otherwise) put me in contact with individuals they thought would fit the study. This was relatively successful, though there were still cases where individuals did not respond, even with their mutual connection facilitating. In other cases, I developed an insider relationship with individuals at some companies through research collaborations and consulting. In these cases, I had the opportunity to develop relationships with those working inside companies, and mutual connections trusted me more as a researcher due to my insider relationships with others at the company. Having access to a company email was also helpful in facilitating trust between myself and those I was trying to recruit. Insider relationships were the most successful means of recruitment and accounted for 42% of all participants.

My difficulties recruiting traditional tech workers highlighted that this group of research subjects is particularly inaccessible to researchers, highlighting to me the importance of understanding their perspectives. When I finally got to speak with traditional tech workers, I quickly uncovered numerous reasons they did not want to respond to me. I describe those in the next section.

Participation Concerns

Throughout the process of both recruitment and conducting interviews, I realized several aspects that made recruitment for this study so difficult. Two major participant concerns arose: (1) concerns about accidentally violating their own NDAs and (2) concerns about purposeful or accidental identity leaks. Though many of my recruitment emails went unanswered, some participants who had agreed to participate backed out before the interview due to legal concerns surrounding fresh controversy at their company. A participant who had initially declined to participate, but later participated after I had built a relationship with her, informed me that my initial recruitment emails had caused a great deal of “backchanneling" (secret conversations that did not involve me, the sender) about whether participation was too risky without having me sign an NDA. Much like similar studies involving industry stakeholders (Holstein et al., 2019; Veale et al., 2018), a number of participants expressed a distrust of researchers and speaking about AI due to fears that their personal and company identities would be leaked to the press. A distrust of journalists amidst a wave of articles covering AI ethics was salient among participants. Some participants also expressed concerns that the academic community was “reactionary” towards industry.

This highlights my position as not only an outsider to traditional tech workers, but as a threat. This shaped how I approached research with traditional tech workers. I became more upfront about my intentions in my recruitment, and often reassured them that I had no intention of publishing information to harm them or their companies. While this makes my research opaque in the sense that each company is anonymized, and thus insight regarding specific company practices is still hidden from the public eye, this helped me to build trust with my participants. It was also yet another factor that likely shaped my research; participants likely held back or altered the way they spoke to me out of concern.

EnVision Data

I have established and maintained a relationship with EnVision Data since September 2021. Data collection concluded in October 2022. EnVision Data is a company with a globally distributed workforce, based in Southern Europe. They provide outsourced data services to clients across the globe for computer vision projects. They provide data collection and annotation services, as well as output validation (e.g., verifying model performance) and edge case handling (e.g., 24/7 human annotation coverage to detect failures). Projects range from object classification, like assigning labels to clothing styles, to facial recognition aimed at identity verification. The data projects relevant to this study can be found in **Table 11**.

EnVision Data is unique in comparison to more traditional data BPOs in that it is also a social enterprise focused on providing remote work to at-risk populations. Specifically, EnVision Data's workforce is entirely comprised of individuals displaced by human conflict: refugees, asylum seekers, and individuals located in conflict-affected zones. They have public labor standards focused on fair pay given the country the worker is based in, intentionally attempting to mitigate unfair labor standards commonly experienced by data workers. Further, EnVision Data supports ethical AI initiatives and desires to contribute to more ethical AI systems; discussions with the CEO of EnVision Data revealed acknowledgment of past projects that might no longer be considered ethical, and which will appear in the findings of this study. Unlike traditional data solution companies, instead of relying on APIs to connect data requesters with workers (Gray & Siddharth, 2019). EnVision Data partners with non-governmental organizations (NGOs) to recruit, train, and manage workers and payments. Further, in opposition to micro-task platforms like MTurk and UHRS (Gray & Siddharth, 2019), which attempt to make invisible the humans behind annotation tasks, HITL makes visible and centers their human work force. They offer training in various areas, spotlight workers on their website, and release yearly reports on the status of their workforce. Given EnVision Data is unique in its mission as an ethical AI company with ethical work practices, they are not representative of data BPOs as a whole. Working with more traditional BPOs (e.g., Appen, Sama) might uncover different insights than working with EnVision Data.

As reported in their 2021 impact report, EnVision Data has over 480 active annotation workers. Workers are based in Bulgaria, Turkey, Syria, Lebanon, Afghanistan, and Iraq, with potential plans to expand to Yemen and Portugal. While the majority of EnVision Data's workforce is based in Central and Western Asia, the majority of their clients are in Western Europe (40%) and Northern America/Australia (25%). The remainder are located in other individual countries (30%), more so in Central and Western Asia (referred to as “the Middle East” by EnVision Data) than in Eastern and Southeastern Asia.

EnVision Data is also focused on providing data to companies and projects they deem beneficial to society, or at minimum, not harmful. Given the company's commitment to providing ethical data services and promoting ethical computer vision uses, the CEO, Irina, regularly attends conferences, workshops, and talks on fairness and ethics in AI. I developed a relationship with Irina when she attended a workshop that I presented on. She desires to continue improving her company and has provided access to her company as a field site to other researchers as well.

EnVision Data also employs Upworkers in cases where their main internally employed workforce cannot meet project requirements. Generally, hiring Upworkers occurs in cases of data collection, rather than annotation. Given data collection requirements are often focused on broadening geolocational diversity, EnVision Data hires Upworkers from specific locales dependent on client needs. I spoke with four freelance data workers that EnVision Data hired on Upwork for data collection projects (see **Table 10**). EnVision Data’s reliance on Upworkers is why I chose to also interview Upworkers as part of this research.

Observations with EnVision Data

I conducted observations of EnVision Data over the course of about a year. Due to the COVID pandemic, borders were still shut during the time of data collection and thus I conducted my observations entirely digitally, despite my desire to visit their physical office in Bulgaria. Digital observations were appropriate given almost all work at the company, with the exception of an in-person training in Portugal, was conducted remotely from 2021 to 2022. Some data workers described going into a physical office in areas where they suffered power outages, such as those located in Afghanistan. However, workers largely worked from home when power was not an issue.

As part of my observations, I was added to the company's Slack workspace. This allowed me to observe general communications between workers and to communicate directly with different people in the organization. I also observed project meetings, specifically the negotiations for the project Xavient, a diverse data collection project that EnVision Data decided to turn down due to its large scale. I was given virtual walkthroughs of the systems the company uses for project management and annotation, and demos of annotations by data workers during my one-on-one interviews with them. Beyond gathering and reviewing all of the company's public documentation, I was also given access to private documentation, such as contracts, client pitch decks, and annotation guidelines. I completed all of the private training modules that EnVision annotators are required to take to better understand how data workers are trained to do their jobs. I was able to compare the formal work environment and training of EnVision Data workers with the more informal and contingent work environments of freelance workers on Upwork. In the context of this study, observations provided contextual understanding of work conditions and expectations, client and project backgrounds, and the specific cultural situations of conflict-affected workers.

I kept a diary of field notes where I recorded observations and thoughts on meetings, demos, documentation, email exchanges, and any other relevant data that arose during my time with EnVision Data. Over the course of my fieldwork with EnVision Data, I conducted 22 semi-structured interviews (*M* = 54 minutes, *R* = 27-121 minutes). Interviews were conducted with EnVision Data data workers (Yasmin through Azyan on **Table 10**), freelance data workers hired by EnVision Data on Upwork (Jaako, Rebecca, Thahn, and Manjola), EnVision Data C-level employees (Zephyr, Thalia, and Samuel), and one of their client representatives (Solange). I did not conduct a singular interview with Irina but met with her 11 times over the course of a year, and regularly exchanged Slack messages. For that reason, I included data from Irina in meeting hours, rather than interview hours. Some interviews also included observations of annotation demonstrations or reviews of old projects. I recorded 398 hours of meetings. One client meeting I was not allowed to record, so I took 9 pages of detailed notes. I recorded 1184 minutes of audio in total. I also compiled approximately 27 pages of emails and over 60 documents, including datasets, reports, contracts, client decks, Upwork recruitment emails, and labeling instructions. I wrote approximately 70 pages of field notes.

Upwork

The final field site that I included in this research was Upwork. Given a great deal of data work is also conducted by freelancers—including some of EnVision Data's work—I also sought to understand the perspectives of freelance data workers. Freelance data workers often work on projects for short terms and are not formally employed by any specific company. As EnVision Data used Upwork to source freelance workers, I chose to recruit freelance participants on Upwork. As previously mentioned, Upwork only hired freelancers for collection projects. Those freelancers who worked on projects outside of EnVision Data were not limited to collection projects.

Freelance participants did not undergo training in the same manner as EnVision Data participants; they were often simply given the instructions for the project, and able to ask clarifying questions from a point of contact or other team members (if given access to the team via platforms like Slack or Discord). Unlike EnVision Data workers, who largely had no experience doing data work prior to joining EnVision Data, the majority of freelance workers had a variety of different data work experiences. Some data workers had worked for more traditional BPOs like Sama; others had worked for smaller data startups. One participant, Malik, is attempting to start his own data annotation company.

Also, unlike participants from EnVision Data, who worked in groups on specific projects with specific companies, freelance workers often worked individually for a number of clients across the globe. Much like **Table 11** of the clients served by EnVision Data, I present a visualization of which countries each freelance participant had clients in (see **Figure 7**). The geographic context of both worker and client is crucial to understanding how workers negotiate their positionality in their work. The figure also showcases which freelance workers are particularly experienced and have worked for a variety of clients across the globe.

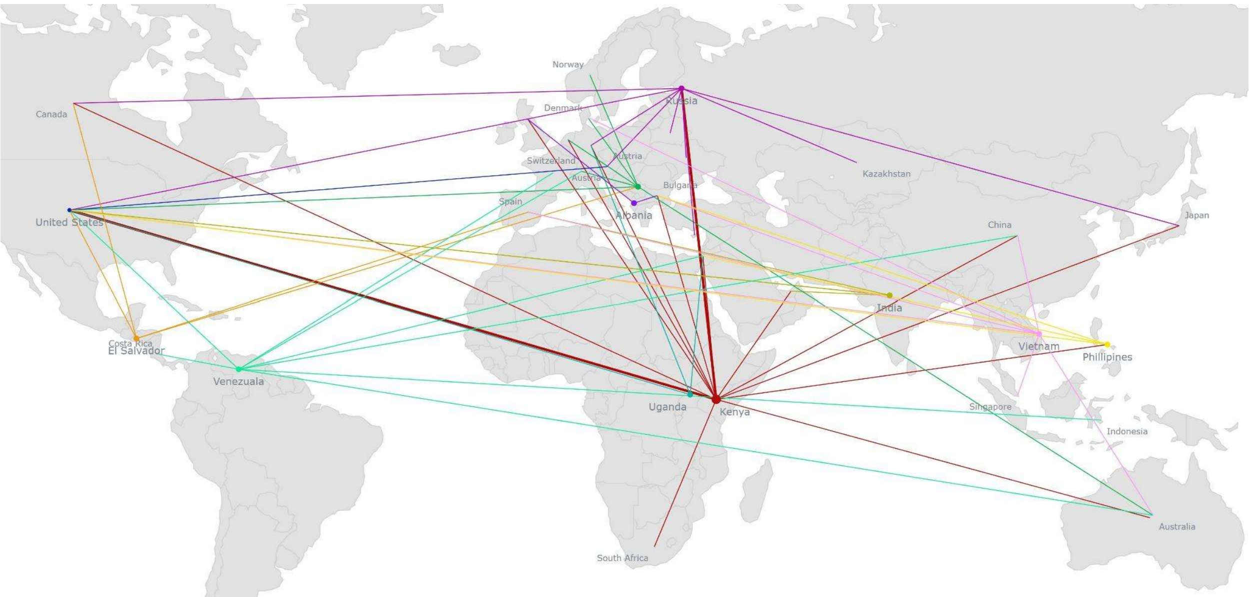


Figure 7. The figure above shows a world map. Dots represent areas where freelance data workers are located. Each line branches from a data worker's country to where their past clients have been located. The dot on Kenya is slightly larger because 2 data workers are located there; similarly, the line from Kenya to the US and from Kenya to Russia is thicker because both data workers had clients there. The map showcases not only how worker/client relationships span globally, but where the majority of participants' clients are located.

Recruitment of Upwork Participants

To recruit participants on Upwork, I posted job ads for an interview about data annotation and collection experience. The job ad requested applicants who had primarily worked on human-centric computer vision and had been hired by companies rather than researchers. The study's IRB approval documentation and consent form were attached to the ad. I posted two job ads: the first month I posted a general job ad aimed at anyone; the second month I posted a job ad aimed at participants in Northern America, Western Europe, and Eastern Asia (regions were defined in Upwork's platform using this language). On both job postings, I used Upwork's “invite” feature to search for relevant workers and invite them to apply to the job ad. I invited 16 people; 5 invites were ignored. I got a total of 48 proposals. I selected proposals based on the relevance of the workers' projects, the number of relevant projects, and their country of origin, in an attempt to get diverse perspectives from both a geographic and work experience perspective.

As can be seen in **Table 10**, I was successful in recruiting participants from East Africa, South(east) Asia, Southern and Eastern Europe, and Central and South America. However, recruiting participants from some geoscheme subregions was challenging. I was unsuccessful in recruiting participants from Europe and Eastern Asia, and only got one participant from Northern America. The participant from Northern America was also an immigrant from Western Asia (Jordan), and thus held a specific positionality as an Arabic-speaking immigrant. I had hoped to include further perspectives from those based in Northern America and Western Europe because there has been little research on labelers from those regions; primarily, Northern Americans have featured in research on "gig work" broadly (e.g., Katz & Krueger, 2019; Wilkins et al., 2022) and content moderation more specifically (e.g., Newton, 2019). I had also wanted to include Eastern Asian participants, particularly those located in China, because China has a massive AI industry and is a hub of data annotations (Beraja et al., 2020; Yuan, 2018). However, I could not locate workers who did data collection or annotation to invite from these subregions on Upwork. It is possible that workers from these regions are less likely to freelance data work and more likely to work for specific BPOs, like Appen China. While I explored options for including Chinese data workers by reaching out directly to Chinese data BPOs and joining Appen China forums, gatekeeping and language barriers were prohibitive. Inclusion of participants from these subregions remains prime for future work.

Data Projects

Clients, whether they hired data workers directly through Upwork or went through an intermediary business processing company like EnVision Data, generally provide documentation on the requirements and expectations of a project. Much like the data workers interviewed by Miceli and Posada (Miceli & Posada, 2022), the data workers in this ethnographic project worked on a variety of data tasks, including data generation, data annotation, algorithmic verification, and AI impersonation (i.e., real-time human-in-the-loop labeling and verification (Tubaro et al., 2020)).

I observed numerous data projects throughout my research. Data projects were primarily from EnVision Data, though some Upwork participants also walked me through their data projects and shared documentation with me. In this section, I describe the documentation used to guide identity-based data work for different types of projects. I discuss different projects in the following categories: data collection and data annotation. Given my in-depth year-long work with EnVision Data, I describe projects by name and provide deeper documentation (see **Table 11**). I give shorter descriptions of projects on Upwork, as freelance participants described numerous projects in briefer detail.

Envision Data Projects

Company Location	***Project Alias***	***Project Purpose***	***Data Workers Involved***
United Kingdom	ChAI	Video interviewing with personality insights	Yasmin, Ghaliyah, Aakrama, Dinorah
France	SensEyes	Face authentication	Jaako, Rebecca, Thanh, Manjola
Switzerland	Emovos	Emotion classification	Wares, Sumbul, Shokouh
United States	CaringHearts	Real-time patient monitoring	Yasmin, Aakrama, Abyar
Bulgaria	Codeguard	Labeling images as adult or not-adult for automatic moderation	Ghaliyah
Japan	Xavient	Collection of a highly diverse human dataset (project turned down)	N/A

Table 11. A table describing all of the EnVision Data projects I observed during my research. The remainder of EnVision Data participants (Sadham, Raiha, Makaarim, Hijrat, Baksish, Azyan) were interviewed regarding EnVision Data's ethical annotation training.

The purpose of this section is to provide contextual information about the projects discussed in these findings, especially to provide an understanding of what was communicated to data workers via the project guidelines. In particular, this section highlights that many projects require data workers to address identity characteristics (like race and gender) and identity-adjacent concepts (like clothing styles and emotions). Yet, project guidelines rarely attend to the identity aspects of this data work and provide no explicit instructions or examples for determining ground truth or for attending to potential biases. As a result of this lack of engagement with identity instructions, data workers are left to fill in the gaps on their own. I attend to how they go about making identity decisions in the Findings of Chapter 7.

Collection

Data generation involves the collection of images from either the “real world” (physical settings) or the web. In the case of human-centric computer vision projects, the data collected was largely images of people or images of concepts that held some form of cultural significance. For example, EnVision Data’s client SensEyes requested a dataset of diverse faces for their face authentication application to be used in mobile banking applications. The client requested that data workers collect selfie videos from people using a web application that they built (see X). The goal of this dataset was to ensure that their face authentication model worked on diverse groups of people. The client wanted the following identity-based distributions in the data: gender (men: 50% of images, women 50% of images) and “human group” (Caucasian: 20%, Asian 20%, African: 20%, Latin American: 20%, Middle East: 20%). EnVision Data worked with the client to determine how best to target the “human groups” in particular and decided to target specific regions through Upwork’s job posting affordances. No instructions were given to data workers on how to determine data subjects’ gender or ethnicity. Another collection project EnVision Data worked on was CodeGuard, a project aimed at labeling images as “explicit” or “safe” (as well as “underwear” as a middle ground category) for content moderation purposes (see **Figure 17**).

Participants on Upwork similarly described that, in the case of data collection projects, they were not given instructions on determining demographic or identity-based information. Some workers were asked to collect non-human images which held cultural or identity-based significance, as well, such as clothing types, food, or even infrastructure. Such non-human image requests often also provided insights into positional perspectives of data workers.

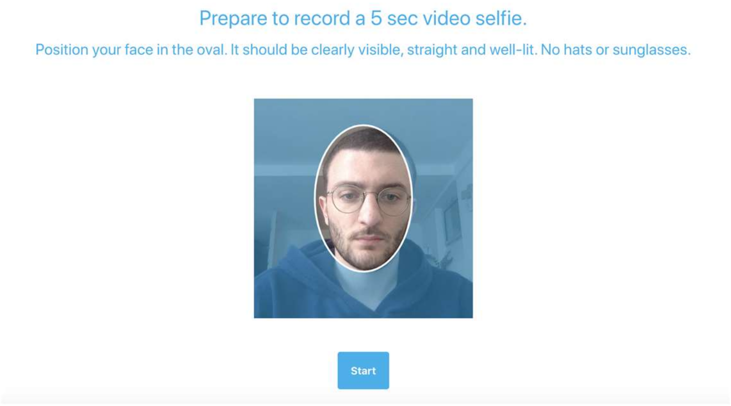
SensEyes Project Instructions

An excerpt of the job posting showcases how requirements were communicated to potential data collectors: “We are looking for people from African and South East Asian countries who can work as “video selfie collectors” for our project. The task of each collector would be to reach out to friends and family and to have 60 unique people record a 2-second video selfie on our app (Webcam or Mobile phones). When you complete 60 unique faces, you will be paid $30.
This project is for a startup that aims to detect identity fraud attempts, developing a software that detects fraudsters who would try to open bank accounts by showing somebody else’s picture for verification. Please note that we, and none of our representatives would ask for personal information like name, email or contact details of people in the video selfies. We simply

need as many and as diverse facial samples we could gather, in order to train our liveness detection solution which will be used to detect a spoof attempt (a fraudster trying to impersonate another person) by determining whether the source of a biometric sample is a live human being or a fake representation (photo, mask, etc.).

An excerpt from messages sent to each individual EnVision Data hired or invited to the job posting: “We'd love to consider you for this project. For the first stage we simply need you to answer these 2 questions:

1. Which country are you from and where are you currently living?
2. Can you reach out to a minimum of 50 people to collect video selfies? If YES, which of the following groups can you collect?
a. Caucasian
b. Asian
c. African
d. Latin American
e. Middle Eastern
Kindly note this information is only needed for the purpose of the project (e.g. regional/diversity).”



Prepare to record a 5 sec video selfie.

Position your face in the oval. It should be clearly visible, straight and well-lit. No hats or sunglasses.

Figure 8. SensEyes project instructions. Text includes a job posting describing the relevant human group categories. Screenshot shows the interface for recording selfies each worker used.

CodeGuard Project Instructions

Task “Extension of the NSFW classifier specificity” would require 350 to 500 representative images for each of the following 11 categories:

- very near close-ups of male and female genitalia and/or breasts/nipples
- otherwise clothed person but with fully or partially visible genitalia
- nude men/gay porn
- nude teen boy selfies (18+ y.o. so it's legal to store such files)
- nude teen girl selfies (18+ y.o. so it's legal to store such files)
- home pornographic content (for example such as the video stream preview images here https://chaturbate.com/)
- unsafe close ups of hands (for example involved into handjobs, male and female solo and/or mutual masturbation)
- vulgar gestures with hands or face/mouth/tongue

- artistic but fully nude images (such as

)

- tattooed bodies (sketchy, but not nude/pornographic)
- close ups of hands (safe images, no pornography)

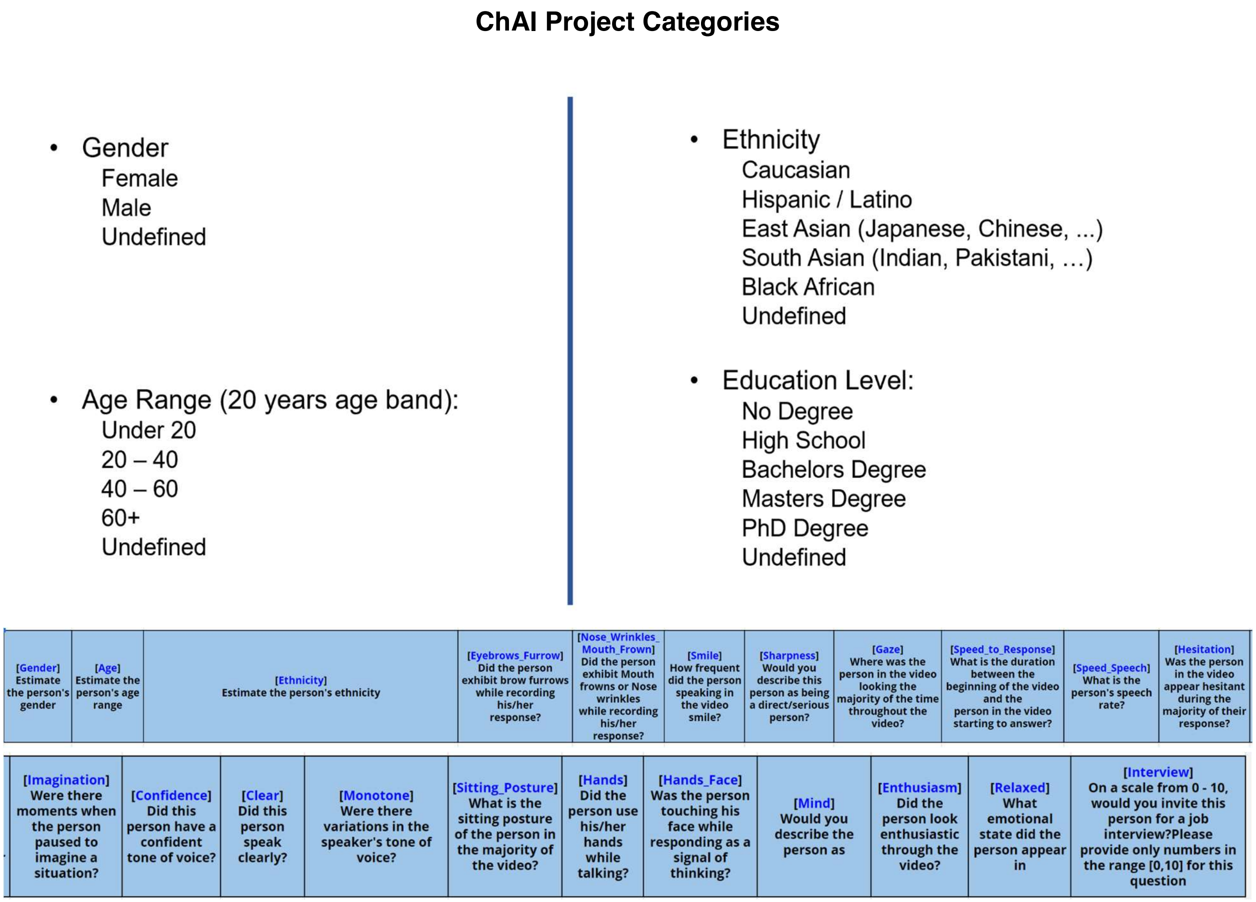
Figure 9. CodeGuard project instructions. Lists out categories to label NSFW.

As can be seen in the instructions for collecting data for both SensEyes and CodeGuard, labeling was also largely built into these products. For example, collecting images of “Caucasians” meant data collectors found images they believed were representative of the label “Caucasian.” Similarly, the data collected for “nude men/gay porn” was representative of that label. However, often data might also be labeled, or further labeled, after being collected. I describe the annotation projects that participants worked on in the next section.

Annotation

Annotation, or the labeling of images with concepts or bounding boxes, was done in the following areas in the context of this study: annotating demographics, annotating emotion, annotating cultural concepts (e.g., clothing types), and annotating the explicitness of an image. EnVision Data conducted three annotation projects: ChAI, Emovos, and Codeguard.

The ChAI project was focused on annotating “psychometrics” (emotion and personality characteristics) to train a digital interview platform. They provided EnVision Data with a dataset of interviews to annotate and a pitch deck with project requirements, which Dinorah then used as the basis for creating an annotation spreadsheet for each annotator to fill out. Annotators were asked to choose from a set list of options in the following categories: gender (male/female/unknown), age range (below 20, 20-30, 30-40, 40-50, 50-60, 60 and above, unknown), ethnicity (Caucasian, Hispanic/Latino, East Asian, South Asian, Black/African, unknown). They were also asked to answer questions about candidate expressions, such as: “Did the person exhibit brow furrows” (yes or no). Finally, they were asked to rate the candidates on a scale of 1-10. Each video had three annotators, and the final annotation was determined through majority rule (i.e., if 2 annotators chose “yes” and 1 chose “no,” the final annotation was “yes”). While examples were given to annotators on categories like “Did the person exhibit brow furrows” (see X), they did not provide examples for gender, age, or ethnicity.



ChAI Project Categories

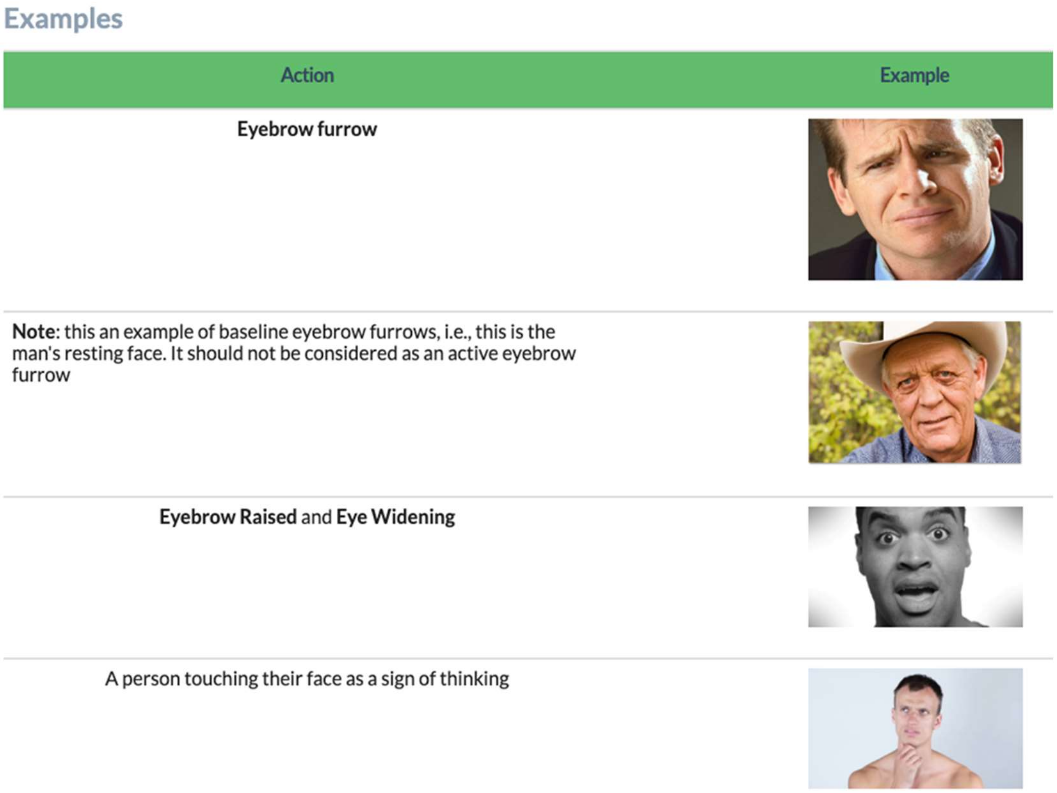
- Gender
  Female
  Male
  Undefined
- Age Range (20 years age band):
  Under 20
  20 – 40
  40 – 60
  60+
  Undefined
- Ethnicity
  Caucasian
  Hispanic / Latino
  East Asian (Japanese, Chinese, ...)
  South Asian (Indian, Pakistani, …)
  Black African
  Undefined
- Education Level:
  No Degree
  High School
  Bachelors Degree
  Masters Degree
  PhD Degree
  Undefined

[Gender] Estimate the person's gender	[Age] Estimate the person's age range	[Ethnicity] Estimate the person's ethnicity	[Eyebrows_Furrow] Did the person exhibit brow furrows while recording his/her response?	[Nose_Wrinkles_ Mouth_Frown] Did the person exhibit Mouth frowns or Nose wrinkles while recording his/her response?	[Smile] How frequent did the person speaking in the video smile?	[Sharpness] Would you describe this person as being a direct/serious person?	[Gaze] Where was the person in the video looking the majority of the time throughout the video?	[Speed_to_Response] What is the duration between the beginning of the video and the person in the video starting to answer?	[Speed_Speech] What is the person's speech rate?	[Hesitation] Was the person in the video appear hesitant during the majority of their response?

[Imagination] Were there moments when the person paused to imagine a situation?	[Confidence] Did this person have a confident tone of voice?	[Clear] Did this person speak clearly?	[Monotone] Were there variations in the speaker's tone of voice?	[Sitting_Posture] What is the sitting posture of the person in the majority of the video?	[Hands] Did the person use his/her hands while talking?	[Hands_Face] Was the person touching his face while responding as a signal of thinking?	[Mind] Would you describe the person as	[Enthusiasm] Did the person look enthusiastic through the video?	[Relaxed] What emotional state did the person appear in	[Interview] On a scale from 0 - 10, would you invite this person for a job interview?Please provide only numbers in the range [0,10] for this question

Figure 10. ChAI project categories. The top image contains different options for demographic selections. The bottom image contains a screenshot of many of the categories annotators were asked to label.

Emovos Project Examples



Examples

Action	Example
Eyebrow furrow	
Note: this an example of baseline eyebrow furrows, i.e., this is the man's resting face. It should not be considered as an active eyebrow furrow	
Eyebrow Raised and **Eye Widening**	
A person touching their face as a sign of thinking	

Figure 11. Emovos project examples. Examples were given to annotators to refer to when labeling emotions.

In a similar annotation project, client Emovos sought emotion classification for a computer vision advertising application. They provided a dataset of public figures and celebrities to be annotated with seven pre-determined classes of emotion: sad, happy, angry, disgust, surprised, fear and neutral.

Freelancers also described a number of annotation projects, including labeling identity categories like gender and race, labeling emotions, and classifying objects which held differential meanings across cultures, like clothing and infrastructure.

Methods

In this final section, I describe the interview methods I used when talking with both traditional tech worker and data worker participants at all three field sites. Specifically, I describe the design of my interview protocol. I then describe my approach to data analysis, which I conducted iteratively to slice my data in three ways (as presented in Chapters 7, 8, and 9).

Interview Design

To understand the role positionality plays in implementing identity in human-centric computer vision products, I conducted semi-structured interviews with both traditional tech workers and data workers. I chose to use interviews to gather rich descriptions of the perspectives, beliefs, and experiences of tech workers. Semi-structured interviews are a flexible methodology that also allowed opportunities to ask clarifying questions, seek out specific examples, and tailor interview questions in real-time to contextual responses (Irving Seidman, 2006; Qu & Dumay, 2011).

Interview questions were designed to elicit descriptions of participants' roles and the products they worked on, how identity characteristics (e.g., gender, race) were embedded into those products, how decisions shaped those characteristics, and constraints and difficulties in implementing those characteristics. Interviews with Macy and Jacqueline (see **Table 10**) were more informal; they were the first interviews conducted and were largely exploratory, towards understanding how to speak about identity with industry stakeholders. Interview protocols were then revised based on difficulties and confusions that arose during these early interviews.

Given identity is a nebulous concept that led to confusions for early participants, I decided instead to talk to participants about inputs and outputs involving human characteristics. I could then drill down into specific characteristics with participants. As the majority of participants worked on largely different projects, and many were in completely different companies, the interview protocol was designed to be flexible towards participants' individual roles and products (see APPENDIX). Since some participants worked at the same company and even on the same teams, I was able to build on questions based off of the context I'd gained from prior participants.

I conducted all interviews using video conferencing software. I both audio and video recorded all interviews, allowing me to capture both the audio interviews and annotation demonstrations. In some cases, I used text chat to communicate questions to participants who wanted to use translation services to better understand questions. Participants would then communicate answers to me back in English. Translation issues and language barriers were present realities during interviews. Chat logs were also saved in these cases, and cases where participants shared links to websites or images with me. All recording was done with participant consent. A limitation to conducting virtual interviews was when some participants' internet connections were slow or unstable. In these cases, interviews were generally interrupted until internet connections were restored.

Data Analysis

I adopted a constructivist approach to conducting and thus analyzing the data I collected (Moses, J. & Knutsen, 2019). I conducted a series of theoretical memoing practices informed by grounded theory (Charmaz, 2006). As data was collected across numerous months and with different participant populations, I took notes and conducted open coding as I continued to collect data. I conducted open coding to understand the range of themes present in all interviews. Documents (from EnVision Data) were not formally coded. Instead, notes were initially taken on documents, focused on piecing together a larger contextual picture of EnVision Data as an organization. Documents were then returned to analyzing interviews and to bolster and support the construction of thematic memos.

As I became increasingly familiar with the data, I began writing theoretical memos focused more acutely on how participants expressed their positionalities—the subjective experiences they described and how those seemed to inform their work. Participants rarely discussed concepts like "positionality," "perspective," or "subjectivity" explicitly. Rather, they described experiences, opinions, and culturally contextual characteristics relevant to how they interpreted project requirements. As themes coalesced, they were informed by observations, interview data, and my own positionality and experience as a researcher conducting a larger project on identity in the computer vision industry. I interpreted not only participants' positionalities, but how those positionalities had an impact on the artifact that the participant was working on. **Figure 12** visualizes how my own positionality acts as a lens through which I attempt to clarify participant positionality and its implications for identity in computer vision artifacts. I describe how I reflect on my own positionality in the context of this study in the next section.

Since participants did not necessarily explicitly discuss their positionalities or express reflexivity, understanding how participant positions influenced their work and practices of identity implementation were unearthed through my own lens as a researcher (see **Error! Reference source not found.**). I began clustering themes from participant-level memos into larger theoretical memos about how worker positionalities inform identity implementation. I supported these memos with multiple examples from across participants. As I refined these theoretical memos, they became the findings of each study.

Interpretive Analysis

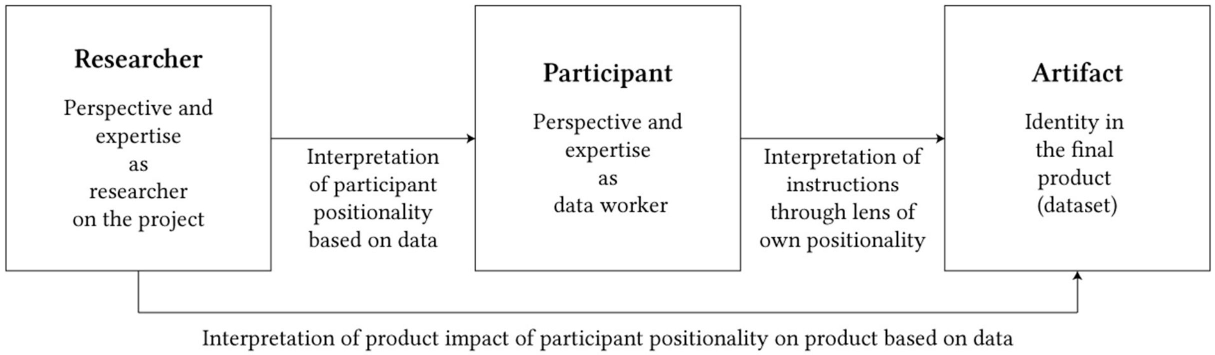

Researcher
Perspective and expertise as researcher on the project
Interpretation of participant positionality based on data
Participant
Perspective and expertise as data worker
Interpretation of instructions through lens of own positionality
Artifact
Identity in the final product (dataset)
Interpretation of product impact of participant positionality on product based on data

Figure 12. A visual representation of how my own positionality shaped my analysis of the data.

Figure 12 visualizes how my own positionality acts as a lens through which I attempt to clarify participant positionality and its implications for identity in computer vision artifacts. I describe how I reflect on my own positionality in the context of this study in the next section.

7
HOW TRADITIONAL WORKERS IMPLEMENT IDENTITY

I demonstrated in Chapter 6 that what the designers of computer vision artifacts value drives how the artifacts are designed. For example, valuing objectivity over contextuality necessitates specific approaches to computer vision, while ignoring other possibilities (see Chapter 5). Regardless of classification results, the value decisions about *how* computer vision should classify identity concepts result in issues of exclusion, erasure, and stereotyping (see colleagues and I's forthcoming work (Katzman et al., 2023)).

Issues surrounding identity concepts in computer vision come down to a perspective on technology design common in human-computer interaction: *computers are designed by people.* Given computer vision is, then, designed by people, identity categories are not simply neutral, and bias is not simply a mistake, but each are the result of the intentional decisions made by human actors. Increasingly, HCI scholars are exploring the way human actors influence the outcomes of computer vision artifacts (see Chapter 5 as well as (Denton et al., 2021; Hanley et al., 2021; Miceli et al., 2020)). Such scholarship highlights opportunities to better understand not only how computer vision is shaped by people, but how people's individual and subjective perspectives influence their decisions. Implicit in this body of work is the acknowledgment that people's *positionalities*—the identities they occupy in the world and how those identities shape their perspectives—influence how they approach designing identity categories.

Value decisions about identity lead to undesirable representations and outcomes for computer vision models. In this work, I explore how industry tech workers situated across a variety of roles, from engineering to research, are responsible for the design of enterprise-level computer vision systems. More specifically, I investigate how workers' positionalities—including the industrial contexts they are situated in, their own values, experiences, and perspectives, and their negotiations with their colleagues—

influence the design of identity concepts in computer vision technologies. In this chapter, I address answers to the following research questions:

1. How do individual workers' positionalities impact the development of identity in computer vision products?
2. How do individual workers negotiate their own positionalities with the positionalities of their colleagues and within the context of their organizations?
3. What failures occur when worker positionalities fail to account for other lived experiences?

In answering these questions, I was informed by the broader data collected in this project. However, in this chapter, I specifically focus on the semi-structured interviews specifically conducted with twenty-four industry practitioners who work on computer vision products (see Technology Companies in Chapter 6). Participants worked at companies ranging from small startups to big tech; they worked in a variety of roles, such as data science, engineering, research, business, and project management. Further, they worked on various types of computer vision products, from video-based interviewing to facial demographics to gesture recognition. Interviews were designed to elicit descriptions of participants' company environments, their relationships with their fellow workers, and their own personal experiences and values.

Findings showcase how the positionalities that workers inhabit influence the way that computer vision artifacts are designed. Workers seek to impact product design given their own positional perspectives about identity, while also being constrained by their fellow workers' differing perspectives and broader company-level contexts like regulation and company vision. Further, findings illustrate the types of positional gaps that arise when workers fail to account for different experiences and perspectives—like the cultural context of the data workers they employ and negative press around identity-based harms post product deployment.

I discuss how worker positionalities are relational, rather than individualistic; they operate within larger contexts in which workers are embedded and their relationships with other actors within those contexts. I conclude with implications for attending to positionality in tech work, at a higher-level contextual level and at a lower-level actor level.

Participants and Analysis

In this chapter, I focused on the 24 participants in industry contexts who worked on human-centric computer vision products (see **Table 12**). Participants worked on computer vision either as their primary responsibility or as a major part of their job (e.g., some participants also worked on natural language processing tools). Participants held a variety of roles at differing levels of seniority (from intern to C-level). Interviewing participants in a wide variety of roles meant obtaining more diverse perspectives around the problem of identity implementation in computer vision.

My analysis of the data focused on how traditional tech workers expressed their positionalities. I thematically coded data for instances where workers described their own identities, how they reasoned through their work, and the outside influences impacting their approaches.

Traditional Worker Participants				
Alias	***Role***	***Company***	***Company Size***	***Location***
Jeremy	Software engineer	Aqueous	Large	United States
Coleman	Principal data scientist	Aqueous	Large	United States
Kaleigh	Program manager	Aqueous	Large	United States
Vasudha	Project manager	Aqueous	Large	United States
Ethan	Senior principal research manager	Aqueous	Large	United States
Callia	Principal research manager	Aqueous	Large	United States
Jacqueline	Lead UX researcher	Maelstrom	Large	United States
Elliot	Research scientist	Maelstrom	Large	United States
Madison	Lead research scientist	Maelstrom	Large	United States
Macy	UX researcher	Exodia	Large	United States
Beiwen	Machine learning research intern	Zeta	Large	United States
Nitesh	Data engineer	Inoculus	Medium	United States
Irina	CEO	EnVision Data	Small	Bulgaria
Zephyr	Chief impact officer	EnVision Data	Small	Bulgaria
Thalia	Chief operations manager	EnVision Data	Small	Bulgaria
Samuel	Chief commercial officer	EnVision Data	Small	Bulgaria
Lynn	Head of data operations	MultiplAI	Small	United States
Kenny	Vice president of business development	MultiplAI	Small	United States
Nicholas	Chief IO psychologist	Resoom	Small	United States
Lydia	Head of data science	Resoom	Small	United States
Kelly	Developer advocate	Phrenx	Small	United States
Solange	AI product manager	SensEyes	Small	France
Siddhartha	Computer vision scientist	Sybil	Small	United States
Aishwarya	Computer vision research intern	Verus	Small	United States

Table 12. A subset of ***Table 10***, this table lists the 24 traditional tech worker participants in this chapter. The table is first organized by company size. Small companies have 500 or fewer employees. Large companies have 10,000 or more employees. Medium has between 501 and 9,999 employees. It is then organized by the number of participants per company. It is lastly organized alphabetically by company alias. Participant aliases were created using the same cultural origins as the participants' real names. Company aliases were randomly generated.

Next, I present Findings on how traditional tech workers went about conducting identity work in computer vision. Specifically, I attend to how traditional tech worker positionalities influenced their work practices.

Findings

Positionality manifested as highly relational and negotiable; both traditional workers and data workers approached identity work from their own positional perspective, but regularly interfaced with many other actors with different positional perspectives. Influencing each participant's vantage point is an intersection of personal, social, and political contexts. Further, these contexts shift over time, as products are created, deployed, and later changed. In these Findings, I showcase how positionality (often implicitly) manifests through relational interactions with other actors—other workers, clients, data workers, competitors, academics, journalists—across product lifecycle stages.

I present the Findings below as follows. First, I introduce how identity is being defined in industrial contexts, including the challenges practitioners face in defining it. Next, I discuss how the company context shapes how workers can approach identity in their work. I then discuss the ways that worker positionality influences approaches to computer vision work. I detail the level of personal interest participants expressed about working with identity characteristics, explicitly. I also describe how workers lament the approaches of their colleagues that they personally disagree with. Finally, I show how individual workers must negotiate their own positional perspectives with their colleagues during development. In the final section, I describe how gaps in positionality arise due to workers having their own limited viewpoints during the development process. I describe unforeseen and undesirable outcomes that become embedded into products as a result of these limitations. I also describe how workers desire more diversity, particularly from colleagues with marginalized identities, to avoid these undesirable outcomes.

How Identity Is Defined in Industrial Contexts

As Siddhartha said: *"Identity is very important, right?"* Identity is crucial to computer vision. Before building a computer vision product, workers define what identity should look like and how it should be scoped, in terms of its categories and data representations. For example, a gender classification model often uses the categories "male" and "female" for gender, and data representations include face images annotated with those categories. Identity was defined in a multitude of ways in the products participants worked on.

Some participants worked on computer vision products that explicitly classified human characteristics (such as Nicholas, Lydia, Lynn, Coleman, Kenny, Kaleigh, Vasudha, Ethan, and Callia). Other participants worked on products which more implicitly required identity classifications in the data, often for testing and evaluation (such as Jeremy, Elliot, Madison, Kaleigh, Vasudha). Finally, many participants worked on products which classified non-human objects which were still imbued with sociocultural meaning (such as Coleman, Kaleigh, Vasudha, Ethan, Callia). For example, clothing items, food, or immaterial concepts like "racy" for content moderation.

The process of defining identity varied depending on the company participants worked for and the products they were working on. As Elliot described, how identity is defined in computer vision is dependent on the goal of the product: *"When you're trying to incorporate information into a model during training, as opposed to just like, sort of understanding generally patterns of like, does this work for people, then you end up needing to do much more rigid things."* The "rigid things" Elliot is referring to are the categories workers define for computer vision products—such as gender categories for demographic classifications. Elliot expresses that in order for supervised machine learning products to work, the categories must be made into something rigid. Something like gender must be turned into discrete categories to be classified. Even products with more vast classification schemas, like object labeling models, still require a set number of categories—like 100 different animals, for example.

In larger company contexts, participants generally were unsure where most products originated from—who exactly began the work on the project and why. For example, Kaleigh in describing where the identity categories in Aqueous' computer vision models could only assume that the use of public datasets, like ImageNet, were the original source material: *"So to my knowledge, the public datasets that went into the creation of the models, the captions were primarily used as is."* She explained that there was likely cleaning and modification, but whatever reasoning the original authors of the public datasets employed in selecting and defining identity categories was unknown.

Though "identity is very important," workers often struggled with how to best scope identity, given the vast possibilities for categorizing identity characteristics. For example, Lynn, in heading the data operations team at MultiplAI, explained how "*intimidating*" it was to determine how to measure racial categories for bias testing in the company's computer vision models:

> *"Well, the first thing that I needed to do, which was a terribly, terribly intimidating prospect, was to sit down and think about how we are even going to test for [underrepresentation]. The kind of resounding sentiment from the team was like, we have to constrain this problem because it's an impossibility criterion if we just allow ourselves to think about every single phenotype and every single appearance of human facial features … So leaving those out entirely, was just based on skin color, I guess, right? And so, you think about, like, where is there maybe a standardized thing that I can steal from and then like, well, there's the US Census, which is highly problematic."*

Lynn's difficulty shows how workers struggle to decide on how to best represent identity for core product needs, like testing, and often make decisions from necessity and use resources that may be historically flawed (e.g., the US Census). Implicitly, Lynn's difficulty with how to represent identity insinuates that representing identity is a given. It was not questioned whether identity attributes need to be included, because they are seen as necessary. In describing the gender combinations, they decided to use for their wedding classification model, Lynn summarized: *"We would have to use our understanding, intuition, best judgment."* In practice, defining identity occurred through client meetings, project meetings with colleagues, and in defining guidelines for data workers to use to collect and annotate data.

In some cases, certain representations were much more difficult to get data for. For example, Jeremy worked on gesture recognition and described how difficult it is to get data on hands which are missing digits. Certain types of identities, like disability, were often seen as untenable, because making them into "rigid things" was much more difficult due to the vast diversity of disabilities. Not only was it difficult to get this data, what fingers to account for ballooned the problem exponentially. Jeremy explained that accounting for the spectrum of human diversity is difficult to imagine: *"This is always something that happens because when you design the dataset, you can't anticipate every type of failure, you try to vary it up as much as possible."*

Lydia was aware that many identities cannot be measured, but machine learning can be used to identify proxies of those identities. For example, she mentioned the infamous Stanford Gaydar experiment*. "I would be interested to understand more of where the data was coming from, because a lot of it might have been other cues in the photo. It's not the person's face. It's like style, maybe."* Elliot similarly described that computer vision connects presentation proxies to defined identity categories: *"The system isn't picking up on if it's a man or a woman, the system is probably picking up on like, what's the length of the hair, where is this person? …There's like, not one way in which a particular gender can* *look."* Therefore, when categorizing identity or assessing bias, computer vision systems are attempting to identify the cues which connect to gender or race categories. Lynn, on the other hand, seemed to describe the lack of non-binary genders in her company's model as a tradeoff:

> *"[We] made a conscious choice not to include things like androgynous or non-binary. I really had nothing to do with not feeling that those groups should be represented. It had more along to do with, you know, if we're training certain features that have an appearance, what your biology says about you, under your appearance, the model is really only prescribing to you what it thinks you look like, based off of the traits it understands to be masculine or feminine."*

Lynn's explanation also reveals her underlying perspective that gender is always inherent on the face in a specific binary way. Her perspective clashed with those of other participants, who viewed binary gender as not "biological" but social. Madison explained how her team created guidelines for annotating gender in more accurate ways, reflecting that gender classification is a visual interpretation made by others*. "I've been calling [self-annotated gender] first-party gender, and … then adding onto that for third-party gender, when it's more about what's being perceived by someone else or by a system."* She explained that she used outside resources from academia to inform these guidelines.

Scoping identity for a computer vision product was a complex and tangled process—it was generally not obvious or known to participants the exact origin of the products they worked on. At the end of the day, we see that identity is defined as a result of the interplay between worker positionality and product requirements, where positionality is in tension with the overriding constraints of product needs. This tension always occurs within the broader context of development, which includes differing personal values amongst workers, economic constraints, client demands, and regulatory frameworks. In the remainder of the Findings section, I detail how worker positionality clashes or aligns with the factors entangled in the overall development context.

How the Industrial Context Influences Worker Positionality

To better understand how workers embed tacit knowledge informed by their positionality into computer vision products, it is necessary to understand the organizational context in which workers are situated. Participants described three characteristics of the organizational context they were in: economic constraints, regulation and policy, and company values.

Economic constraints heavily influenced how workers were able to approach identity in computer vision development. Workers were enabled or constrained to approach identity in specific ways depending on the company they were working in. In particular, workers discussed how the economic power of their company shaped what the focus of their work was. Many felt that lower economic resources at their companies hindered their ability to conduct more in depth or expansive identity work. The workers discussing economic factors were those in small companies where money was a constraint, whereas those in larger companies did not discuss economic costs or disadvantages. For example, Kenny, the Vice President of Business at MultiplAI, a small computer vision startup, compared his company's ability to invest in new data with that of Google:

> *"So the problem is that, like Google, for example, when they're building their vision platform, they have a ridiculous amount of money … They had a team of seven individuals use their platform and labeled data for like a year, they pay these people over $100,000 each. We can't do that. Like we cannot do that."*

Kenny attributes the ability to engage more deeply with data collection, annotation, and research to economic power. Lynn, who worked at the same company as Kenny, expressed similar reflections about company size and economic power:

> *"It's only 100 people. So when you have a startup that's that young, you know, you have maybe a year's worth of cash in the bank, or however much it is, it could be a year, it could be six months, it could be 10 years, you sort of think differently about business as time goes on."*

In addition, Kenny explains that because MultiplAI is so small, they take a "business first" approach to building products:

> *"I don't think we think of use cases. I don't think that's how startups work unfortunately. In the position that we are in at a … some 100-person company, we take what we have, we repeat it, and then within those environments, if people ask us to create new applications for them, we're absolutely doing them. But we don't have a huge research or development team that can go out and just build new AI … The reality is that we build products based out of market need, right? … And in the early stages of the company's infancy, there was a significant amount of inbound demand for gender identification."*

Client demand is the most important factor when designing computer vision. Kenny explains that clients' visions of what identity should be in the product—like gender classification for marketing purposes—drives how identity is then designed. Nicholas, working at a similarly small company as Kenny and Lynn, weighed the same tradeoffs between business and academic research. He said that *“we're a business [so] we need to be profitable, so there's a fine line between [profit and research].”*

Lynn further reflects on how the startup environment made it difficult to more deeply consider the impacts of identity classifications on different potential user groups:

> *“The strains of time and money were very impactful. It was a real challenge to think about, just feel the consequences of one group or another … This comes back to it being his company without any regulation at all, guiding us. And without any real kind of oversight from like, you know, we're not Microsoft, where Microsoft has a team of probably 100 compliance officers and ethicists [who are] responsible, they have processes, any new technology is going to have to run through them first for governance. But we weren't like that. We were, you know, the, the number of people working on [the demographic classification model] was … like six people … And so, we just tried to use our best judgment.”*

Lynn described how there were no formal processes or regulations governing how to approach demographic information in MultlplAI’s classification model. As such, she and her five other colleagues relied on intuition about the best way to approach identity. This included discussing how to avoid major PR issues. She described how they were familiar with Google’s *“gorilla thing”* and they wanted to avoid those types of outcomes. Lynn also discusses how larger companies have the ability to hire policy professionals and ethicists, taking the burden off of development teams to make all of the decisions about what is or is not ethical. She also expressed that startup environments come with less formal regulation than large companies, making it more likely that they will approach identity work from a *“best judgment”* standpoint.

While Kenny insinuated that MultiplAI would take on most model building projects for the sake of income, other small companies did not always have the ability to take on clients which would help their bottom line. This constraint seemed especially true for data companies, rather than model companies. For small social enterprise companies like EnVision Data, resource constraints meant being unable to take on projects they did not have the workforce to accomplish. One big name client came to them with the goal of creating the most robust diverse dataset of human faces on the market. However, it was not possible given EnVision Data’s small size. They did not have the internal workforce necessary to create such a diverse dataset. Further, EnVision Data’s main goal is to employ data workers from conflict-affected countries. Given this project required extensive data collection, it did not directly serve their main workforce, since it would largely require hiring freelancers from around the world. Beyond purely economic constraints, some projects did not align with company values.

Kaleigh, who worked for a large company with vast economic resources, did highlight that economic resources do not rid them of all challenges:

> *"It is absolutely spot on that, yes, big companies have a lot more resources. But we still need and welcome help with more practices about how to do this well, how to measure things well, how to mitigate well, how to do participatory design well, because I think a lot of that work could directly translate into product changes."*

Kaleigh described that, even with economic resources, her company and her team did not necessarily know the best way to approach identity. Such approaches were still social enterprises, with humans making informed decisions guided by best practices. Beyond needing more guidance, Kaleigh also described how, even in a large company, teams had to request budget allocations:

> *"So, within my team, because we've been able to spend the last year making some progress on fairness and responsible AI initiatives and showing the importance of it … we put in a budget ask that was double our previous data budget asked. Because that's what it's going to cost to actually do fairness, at least doubling data budgets, if not more than that. And that doesn't even include all of the extra headcount that's needed to be able to manage data collections in a much tighter way."*

In order to approach identity in computer vision from a responsible standpoint, Kaleigh expressed the need for more money and more team members. Yet requesting an increased budget also comes with the burden of proving the money was necessary. She had to spend the past year documenting the progress her team was making on fairness and responsible AI initiatives. She also had to showcase what made these initiatives important. Even though her company had vast amounts of economic power in comparison to MultiplAI, it did not mean focusing on identity was a company priority.

To summarize, economic constraints were viewed as playing a major role in how identity was defined in computer vision. Workers at smaller companies with less economic power felt that they could not approach identity in more robust ways that they valued. They perceived larger tech companies, like Google and Microsoft, as having a great deal more freedom given their economic resources—and they envied them for it. However, workers at large tech companies also acknowledged that there were still major challenges to implementing identity features. While they had more resources, they also had to prove the need for those resources.

Regulation and policy also influenced the context in which workers approached identity categories. Regulation occurred in the strict legal sense and also through localized company policy. Lydia, the Head of Data Science at a small company focused on providing AI interview tools, explained that the company is bound to fair employment laws in the United States. Given that the company's approach to identity is built on the concept of fair hiring practices, the categories used are derived from a legal perspective. Lydia explained their company's approach:

> *"We try to predict age, race, gender, and we see how well we can predict it. So, we're saying, is there anything in this data that is telling us the gender of the person? And if there is, we want to take it out. We just wanna remove it … We just say, okay, we have to follow the guidelines that the EEOC sets for us, which is just for all job assessments."*

Lydia's company is subject to strict federal regulations, because the product they provide impacts hiring and could potentially lead to illegal unfair hiring practices. On the other hand, many companies also have their own internal policies. While not formally governed by legal requirements, company policies are still often binding. Madison described the ethical AI policies her team is expected to follow:

> *"And for each of those [ethical AI policies], it will have, essentially, our values as a company. And so, when you're doing ethical deliberation, basically what you do is you say, for each of my values, how am I meeting that value? How am I straying from that value? And what are the benefits and the risks in light of that value? … You can use that in doing sort of decision making around what should and shouldn't be released."*

Madison is describing how these overarching principles act as a policy. Workers at her company are expected to align their decisions with these policies. Her company's policies empower workers to push back on products that do not align with the company's ethical AI principals, because they have been formalized in ways that company values often are not.

While regulation was not commonly discussed by participants, Lydia and Madison's descriptions show how regulation—whether through legal frameworks or company policy—also plays a role in how workers approach identity categories for certain products. Some workers might even fall back on regulatory guidelines for demographic categories rather than expanding them. For example, Lydia's company only attends to EEOC frameworks for gender through the lens of male and female, because the law does not currently protect non-binary people in hiring. Therefore, workers like Lydia did not attend to expanding gender categories beyond the binary. While regulation and policy might maintain certain legal and ethical standards for identity that could go otherwise overlooked, it also risked becoming a constraint which narrowed worker thinking to more rigid categories.

Company values also influenced worker perspectives on identity. Workers expressed instances where their values seemed to align or differ with their view of their company's values. Some participants seemed to express an alignment and adoption of company values. For example, Nicholas told the underlying story of why the CEO founded his company. He said the CEO had been rejected by a large banking company because he did not go to a prestigious school that the company recruited from, which led him to create his own business focused on giving people a more meritocratic chance at landing a job. Nicholas said:

> *"So, the premise of [company] if you think about it was everybody should have a fair shot for a job that they're qualified for. So, if you're qualified for a position or a job, uhm, you should have a fair shot for it, irregardless of all the EEO [Equal Employment Opportunity] categories, like race, gender, age, but anything else too. Uhm, it doesn't matter who you are, as long as you're qualified you should have a shot."*

The story Nicholas told about the founding of the company and the CEO's vision communicated an alignment and identification with the goals of the company, which provides computer vision software for video interviews. The software is trained to analyze facial expressions and tonality and link those analyses to certain characteristics their clients are looking for when hiring, such as personality characteristics. Demographic classification is not an explicit feature of the software but is used to measure and mitigate bias in the system. Nicholas' alignment with the company's values informs his own approach and view on classifying identity in computer vision—that it is positive because it supports the company vision of a fair job interview. *"Along with the same guidelines and principles that we have as a company, everybody should have a fair shot for a job that they're qualified for … we still live by that single value, everybody should have a fair shot."* Lydia, who worked at the same company, similarly expressed support of the underlying vision of their product. She said: *"I really do believe that we're what we're doing is a huge improvement upon something that's really not ideal [hiring]."*

Siddharta, who worked at a different company that also worked on affective classifications, Sybil, also expressed joining his company because he felt that the vision of the company was "ethical." *"You know, it is one of the reasons I joined Sybil, and not the other places I had offers from, because so* *ethically, yeah, yeah, I would say the management, and everyone is pretty strong on what they want to do."* Siddhartha contrasted this with other companies, which he felt used computer vision for "scary" purposes, like highly targeted advertising. This contrast was intriguing because Sybil provides affective classification for targeted advertising, a use case that Siddharta expressed was a violation of privacy and something he personally disliked. However, because he was working on research on computer vision for autonomous vehicles, he was distanced from his company's main product.

Some participants seemed to be explicitly against what they felt their own companies' values and priorities are. Lynn, who initially joined MultiplAI because she felt it was a "social good" company laments how she feels the company has changed:

> *"We were a social good company, you know, we founded clarify for good, there were four of us doing charity, mentorship, you know, helping free products for students, you know, trying to partner with NGOs, and researchers and just offer, you know, our services to help make the world a better place. And we had executive sponsorship, it was like a bright spot in a company, everybody was really excited about it and things like that. And that was a big part of our identity when we first started, you know, it was, it was really lovely, you know, looking at the article that was written about me, you know, the company has become a weapons company wasn't something that any of us expected when we first joined."*

Lynn's values sit in opposition with the company's market-based approach that Kenny described, building any product they legally can. Elliot expressed similar beliefs about their company, describing feeling like an "*outlier*." Yet they also describe finding colleagues at the company who also do not believe in the overall capitalist mission of the company to put profit before social good: *"I've been finding lots of other people where we are like, we're in this company, but like, we don't want them to be doing a lot of what they're doing."* Lynn actually left the company due to differing values. Such instances represent how the company context might drive workers with certain values—in Lynn's case, a focus on social good over profit—away and potentially lead to a company full of workers that operate from a status quo perspective.

Much like regulation and policy, whether workers agreed with their company's soft values colored whether they felt they could approach identity the way they desired. Workers like Nicholas identified deeply with the values of their companies, so he felt his own positional perspectives were being represented in his work. On the other hand, workers like Lynn disputed the value of her company so greatly that she eventually left. She felt her own positional perspectives could never be properly represented in MultilplAI's approaches.

Ethan emphasized that the overall context of research and development occurs in relation to not only the company, but the broader societal context the company sits within. He described how the broader social context, beyond the development of the product itself, influenced how workers approached identity. He said:

> *"Research doesn't happen in a vacuum. And just like any, you know, academic research or anywhere else, the kinds of questions you're asking, there's a reason you're asking those questions. And it's, it's driven by societal concerns, by company concerns, by what you can get funding for, by … all of these kinds of things. And they all come together to come up with what you're asking."*

To recap, participants felt that their positionalities were either constrained or enabled based on the context of their companies. The contexts influencing positional perspectives include: the economic power of their company, legal and policy landscapes, and whether their company valued certain approaches to identity in computer vision. Finally, the broader social context that exists outside of their companies also influenced their outlooks and approaches to identity.

How Worker Positionality Influences Product

Now that the contextual factors in which workers are situated have been established, I will present how workers apply their own personal perspectives to their work and how they negotiate those perspectives with other workers and clients. First, I will describe the personal interest some workers have in defining identity, as well as instances where workers felt no personal attachment to their work. Second, I will describe how workers negotiated their own positionalities with that of their colleagues.

(Im)personal Stakes in Computer Vision Work

One of the ways worker positionalities became particularly apparent was when workers discussed their own personal interests and stakes in defining identity. Many participants attributed an interest in working on identity issues for computer vision to their own personal values and affinities with certain identities.

Kaleigh, whose primary role is overseeing fairness initiatives across multiple computer vision products at Aqueous, described how the majority of resources come from people who are personally passionate about fairness. As a concrete example, Kaleigh is working on guiding product teams to update their approach to gender categories and concepts across computer vision. In doing that work, she is identifying and contracting researchers with gender-specific expertise to bring product teams on board with the changes. *"Right now, getting a lot of fairness help in cutting edge research has been just finding researchers who are passionate about this and willing to devote their time to it, and [who will] help us bridge the gap between what already exists and what we need for product."* Kaleigh's helping connect product teams to appropriate research resources in instances where product needs to move from the status quo ("*what already exists*" in product) to what product should be ("*what we need for product*"). Those participants who contributed to the goal of updating identity in product had a personal passion for doing so.

Beyond personal interest or passion, specific affinities with group identities played a major role in workers' approaches to their work. Vasudha (as a person of color), Kenzie (as biracial and a child of immigrants), Elliot (as non-binary), Lynn (as a wealthy white woman) and Madison (as a woman) all recognized how their own identities played a role in their work. Madison described how researchers at Maelstrom prioritize certain projects: *"Initially deciding what projects to focus on … is at least partially informed by people's identities and who they are, not only as a researcher, but also who they are as a person in life generally,"* she explained. She went on to reflect on a specific example of a colleague who is working on improving machine learning classifications for LGTBQ people:

> *"One example in particular is one of my coworkers on my team is gay … So, they felt a personal interest in this, as well as a professional interest. So that's one way that … that identity is manifesting. There's the actual, like corporate or tech, kind of [way] like, what are the terms? How do we deal with them? But that's also tied to the individuals doing [the work] and what they want to prioritize."*

Beyond driving this colleague to work on a project focused on LGBTQ identity, being LGBTQ also informed their approach to their work. This worker collected data at pride from people directly to both avoid online trolls and to work directly with the LGBTQ community. Madison felt this *approach "embraced identity by the horns,"* rather than letting it be implicit or neutral, as simply demographics or data points.

Much like Madison's colleague, Vasudha brought specific expertise to the table due to her own positionality as a person of color. She explained that historically, product teams at Aqueous have approached evaluating racial biases in facial recognition models by using skin tone. Yet, given her own experience, those categories were ineffective because they were far too static:

> *"At least with me, if I go swimming in the pool, … because of a brown skin tone and a lot more melanin content, if you leave me under the sun for 20 days, I have like five shades darker skin tone … So, when you think about comparing those, skin tone didn't really make any sense … We realized very quickly that skin tone wasn't something that we could test on, especially taking into consideration, sort of the aging aspect."*

Vasudha knew from her own personal experience that, if they were to use skin tone as a metric, the system may be unable to recognize her face over periods of time as her skin tone fluctuated. This personal knowledge also informed user studies her team went on to do, which showcased a need for information beyond skin tone, including facial structures. Of course, this did not make the decision for which categories to use necessarily easier or more concretely correct. She continued: *"We pivoted to ancestry background … [but] there are issues that come up with 18 demographics. Why not 24? Why not 36? Like, there are just a lot of these questions that come into play."* Vasudha's personal identity made asking questions about the viability of skin tone more obvious to her, but she didn't posit herself as able to determine what the best course of action was. Based on my analysis, I found there can be financial constraints to increasing the number of categories, let alone logistic constraints for defining and collecting the data. While Vasudha wasn't sure about the reason for 18 categories, she acknowledged that there are tradeoffs in designing categories for identity.

Irina, the CEO of an ethical data company, described how she would screen potential clients for data projects. She would assign each potential client an impact score, ranging from 1 (the project contributes to social good) to 4 (the project causes active harm, like military projects or content moderation projects). Her policy was to never accept projects with a score of 4. She described how, in cases clients came to her and she felt the projects were too harmful, how she went about rejecting them:

> *"I remember there was this really problematic one about intelligent weapons … You know, you don't want to offend the people and tell them that they're horrible people and they shouldn't be building this AI. Usually what I say is that we're a social enterprise and given that the majority of our workforce comes from conflict affected countries we're not able to perform such type of labeling. We have some Palestinians as well and this is also a personal preference of mine that I've instilled in the company to reject any project from an Israeli company, so in that case we tell them we work with … Palestinian refugees so we prefer not to work with Israel. So, in that case people do get kind of offended."*

Even though her business requires making a profit, since her business is also focused on social good for her workers, she is strict about which projects to take on. At the same time, she acknowledges that the clients will simply go to another data provider and the datasets and models will still be created.

She told me that she connected the rejected weapons client with iMerit, who happily took the project on. Lynn also described how she would assess the ethics of a project:

> *"There were other use cases where people were trying to classify based off of stereotypes, that was a fun project … [The client] wanted to build a model that would recognize something like that, which was a horrible, horrible project for good cause, like, believe it or not, I can't go into detail about it. But we took that project on because he believed in the message that they were actually trying to."*

Operating beyond the minimum requirements of the law, Nicholas explained that Resoom also tries to measure biases otherwise not protected under the EEOC (U.S. Equal Employment Opportunity Commission) *"from a sort of goodwill perspective."*

Many workers also simply have not been exposed to thinking about identity critically, even though they may reference their own social identities in their work. Kenzie described becoming interested in fair machine learning after attending a software engineering bootcamp where famous fairness researchers, like Timnit Gebru, spoke. She had not been exposed to discussions of bias in technology prior to that. *"It blew my mind um and yeah I just became very interested in it,"* she explained. She implied she has also since reflected on her identity as Brazilian and the diversity of Brazil's relevance to its use of AI.

The positionalities people brought to their work were not static, and changed over time as workers were exposed to new ideas and adapted to their company context. As Vasudha explained, when she was in graduate school, focusing entirely on research, she was not considering a business context. She explained that simply relabeling and retraining models at the industrial context is often economically untenable, especially due to labor costs:

> *"When I was in grad school, the mindset was different. Because I'm not thinking about dollars, I'm not thinking about how much it's bringing in or what are the ship timelines and things like that. So, you put it in a very practical aspect of this is a product, a company is trying to monetize or help customers build experiences through it. You have a very different lens where you don't have infinite time or infinite resources."*

While workers influenced products, workers were also influenced by their surroundings as well. The above examples showcase how workers make subjective judgments about what is ethical or "*goodwill*." Such subjective judgments reflect their own positional perspectives, informed by their experiences and beliefs. They relied on their own familiarity with identity concepts—either because they personally valued concepts like equality or ethics, because they identified with the identity attribute in question, or because they became exposed to values they later came to internalize.

On the other hand, participants also seemed to lament values or perspectives that differed or clashed with their own. There seemed to be an overwhelming perspective that those in heavily technical roles—like engineering and data science—had an interest in tasks rather than social implications. Kaleigh described how machine learning research teams often explore novel problems out of personal interest, which then later become embedded into products. *"It's what [they are] interested in … and then sort of after the fact it might make its way into a product if it seems promising. … they have flexibility to decide what problems they want to go after."* This approach was the case for one of the computer vision products under her purview as a program manager, a mobile application for real time classification for accessibility purposes. The researchers at the time created the product to classify gender and age because they had felt it would be useful, though they did not assess utility in any empirical way. Coleman, who worked on a prototype of this accessibility product for a research project, described choosing human characteristics that seem useful. *"We tried to make sure there are certain things like person names … Because otherwise a human might not find it useful anymore."*

Much like Kaleigh described, Coleman took a utilitarian approach to his work on identity in the product. Coleman described how the "*part*" of himself that is trained as a machine learning researcher desires to build new models and focus on improving methods, perhaps at the expense of fairness:

> *"So, I mean, part of me is a researcher who wants to get basically whatever data I can get my hands on, toss it into the data grinder and build models. And once they show a very significant improvement over the state of the art, that for me is a paper and is potential progress on my methods, right? And this can become an end to itself in the sense that I can ultimately become very blind to how people might use that system. And ignore the fact that maybe I've just produced something which is very, very biased towards certain things, which have optimized my scores."*

Coleman is highlighting that his own training, as a machine learning researcher, is not sufficient to deal with bias in model deployment. He offers an implicit commentary that those with a machine learning research focus have to be pushed to think more deeply about identity bias, against the disciplinary norms they are encultured to. In some cases, machine learning researchers may simply not view identity bias as relevant to their position and expect others in the company to handle it.

Similarly, Nitesh described not really knowing anything about how identity tags are selected or filtered in the image indexing system he worked on. He stated that it was not his responsibility, but rather the responsibility of the data science team, to consider how identity is represented and to mitigate biases.

Much like Coleman's motivation to pursue state-of-the-art modeling, Nitesh's major motivation for pursuing a career in computer vision was solving technical problems, not social ones. Nitesh was so out of touch with how identity was implemented in the model he was engineering that he was entirely unaware that slurs were associated with targeted subgroups on his company's public facing website.

Callia, Cole and Kaleigh's colleague, took a more personal approach. She talks about having a blind child and spending a lot of time in the blind community, saying these issues are familiar to her—and implicitly personal:

> *"You know, [they] made them quite innocently thinking, 'Well, you know, this is what you do to do computer vision, without really thinking about the larger context of his decisions.' And I would say, actually, in my experience, eight years of computer vision, the most harmful decisions for the computer, for the user experience are usually made are those off the cuff decisions that our researcher makes by themselves, or engineer makes by themselves, thinking that it has no impact on the larger experience ... So [my manager] said, 'Well, either you take it on, and you lead this project and figure out how you can make it not silly, or you let them do it.'"*

She believes that projects on computer vision that have largely been driven by machine learning researchers require intervention from more socially focused colleagues, since the focus is more on being able to do a task than it being "*human-centric*." She discussed the utilitarian approach the technical team was planning to take to the accessibility application as potentially harmful. Her description indicates a sense of personal responsibility to ensure the project was human-centric, a moment of intervention to prevent the product from becoming too *im*personal.

Much like Callia, Elliot critiqued the technical approach to scoping identity. In particular, Elliot was concerned with how categories like race and gender are represented—implicitly, because of a personal stake in gender as a non-binary person:

> *"I honestly, it's a bunch of like … you know, predominantly like predominantly white, cis, male engineers who have not thought too much about identity, just kind of like treating things as fixed and trying to label it … Oh, well, gender is binary. So, of course, this is something I'll just incorporate, as opposed to, like, you know, thinking that through like a little bit more."*

Elliot's perspective explicitly questions the positions that many engineers inhabit—white, cis, male—and assumes they lend to a less thoughtful approach to identity in product. Interestingly, Callia expressed that gender representations in computer vision weren't as big a deal as some others were making it out to be—her focus was on accessibility for blind people, and she thus advocated a binary gender. These clashing perspectives also highlight that personal affinities and experiences with certain positions motivate workers to approach problems in very different ways.

Vasudha points out that, just because technically focused workers might prioritize metrics over social impact, they are "*not malicious or necessarily bad people*." *"It's just that they have a problem. If a metric, they optimize the problem until the metric score goes up. That means for them oftentimes, the implicit bias in that metric is something they don't even consider."* Vasudha highlights that technically focused workers, like machine learning researchers, are often unable to see issues of implicit bias. This insinuates that such workers are simply approaching their work from a very different positional vantage point, one which prioritizes model performance over identity biases.

Traditional workers made their own positionalities clearer through their disagreement with the perspectives of colleagues. Largely, it was those with an acute interest in engaging critically with identity who expressed unfavorable opinions about colleagues they saw as uneducated or unengaged with identity. These workers tended to assume it was the technically-focused colleagues who weren't engaged in critically thinking about identity, insinuating that the position of a technical worker—like engineers and data scientists—was distinct from those less technical workers—like researchers or policymakers. The distinction between these two positionalities—the workers who "care" about identity and the workers who do not—indicates highly different approaches to identity work in computer vision.

Negotiating Positional Perspectives with Others

As demonstrated by how different workers disagree with each other's outlooks, designing identity in computer vision is a team endeavor. Just as decisions about identity were not made "in a vacuum," they were also not made by singular individuals. Beyond the role each individual's positionality played in motivating their interest in specific work and guiding the decisions they make in conducting their work, participants regularly had to contend with the positionalities of others.

Most often, workers were collaborating with those inside their own teams. Generally, workers seemed to share values and perspectives about identity with those colleagues on their direct teams. Elliot compared their own team with how other more product-focused teams approach their work: *"Within Maelstrom, my team is probably like, I think the best team in terms of like, thinking a little bit more* *critically about machine learning systems as sociotechnical systems, as opposed to just kind of like, algorithms where data comes in, and data comes out."* Elliot is describing the focus of their team, which is specifically designed to address issues of ethics in machine learning. The team focuses on understanding and creating tools to address issues of accountability, transparency, ethics, and analysis for machine learning systems. The description of the team indicates a certain direction that individual members can play in the computer vision space, focused on critique and solving ethical issues, rather than necessarily building computer vision products for the company to sell. Rather than being surrounded by pragmatic engineers focused on ensuring the best possible product, Elliot is surrounded by fellow researchers who care most about ethics. Given that all members of the team focus on ethical issues, it likely shapes their worldview in how they approach their work. Given the way that Elliot presents this information, as a collective ("*we're interested*"), Elliot sees themselves as part of this larger mission with an ethical focus. They agree with their team members and their approach within the company.

Yet, workers also acknowledged that having a team that understands and welcomes your perspective is not always the norm. Macy explained how words like fairness and ethics tend to make people's eyes glaze over or take people aback because *"they don't feel like they can do anything about it."* Instead, she tries to incorporate concepts while avoiding the terms. Company approaches might differ based on the size of the company. A huge company might have teams that act as independent companies. She also feels that companies are taking a *"customer first"* approach but the companies are the ones defining the customer—it is similar to critiques of "users."

Elliot described that *"[it] very much depends on your chain of command, also … Like I feel very fortunate to be in a very supportive chain of command … Lots of individuals might not be in such a supportive environment."* Elliot describes not being concerned with retaliatory efforts against them for their perspectives but implies such retaliations may occur in other parts of their company. Retaliations from those higher on the "chain of command" indicate how some workers might use their power over others to quash specific positional perspectives. Madison described how the company and those in it were not supportive of addressing identity-based biases in product, and so it was an uphill battle with those in more powerful positions who didn't value her perspective:

> *"Initially, there was just no infrastructure … So, you were just fighting against skepticism, tied with the fact that there was no requirement for doing it. So that was just like, a lot of sort of pushback, and then making the case that it's worthwhile can be hard … So yeah, so I think like, originally, when we were hitting barriers, just from people just like, not being familiar with some of the basic ideas, and just being like, I don't have to, why should I kind of thing."*

She also described appealing to perspectives she knew were more valued, identifying ways to instill her own beliefs into accepted approaches:

> *"You know, you could also make the business case that this is about inclusion, this is about reaching more people and making more people feel comfortable. Which works a little bit better, because then you can also tie it to expanding your user base or whatever else."*

Madison describes, in early stages of establishing her team's focus with management, appealing to priorities she knew that management cared about, like growing potential user groups for products.

Of course, in industrial contexts, people often collaborate with others outside their core teams. While Kaleigh says it is more common for research teams to work in a silo at Aqueous, *"not really necessarily even thinking or caring about what product it will go into initially,"* there are also instances where research and product teams work together from the get-go. Kaleigh describes how different types of teams often work together to influence the outcome of a product. *"I know of some research groups that work very closely with product teams. And then the focus from what I've seen is much more on what's the research that will directly lead to certain product improvements or help us figure out the next thing."* In the cases of collaboration between research and product that Kaleigh describes, team members actively learn from one another's perspectives and expertise to shape identity outcomes in product.

However, positional tensions seemed to occur more often when working with colleagues on other teams, particularly when those teams had very different roles and goals. When teams work together on products, they often have different viewpoints on how to approach identity. As Callia explained, from her perspective as an accessibility researcher, *"there's a significant amount of negotiation to orient slash reorient [computer vision engineers] in a way that accounts for the human experience."* Callia expresses viewing her position, and other human-centered researchers, as oppositional to technical researchers, who are mostly focused on solving technical issues and not accounting for human experience. Elliot similarly criticized machine learning colleagues about their approach to racial identity as attribute-based rather than something sociopolitical:

> *"Basically, telling machine learning researchers that, like if they're going to be in the business of like, making predictions about people and like, affecting things in the world, they very much need to adopt the like, kind of, you know, former strategy of like, understanding racial dynamics, as opposed to, you know, this is some fixed attribute of an individual and we're just going to control for it after the fact."*

Clashing perspectives also became evident when workers discussed disagreeing with identity approaches in other products their company provided. For example, both Jeremy and Siddharta criticized identity classification models, which they did not directly work on, but their company provided as a core product. Jeremy claimed his team would push back against demographic classification in his gesture recognition product, stating *"to me [it] seems to be either dangerously uncomfortable, or at the very least."*

Siddharta, who works on affective classifications for autonomous vehicles, more explicitly named how it violated his own personal values:

> *"For example, if gender is one of such information, and if you're infusing that to the model directly, then for a transgender person, when they are driving the car, the model might not provide the right answer or might ask for the gender of that person. And can that that that probably is not the right thing to do. So that is why we don't explicitly use such information directly in training … I think it's mostly personal for me … Personally, I don't want to divulge that information. I don't want to give away my own information to everything that I do."*

Siddharta is expressing first an awareness of how certain demographic classifications might harm certain user groups, and second his own personal beliefs about individual privacy rights. While both Jeremy and Siddharta used identity information in their work for evaluating model bias, their beliefs implicitly contradicted those of the colleagues in their companies working on demographic classification models for advertising clients.

Kaleigh contended with negotiating different perspectives regularly in her work, as product teams pushed back on changes to gender in computer vision. Kaleigh explains how workers' personal interests become tied up in product, making it "emotional" for them to change them or let certain features go:

> *"Product teams … have a lot of investment in them , because they've been working on them for years, and they've been committed to them for years. Some are really, really good about recognizing that things have changed and we need to conceptualize this differently. But for others, I was in a meeting recently, where it was quite emotional for them to let go of a feature that doesn't align with our responsible AI principles and values anymore, even though it's a feature that they've been working on for years."*

Kaleigh explained that, though her team had the power to step in and make executive decisions about changing identity in computer vision, they did their best to work with teams to avoid causing internal conflict. *"We try to avoid the sort of a mallet approach of just saying this is the way it is. And it's tried to reserve that as a last resort, if sort of bringing them along hasn't helped, and thinking through it together hasn't helped."*

Madison explicitly attributed difficulties implementing more fairness ideas in computer vision to "politics" around the identity groups that dominate tech:

> *"If you have a group of women saying how they think the technical aspects should go, and it's what you're not used to hearing, there's like, no way they'll be taken [seriously] … But then if you have people who are white or Asian men, maybe more fitting like the traditional personalities of who tech people are, putting forward ideas, it does get a lot more traction."*

Madison describes her ideas being taken less seriously in tech spaces, whereas those who inhabit identities more traditionally associated with tech are listened to. She describes how she has received retaliation from colleagues through bad performance reviews, which impacted her career trajectory. She believes that such perspectives are so deeply ingrained in people's approaches to their work that they aren't even necessarily aware they are being biased. She states that teams *are "more comfortable"* with Asian and white men, insinuating that certain social positions have more power to make decisions in her company. *"I guess, it's just that, the desire to maintain a white, Asian, cis, male view of the world is so strong, and people don't even realize they're doing it."* In order to ensure her perspectives are listened to, Madison describes relying on white and Asian male allies to communicate her ideas for her:

> *"We have this white male front. Like if you have someone that people feel comfortable with, and you're like, I'm just with that guy. That can be a lot more effective. I mean, for better or worse, right? Like, to me, it makes the argument that like, now you're just reinforcing this idea that these people should be leaders and you're like a supporter. But then on the other side, … I want this idea to be out there sooner rather than later. So, it's kind of like you have to decide what's, what's the pros, what's the cons and which ones you want to prioritize the most."*

Others' positionalities were not only seen as a barrier which had to be overcome, but also as a resource for improving approaches to identity. Kaleigh describes bringing in consultation for particularly thorny identity issues:

> *"For some of the highest risk things that are really, really sensitive, then we'll also bring in some of the company's top responsible AI leadership to consult as well. So, there's a lot of different voices that come in and depending on some of the products and particular scenarios, we'll also where we can try and pull in perspectives of people with different backgrounds, we're making sure that we're considering that as well."*

While positional tensions could cause stress, infighting, and even retaliation, it could also push teams to think outside of their comfort zones.

Nitesh described how he *"support[s] diversity"* because *"you get different approaches to a problem."* He felt that *"constructive conflicts"* are reflective of the *"complex world"* that products are meant to serve. In particular, he highlighted that teams with diverse training—such as from diverse academic backgrounds—would be beneficial.

Beyond whether to include demographics, like gender and race, are conversations about *how* to include them. Participants had differing perspectives on how to approach identity categories. Madison explained the difficulty of defining unstable identity constructs*: "Just deciding gender [and] race, such unstable constructs, like, different cultures have totally different ideas of what these are, and it changes over time, and all this stuff."* Therefore, when assessing bias on identities like gender or race, they are attempting to identify the cues which lead to gender or race classifications.

While workers like Elliot, Siddharta, Madison, and Lynn viewed gender as difficult to define from visuals, Kenny viewed it simply as "a logical human decision" on behalf of the data workers annotating gender in images.

Given that many of the companies that participants worked for provide computer vision solutions for clients, teams of workers also negotiated different positional perspectives with client representatives. As already demonstrated throughout these Findings, some workers expressed value differences around certain computer vision tasks (like identity classifications and privacy violations). Beyond different perspectives on computer vision tasks themselves, regional differences were often a source of positional differences, as tech companies generally operate at an international scale. Some workers were even called upon to act as intermediaries between company and client representatives due to their positions. For example, Kenzie, who speaks Portuguese, describes having to translate the requests of a Brazilian client—even though she only recently joined Phrenx.

However, beyond relying on positional experiences like country of origin and language, workers often had to negotiate cultural expectations about identity in computer vision products. For example, Nicholas said:

> *“We've been asked to do some things in prediction, that wouldn't be good to do in the US. But would it be okay to do so in that country? And we've just said: No, we won't do it … To me, it's not the right thing to do. So, we're not gonna do it. But to do it would be really bad for business if it got out.”*

Nicholas and his colleagues had considered specific requests that might be viewed as unethical or inappropriate in a US context and decided against implementing them due to numerous shared values. One of those values was simply morally disagreeing with the request, but the other was not wanting to harm the company’s reputation in the US by accommodating the desires of clients outside of the US. Nicholas also described the moral dilemma of whether it's appropriate to instill US-centric ideals into products meant to be deployed in other countries:

> *“We're an American based company, does that give us a right to sort of inject that attitude on another country and their hiring practices that don't follow that? We have the right to not do business with them, but do we have the right to inject our beliefs and change algorithms for them?”*

Lydia commented on a specific instance where she and colleagues made a decision to include bias mitigation practices despite a client being unconcerned by bias:

> *“Different countries have different laws around bias and fairness … like, Japanese data is usually really sexist. And like, they don't care about that in their country. So, like, the customer doesn't care. Whereas we would want to mitigate [gender bias] … So, I guess for me, it's just kind of like having that conversation where there's cultural differences of where that balance should be.”*

Lydia is not only elucidating the differences between herself and her client, she is expressing her own view of Japanese culture from the positional vantage point of a United States citizen. She is packaging both the views of her client and the views of another culture together in her assessment of how to approach gender in the product.

Further, approaches to identity shift over time, as individual workers, their colleagues, and broader cultural conversations around identity concepts change. For example, Vasudha explained that their annotated dataset for evaluating models initially *“started off with the four or five ethnicities, which is what you see in most surveys.”* Over time, through conducting user research, they expanded to eighteen ethnic categories:

> *“So, it's a combination of us using the research that's present. And us deploying resources to go identify what could be those different reasons where product teams could or would want those groups to be separated or put together? So, it's a combination of that we work with responsible AI teams to come up with that plan.”*

Madison also expressed that those outside of tech—like activists—can push product approaches in new and more positive directions: *“It's sort of a system where we can help one another a little like, so activists help to really push the headlines in a way that can inspire slash force companies to do stuff they might not otherwise do … it's a system that is the activists, externally, the reformers internally, and then the corporation itself.”*

As described in this section, determining how a product should be designed requires negotiation between many different actors. Participants negotiated their perspectives with their colleagues, but also clients and users. Furthermore, interfacing with others acted to mutually construct how individual participants interpreted identity. As workers encountered different viewpoints from their own, their own viewpoints grew and sometimes changed—what Nitesh described as “*constructive conflict.”* Representations of identity in computer vision are therefore not the perspective of a single person or even a single team, but a multitude of actors within and outside of the development context.

Positional Gaps that Arise During Product Deployment

As demonstrated by the Findings thus far, participants approached their work from their own positional perspectives. Gaps in their individual perspectives were revealed and negotiated through collaborations with colleagues. Yet, workers situated within tech company contexts are often unable to predict how their own positionalities, as incomplete images of the world, might result in positional gaps in product design. These gaps often only become visible when products are tested or even deployed and begin to negatively impact the people who come into contact with them. As Coleman stated, *“As we move out of being in our research Ivory Tower to an actual production group it’s become much more painful.”*

Unforeseen Outcomes Due to Positional Gaps

Once products were finished being developed, identity issues sometimes arose. These issues were largely caused by the development team being unable to foresee them, because they did not occupy identity positions that would make such issues obvious.

Sometimes, these issues were caught before deployment—particularly for companies with the resources to do internal testing. Madison described a scenario where a product team of primarily men had designed a computer vision wearable. Before deploying the product, they conducted internal testing, commonly referred to as "*dogfooding*" in the tech industry. *"As soon as they had the dogfooding of the kind of necklace version … the women realized that the camera was like on their breasts ... But all the original designs had been developed by flat-chested men. So, they hadn't even thought of that,"* she explained. This example highlights how the positions the men inhabited made it difficult for them to automatically recognize their product was uncomfortable for people with breasts. Further, those with breasts were quickly able to recognize that the product did not work for their body types: *"[The women] were able to discover that because that's part of who they are."*

Many positional gaps are embedded into products because neither the clients nor the workers even realize their perspective could be biased. Irina, the CEO of EnVision Data, described the history of the ChAI project, focused on providing AI-generated insights about video job interviews. She said that the client had originally gone to another annotation company but ran into strong cultural biases in the data annotation. *"They were working with an Indian outsourcing company and people were much more favorable towards Indians,"* which led the company to seek re-annotation of the dataset. The client had not expected the annotator to have a deeper understanding or affinity for those applicants which shared the same ethnic positionality as them. Therefore, the client decided to create new *"objective gestures"* categories in an attempt to mitigate, or at least measure, those biases—such as whether the person in the video *"exhibits brow furrows"* or *"nose wrinkles."* Irina felt that these were still very subjective categories. And further, much like the last company the client went to, EnVision Data had a homogeneous group of annotators*. "We were still working only with … people from Middle Eastern origin … We didn't use a diverse group of annotators."* While Irina did not notice any specific biases favoring workers in the annotations, she still felt there might be some gaps caused by the annotators being ethnically homogenous, rather than diverse.

In other cases, issues only arose once the product was publicly deployed, and users of the product encountered issues. Often, given the issues were relevant to identity characteristics, users found these issues offensive. Coleman described that many gaps might not be addressed *"unless enough people scream"* about them. As someone who worked on both text and image-based machine translation, he encountered a number of accidents that resulted in major PR problems for the company. For example, he described how religious entities and names are mistranslated. In one specific case, formal Russian names were misgendered as female instead of male. *"[Users] somewhat assumed there was a specific model trained to discriminate against themselves personally."* Some of the accidents resulted in workers at the company even receiving personal threats. Coleman lamented these mistakes, though he also felt they were a learning experience for him and his colleagues. *"[It] was a very healthy shock for our ecosystem, because now people have understood, okay, if we do this, we then at least have to give set up some monitoring for the first week … maybe have a new release to be able to very quickly un-deploy if something bad comes up."*

Nicholas reflected on the high-profile mistakes that some big tech companies have made around identity. Given his focus on fair hiring practices at Resoom, he felt that the incident in which Amazon's resume classifier discriminated against women was "*obvious*."

> *"My take on this as these companies, tech companies, if they're creating AI for a purpose, like Amazon was, they need to have an eclectic team involved. Amazon had a bunch of data scientists. They didn't have a huge team of IO psychologists at Amazon, they didn't draw on that expertise. They have they probably wouldn't have gone as far down the road as they did without checking for biases discrimination."*

He felt that workers with expertise beyond data science would have meant avoiding the issue of gender discrimination entirely.

Madison felt that, while robustness testing is useful in the constrained context of improving models, it often ignores the sociohistorical reality of identity:

> *"[Defining demographics] can be useful as a kind of general robustness testing for models. But it's naive to the fact that these sort of culture specific subgroups are the ones that really carry pain for people that carry historical discrimination. Like we can be bottom up, and that's cool, technically, but it just totally ignores the reality of real life and like, you know, real identities that get tied up in all kinds of stereotyping and prejudice."*

Workers occupy specific positions that allow them to contemplate and imagine how products might be used and what errors might occur. Because workers occupy specific positions, they also often encounter gaps in their experiences that lead to errors and issues. As I will discuss in Chapter 9, traditional workers, who largely control the process of identity development in computer vision, often come from more privileged backgrounds, likely contributing to unforeseen and offensive identity outcomes.

Multiple Voices Improves Approaches

While negotiating positional differences can be difficult, most participants expressed that a diversity of voices improves how people approach identity in computer vision. Having a diverse workforce means having a diverse set of positional perspectives to resource from.

Madison described how, when building her fairness team at Maelstrom, that she had accidentally ended up with a diverse team. *"I ended up with a team of people who are like, diverse along a lot of the underrepresented categories in tech. So, like, gender, sexual orientation, color. I don't know, religion. But it was weird, because I wasn't trying to do that,"* she explained, assuming that people from more marginalized backgrounds might just be more interested in fairness issues. *"It's a lot easier maybe to build up a team that has knowledge about different kinds of identities, because of real world, lived experience having different identities."*

Lynn expressed the difficulty associated with diversifying data so that computer vision works appropriately on diverse groups of people, in this case interpreted as diverse demographic groups:

> *"What groups do we start with? What groups do we prioritize? And it's a horribly kind of like, it's time I had this really sinking feeling of, you know, me as this like white woman with a fancy education and like a great job. And I'm sitting here thinking about, like, where this technology is going to be used is unknown, right?"*

Tied to the social and technical (budget and collection constraints) difficulty of prioritizing certain demographic groups in data collection and testing is an expression of doubt that she is the person qualified to tackle these issues. This doubt is not due to her skills, but her own racial identity, gender identity, and class identity. She is viewing herself as in a position of privilege—a privileged positionality—that limits her qualifications to improve fairness for marginalized groups. Yet her awareness of her own position provides a different perspective to apply to her work than those who are unaware or have not thought deeply about their positionality.

At the same time, deploying and getting feedback from the people interacting with product was seen as positive, because some of the gaps workers missed might come to light. As Coleman said*, "the more diverse your pool of users is, the more diverse your training data augmentation is."*

Lynn, whose views on computer vision changed as she was exposed to ongoing fairness conversations, expressed new beliefs that identity categories cannot be ascribed visually to the body:

> *"The more I learned about this field, the more I agree with, these notions of race and so many aspects of your identity, your body, that's not part of your face at all, … we can't possibly glean your race from just a photo of your face. … it's hard knowing all this stuff, like when you get right down to it, learning all the lessons that I learned in the last two years at MultiplAI [has] changed my life dramatically. And it's turned me into an activist, right?"*

Much like Kenzie, Lynn had also changed her own perspectives as she was exposed to new ones.

Diversity in traditional tech workforces not only helps to cover positional gaps which might otherwise be overlooked, but it helps the other workers around them to learn and develop their own perspectives further. Participants themselves valued colleagues with different perspectives and identities to them, in contrast to their lamentation of certain positionalities as described in above.

Minoritized Workers are a Resource for Positional Gaps

To avoid positional gaps, many participants described how fellow workers from marginalized positionalities were a resource for vetting or feedback. Colleagues from marginalized identities could name issues with products that others on the team were unable to see due to their limited positional standpoints.

Lynn described how there was debate at MultiplAI about the language used in the demographic classifier. While Lynn found the model itself "*problematic*" because it was assigning categories to others based on their appearance, she felt that the language they used lessened the negative impact. While initially the demographic classifier used "sex" and "race" as demographic categories, they pivoted away from more concrete concepts to more appearance-based language:

> *"Instead of saying male or female, we used masculine or feminine, more of a descriptor than a prescripter. Yeah, there was a big debate about it, the whole company was involved. … we had some genderfluid people in the office, we had some trans people in the office. So, their opinions were really important to us. And in the end, we delayed the launch, so that we could actually focus on those kinds of things."*

Similarly, Elliot described how it was common practice to rely on identity-based "resource groups" at their company:

> *"[There are] groups of [people who work at Maelstrom], who like, share some facet of their identity. And so, product teams will often consult with those resource groups. And, and be like, okay, hey, like, I'm building this product, can I like, consult with this, you know, group of people who identify in some way to understand like, is this meeting your needs, and this is kind of like a way of, you know, before doing some kind of, like, external user testing, being, like, I have access to, like, you know, all these [Maelstrom*

> *employees], you know, who, like, represent some population of the world. And so, you know, we can use them."*

Traditional workers viewed colleagues from different identity groups as valuable resources to improve the product. In Lynn's case, some workers organized to address a gender gap they felt was problematic and delayed the launch of the product. In Elliot's case, some sought to test products with identity groups they felt were relevant. Particularly for participants working in large companies, employee resource groups for workers with specific identities (e.g., LGBTQ) were viewed as valuable resources for testing products before deployment. In both cases, participants expressed that those occupying positions outside of their own limited viewpoints could make products more robust and more inclusive.

Discussion

Traditional tech workers all bring their own identities to the table. They operate from their own positional perspectives when conducting identity work for computer vision. These perspectives are influenced and constrained by the industrial contexts that workers are embedded in. For example, workers in smaller companies are often more constrained in how deeply they can engage with identity than those in large companies, due to lack of economic resources and incentives.

Positional perspectives are especially evident in the personal—or impersonal—reasons participants expressed for engaging in identity work. Of course, their individual positional perspectives may or may not align with those of their colleagues. Workers often framed disagreements with their colleagues as negotiations, sometimes implicitly expressing displeasure with their colleagues' worldviews. Further, given each worker brought their own positionalities to their work, teams had gaps in positional worldviews. Such gaps would lead to unforeseen and undesirable outcomes when it came to product design, such as offensive classifications or hardware that only worked for some bodies. Aware of their own positional gaps, workers were proponents of diversity in tech, and would often attempt to use colleagues from minoritized identities as resources to augment their own limited worldviews.

In the remainder of this Discussion, I discuss how individual workers' positionalities are mutually constructive and informed by actors across a variety of contexts. I discuss how the company context workers are directly embedded in, the broader development context of product development, and even those outside the development context influence worker worldviews. Further, I discuss how the macro social context in which workers are broadly embedded when conducting their work is ubiquitously implicitly influencing their positionalities when designing computer vision. I conclude with implications for attending to positionalities at both the context level (company, development, outside development, macro social) and the actor level.

Positional Approaches in Context

Positionality is the complex, mutually constructive relationship between one's identity and how they view the world around them. Every individual occupies a specific position in the world, impacting how they are viewed and treated; as a result, they view and interact with the world from a specific standpoint (da Silva & Webster, 2018; Rolin, 2009). The traditional tech workers who develop computer vision are no different. In implementing identity characteristics for industrial-scale computer vision, workers implicitly rely on their own positionalities. They shape identity in computer vision from their own standpoints. Often, their own interest in identity in computer vision stems from their personal values and their own affinities with identity characteristics. On the other hand, those who do not explicitly value fairness for identity groups or identify strongly with specific identity characteristics express little knowledge or interest in identity issues in computer vision.

However, the process of developing identity in computer vision is not simple and straightforward. Individual workers do not make individual decisions about how best to implement identity. Rather, workers operate within a complex environment, informed by many different contexts. **Figure 13** illustrates the contexts in which each individual traditional worker sits.

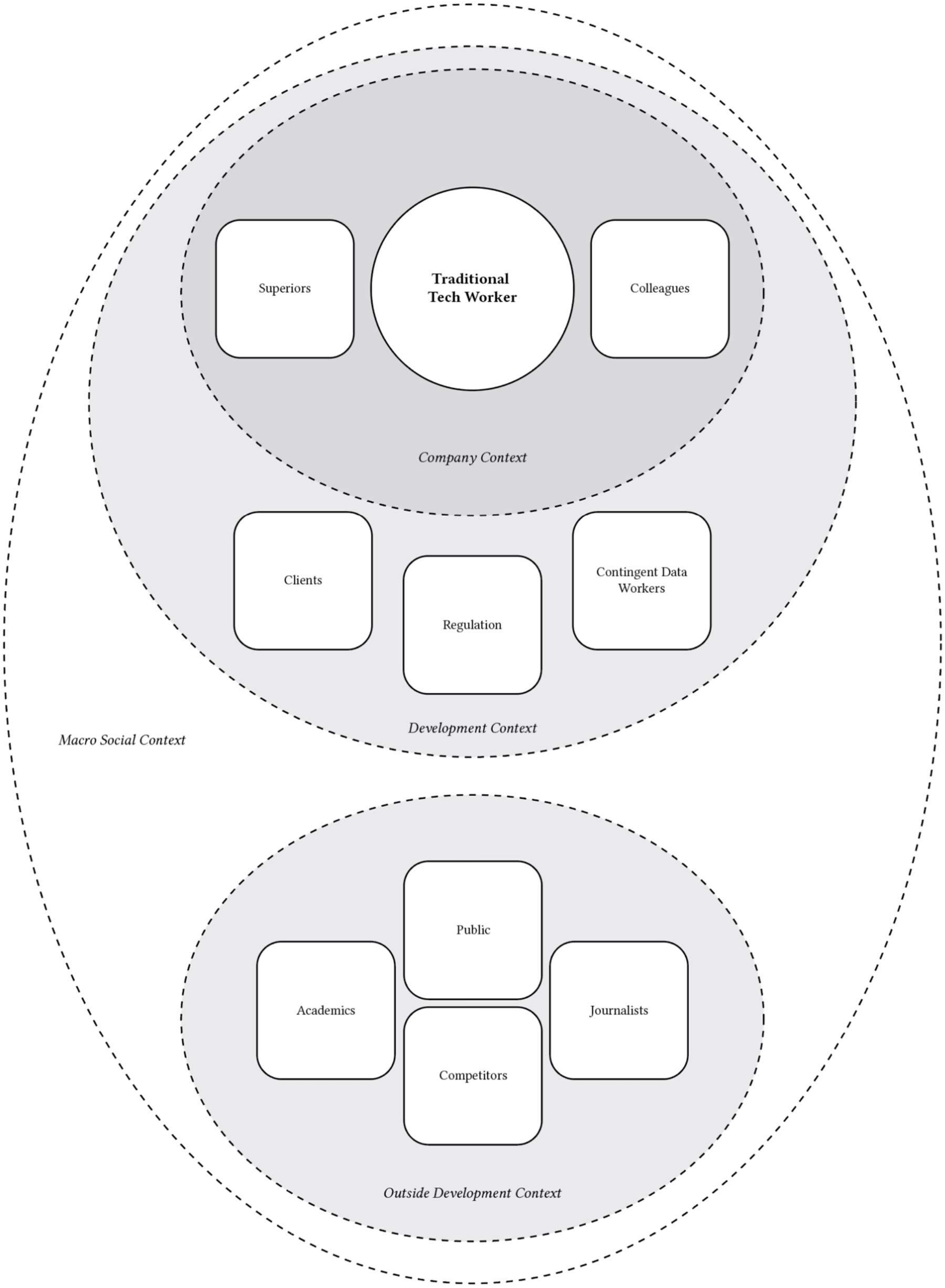


Figure 13. A figure depicting all of the different actors involved in developing computer vision. Each actor is placed in the relevant context it is involved in (e.g., academics are outside the development context, traditional tech workers are in the company context).

Individual workers are embedded within a specific **company context**. Both the economic power and the values of their company impact the approaches workers can take to identity. In some cases, companies also have specific policies governing their approach to developing identity concepts in computer vision. Workers in smaller companies with less economic power often have limited access to resources, such as teams dedicated to assessing the ethics of a project or the ability to collect robust and diverse datasets. Further, such companies are often driven by market demand more so than ethical policies. Small companies like MultiplAI adopt a market-first approach, serving client demands first and foremost and thus deprioritizing more nuanced approaches to identity concepts like gender. Meanwhile, larger companies like Aqueous have core policies driving their approaches to AI, including computer vision. They have teams dedicated to ensuring the fairness of products and whether products align with company policy. They even dedicate resources to overhauling identity concepts as outside perception about them changed. For example, how Kaleigh is overseeing updating gender in Aqueous' core computer vision product.

The relationship that individual workers have with their company also showcases their positional perspectives. Some workers expressed alignment with the values expressed by their company, indicating that they shared those values. These workers seemed to have a positive perspective on the type of work they could conduct within their companies, because their company likely valued their approach. On the other hand, some workers disagreed with the values of their company. Workers like Lynn did not feel empowered to approach development the way she felt it should be done. She disagreed with the company's market-first approach and desired more nuanced, careful, contextual approaches to identity in computer vision. Even while she expressed relatively reductive beliefs about gender herself, she expressed pride in her trans colleagues who pushed back on the initial representation of gender in their demographics model.

The contrast between Lynn's description of gender classification and her description of her colleagues also highlighted her own positional perspectives. To Lynn, biological sex was still always evident on the face. It is possible that her colleagues held very different perspectives on gender. Within a company, workers might have differential views from the others they are working with—both their colleagues and their superiors. Choices are also not made by one person, but by numerous people, with varying positional perspectives. Traditional workers need to negotiate their own positional perspectives with those of their colleagues. As such, identity in a product may shift and morph as it takes on numerous perspectives during the development process. For those with very different views, whoever is given the most decision-making power is likely to be most influential. At Aqueous, Kaleigh, in her position as a manager, was given power to veto product team decisions. However, in other cases, product teams or engineering teams might have more power to make final decisions.

Often, the **development context** was larger than a singular company. In some cases, clients, regulations, and data workers were involved in the process of developing a product. Clients bring their own positions to the table through their demand for specific features when they hire a company to develop a product. Once more, this was evident in MultiplAI's early clients demanding gender classification for marketing purposes; this led MultilplAI's initial model having gender embedded as a feature from the company's infancy. In other cases, a product that was developed years ago was relied upon by clients. Even though Aqueous was interested in updating gender in their models, they also had clients that had been using it for years. Some product teams would utilize client reliance as a reason to cling to older models of identity classification that they were attached to as initial developers.

While computer vision is subject to little federal regulation, some use cases of computer vision products necessitated compliance with federal laws. For example, Resoom's uses computer vision for assessing job candidate interviews. Thus, the model is subject to the U.S. Equal Employment Opportunity Commission's regulations. Regulatory requirements govern the types of identity groups workers must attend to in the design of computer vision. It necessitates that workers design for specific categories and ensure some level of fairness for those categories. For example, the EEOC provides a list of the "minimum" categories which must be attended to for "race/ethnicity": "White; Black or African American; Hispanic or Latino; American Indian or Alaska Native; Asian; and Native Hawaiian or Other Pacific Islander" (*Introduction to Race and Ethnic (Hispanic Origin) Data for the Census 2000 Special EEO File*, n.d.). While workers might push to attend to race beyond these six categories, they are not required to do so. Meanwhile, the gender categories as of this study were only "male" and "female," which were the only categories Resoom attended to in their model. However, the EEOC plans to add "non-binary" to its list of gender categories, likely forcing Resoom to include "non-binary" gender options in their fairness mitigation strategies. Regulation has the ability to force workers to attend to identity in ways that otherwise would not, but also has the potential to limit whether workers attend to categories beyond necessity.

Further, many companies will hire data workers for their data needs, prior to being able to train and evaluate models. The positionalities of those who curated and labeled the data for training models add another layer of complexity to defining identity. EnVision Data is an example of a company that provides data services for computer vision. While traditional tech workers define identity concepts early on during the development process and constrain data worker positionalities via instruction guidelines (see Chapter Three), data workers also introduce their own positional perspectives on identity while conducting data work (see Chapter One). Traditional workers act to control for data worker positionalities, attempting to maintain their own perspectives. Often, these interactions further expose how traditional workers impose their positionalities on identity concepts. As they come into contact with data workers who view identity differently, especially when they live in a different social context, traditional workers attempt to reorient data workers to their positional worldviews.

Beyond the development context itself, many traditional workers are attuned to those **outside the development context.** They were aware of ongoing conversations about identity in computer vision among the public, the press, academics, and their competitors. The press and the public were often viewed as sources of contention. In some cases, participants viewed public outcry or poor press coverage as a lesson for what not to do. For example, Lynn described learning from critical PR coverage of big tech when she was otherwise unsure how to proceed. In other cases, workers seemed to denounce the perspectives of the public. For example, Coleman spoke sarcastically about public outcry towards model mistakes, even if they were opportunities to identify issues otherwise unseen. Academia was also viewed as a resource for workers who didn't want to rely solely on their own intuition but wanted to implement best practices.

Overarching the development of computer vision itself, workers are always influenced by the **macro social context** in which they are embedded. Workers brought to the table opinions and beliefs shaped by the institutions they had learned from and the society in which they were embedded. Many participants expressed beliefs that were learned from their education, before they came to industry. For example, Coleman expressed having an internal desire to “toss [data] into the data grinder and build models.” This approach stems from how he learned to conduct research. Many disagreements seemed to stem from different disciplinary training which led workers to fulfill different roles. Workers in more social scientist roles expressed a dislike for approaches that prioritized engineering goals.

Further, participants acknowledged that identity concepts were culturally contingent; they could differ across cultures. Lydia felt that her clients in Japan didn’t care about sexist data, but it is also possible that views on sexism simply differ between the U.S., where Lydia is based, and Japan. As all participants came from Western countries in the Global North, their approach to identity was informed by what was culturally familiar to them through their socialization.

As I’ve broken down through this discussion, individual traditional tech workers occupy specific positions, informed by their own personal experiences and identity affinities. However, they do not operate in a silo. Individuals are not solely responsible for the development of identity in computer vision. Rather, they negotiate their own positional perspectives on identity with various other actors and concepts across four contexts: the company context they are embedded in, the development context the product is being developed in, outside the development context where others weigh in on computer vision, and the larger macro social context in which all of this work is embedded. These different contexts, with all of their different actors, act as a cyclical feedback loop, informing the positional worldviews of those involved.

Attending to Positionality in Computer Vision

The influence of positionality on the development of computer vision is unavoidable; it is not an issue that needs to be solved. Given that humans are involved in the design process of identity characteristics, they will always bring their own perspectives to the table—perspectives which are influenced by the contexts in which they are situated. Like a snake eating its own tail, workers are constantly influenced by context as they themselves influence those contexts. Rather than attempt to “solve” positionality, viewing it as a subjectivity which should be stripped for the sake of objectivity, practitioners and researchers alike can explicitly attend to it. The goal would be to identify and attend to gaps before they become unforeseen outcomes. I propose attending to positionality from two perspectives: (1) attending to context and (2) attending to actors within those contexts. Both perspectives are not mutually exclusive; they can be attended to in tandem, or in relationship with one another.

Attending to Contexts

Given that contexts mutually influence one another, one can imagine starting at the highest-level context—the macro social context—or at the one most constrained to the development of the product—the company context. Understanding how worker positionalities are shaped and constrained by context can ground research on AI development in industrial context in the social, cultural, and material conditions of work. Companies might also consider adopting contextually informed approaches to improve identity practices and enable their workers to more explicitly contend with positionality, which has been otherwise implicit. I describe opportunities for both researchers and practitioners to examine how the macro social context and the company context affects identity development for computer vision.

Starting at the **macro social context** would provide opportunities to understand the way identity categories are constructed in society, and how those constructions influence the way traditional workers attend to identity in computer vision. Starting with the macro social context in which development is embedded opens up opportunities to understand more about how social categories of identity are structured before attempting to define them for technical systems like computer vision. Further, examining the social context of development might illuminate how categorical histories influence the way that traditional workers approach identity problems in computer vision. They might be taking identity categories for granted, treating them as given, while the categories themselves are socially contingent. For example, workers in the United States might be making decisions about identity that are untenable in other contexts, like India. Tech workers attempting to define identity for more constrained environments, like medical contexts, might fail to account for how domain experts use identity information. If workers first contend with the social context governing identity, their perspectives may become more grounded in the specific context of use rather than personal experience.

Each of these identity categories also has a social and political history attached to it. Examining the history of a category can reveal normative and prejudicial assumptions about the people grouped under a category. Examining social categories can also reveal which types of identity categories are perceived as rigid and which are not—for example, in some cultures, gender is viewed as rigid and binary, while in other cultures, gender is viewed as fluid and non-binary (Driskill, 2016; PBS, 2015; Singer, 2020; B. Vincent & Manzano, 2017). After examining the history of institutions like the U.S. Census, workers might choose to explore more community-grounded approaches to categories like gender, race, or ethnicity. They might assess what identity categories they have chosen to be rigid and why. Given critiques of computer vision reflecting narrow perspectives on identity (e.g., Chapters 3 and 4 and (Bennett et al., 2021; Hanna et al., 2019; Keyes, 2018)), contending with the macro social context of identity before development can expand the narrow positional worldviews of traditional workers. They might instead consider other ways of viewing the world.

One might also take a step down, to examine the **development context** of a specific computer vision product. Attending to the broader development context might mean examining how the role of regulation influences approaches to identity in technical artifacts, much like how the identity categories outlined by the EEOC influenced approaches at Resoom. It might also mean understanding how the current landscape of B2B businesses in the computer vision space has constructed status quo approaches to identity in the computer vision industry, potentially shaping the way individual workers are primed to think about identity problems in the field.

Starting with the **company context** would mean grounding understanding identity in the larger industrial context shaping the project. The companies that individuals work in heavily influence how workers can approach identity concepts. Economic conditions, internal policies, and company values influence whether workers could approach identity the way they desired to. Focusing on each of these three factors can reveal how company context shapes the positions workers occupy. For example, understanding the economic conditions of the company reveals the resources which are and are not available to workers. Workers in smaller companies are often unable to access the same resources, like having multiple researchers focused on developing best practices. Similarly, internal policies established by companies might provide justifications for certain approaches to work, while denying other potential approaches. Kaleigh, for example, would often rely on her company's policies to justify pushing for better approaches to gender in their computer vision model. Other companies might adopt policies which do the opposite, prioritizing, perhaps, the technical over the social. A lack of company policies might also mean workers rely more on management, intuition, or market incentives. Finally, company values might invite certain types of workers to succeed in their approaches more than others. Lynn and Kenny showcase this possibility. Lynn felt her company did not value the same things she valued and left the company. Meanwhile, Kenny seemed to embrace the same company's values of a market-driven approach in his role as the vice president of business; he prioritized bringing in clients over developing nuanced approaches to identity like Lynn. Not only does understanding the company context benefit research—and critique—by grounding it more acutely in the realities constraining and enabling certain workers, but companies can also benefit from understanding how their own company culture shapes development. Companies might also consider shifting priorities to better allocate resources or develop policies for scoping requirements for identity concepts, especially given changing legal landscapes and public perceptions around AI.

Finally, understanding the context **outside development** can reveal how those uninvolved in the direct development process of computer vision can still influence worker approaches. Many workers were aware of how the public, journalists, academics, and corporate competitors perceived identity in computer vision. Workers could often use these outside perceptions to influence their colleagues or managers. Further, companies often responded to these outside influences, creating or updating policies, reallocating budgets, and developing new company identities. Assessing how identity is being discussed outside of the development context can benefit researchers attempting to understand current practices in industry and can benefit companies trying to understand the broader conversations around identity.

Attending to Actors

A more precise method for attending to positionality in computer vision is to attend to the different actors involved in the development process. Attending to specific groups of workers is certainly not a new approach in HCI (e.g., Daigle, 2003; Ghode, 2019; Heikkilä et al., 2018; Meissner et al., 2022; Muller et al., 2019). The work at hand took this approach—examining how traditional workers expressed their positionalities during the development of computer vision to understand how positionality shaped how identity is embedded into computer vision artifacts. There is still further opportunity to engage with the ways worker positionality influences identity practices in technology development for both researchers and practitioners. Attending to different positional actors can further reveal how perspectives shape identity outcomes in AI.

Much like this study, one might consider grounding understanding in **worker** positionalities. Beyond broadly understanding the role of positionality in identity development, there are still many opportunities to create better practices for documenting and attending to positionality. For example, one might center the positional perspectives of workers to better develop policies for explicitly engaging with worker positionality during development. In this study, the process of defining identity for computer vision projects was not explicitly part of development approaches. Workers did not explicitly engage with how their own positionalities influenced the way they perceived or implemented identity in computer vision. Given different types of workers have different perspectives, it could be fruitful to focus in depth on how specific types of workers reason about identity. One might compare researcher approaches with that of engineers, for example. One might also choose to examine the role of management in defining identity, to determine whether and how often identity comes from a bottom up or a top-down perspective in industrial contexts.

Another opportunity for deepening the understanding of how traditional worker positionality is influential in developing computer vision is to understand how *other actors* influence the perspectives of core traditional workers. To do this, one might examine the perspectives of actors within the development context but outside the core company context—like clients, data workers, and regulators.

Clients, those who request computer vision services from other companies, have their own expectations about identity. Understanding the positionalities of clients can also reveal why identity in computer vision products is designed the way it is and may reveal points of intervention for shifting design practices. For example, Kenny explained that marketing clients drove the use of discrete identity categories like binary gender early on at his company. Talking with client representatives in marketing contexts can reveal what worldviews drive their desires for such discrete categories.

Much like clients, there is also further opportunity to understand the positionalities of **data workers** who provide data services for computer vision. Data workers, as underpaid and largely invisible in the development of AI (Gray & Siddharth, 2019; Miceli et al., 2020; Miceli & Posada, 2022), are still crucial to its development. As scholars increasingly examine the ways data workers are disempowered in the development of AI, they might also examine the ways data workers make decisions about identity in their work. Data workers, as they interface with traditional workers, might influence the way traditional workers consider identity in data.

Finally, understanding the role of **regulation** and its implications for identity in AI can reveal alignments and gaps between those directly within the company context and those outside it. Understanding the perspectives of policymakers and their interactions with tech companies in designing policy can reveal their values, experiences, and perspectives. For example, do council members at the EEOC even consider how their decisions influence identity categories in AI? When creating facial recognition laws, what positional perspectives are policymakers bringing to the table? Given regulation influences and constrains the way traditional workers engage with identity categories, expanding understanding of identity development beyond solely product can paint a richer picture of the many positional worldviews influencing identity in computer vision.

One might also consider engaging the positionalities of those outside of the development context of computer vision altogether. Given products are deployed and impact actors outside of their development, understanding the role of the **public, journalists,** and **academics** might reveal different positional perspectives on identity than what traditional workers exhibit. Further, examining the relationships between outside actors and how they communicate with those in a development context can further ground knowledge on how traditional workers perceive and are shaped by outside actors. One might also consider examining the differential approaches between one company and their **competitors**, to try to understand how they influence one another's outlooks and mutually construct worldviews.

Attending to actors within different contexts provides both researchers and practitioners with a number of opportunities to better understand the role of human positionality in computer vision development. It can help to identify positional gaps, before they become unforeseen and undesirable outcomes when products are deployed. Such knowledge can lend to hiring decisions within companies, better practices for attending to positionality and documenting identity decisions during development, and more contextually informed research that extends beyond simple but untenable recommendations focused on improving industry practice.

Conclusion

All individuals have their own positionality, the perspective that they hold as a result of their own identities and interactions with others and the world around them. In the development of computer vision, positionality is critical to how workers approach defining and implementing identity. In this Chapter, I showed how the practices of traditional tech workers—like researchers and engineers—reflect positionality. Not only do their practices reflect their own personal values and experiences, but they also show when workers have differential worldviews and must negotiate them with their colleagues. Further, I show how the contexts in which workers are embedded shape their positional approaches to computer vision, sometimes enabling and sometimes constraining their perspectives. Finally, I showed how positional gaps within tech workforces can lead to unforeseen outcomes around identity issues in products. Workers, acknowledging their own limitations, advocate for more diverse workforces that they can use as resources for improving identity approaches within their companies.

I discuss how examining the positionality of tech workers reveals that workers are influenced by the many contexts they inhabit and are surrounded by during development. Further, I discuss how positionality is mutually constructed and constantly evolving due to the many actors in these contexts. Positionality, as a subjective and value-laden reality, is not an issue to be solved. Instead, it offers opportunities for explicit critical engagement so that researchers and practitioners can attend to positional gaps before they become undesirable and even offensive outcomes. I thus propose implications for more deeply engaging with positionality across contexts and actors in the field of computer vision development.

8
HOW DATA WORKERS IMPLEMENT IDENTITY

As I have previously described in Chapter 2, as machine learning technologies, computer vision models rely on data to “learn” what to predict. While some models may be trained using unsupervised techniques (e.g., Chang et al., 2018), most modeling is done using human-curated datasets. Individuals collect, clean, and label visual data, like images and videos, to train and evaluate computer vision models. This 21st century model of work has grown into a global labor sector coined *data work*. Big tech and startup technology companies alike hire huge swaths of *data workers* to produce datasets for machine learning. These data workers are largely hired from countries in the Global South, lack the benefits and protections of traditional labor, and are often subject to incredibly low wages (Birhane, 2020; Perrigo, 2022a; Ramnani, 2022). Yet, given pervasive issues with inequitable computer vision model performances (e.g., Barr, 2015; Buolamwini & Gebru, 2018; L. A. Hendricks et al., 2018; Zhao et al., 2017), industry practitioners and researchers alike have sought to address bias at the data level, adopting a “garbage in, garbage out” perspective (Stuart Geiger et al., 2020). One salient approach to dealing with data bias is to attempt to control the potential biases introduced by data workers—through methods such as instruction manuals (as seen in (Miceli & Posada, 2022)), consensus (Z. Wang et al., 2020), inter-annotator reliability (Davani et al., 2021), and performance testing (Geva et al., 2019). Despite being incredibly undervalued, data work is crucial to both enabling computer vision work and ensuring that work is fair and ethical. Beyond monetarily undervaluing data workers’ contributions to computer vision, data workers’ perspectives are seen as a liability that needs to be carefully controlled (Sambasivan & Veeraraghavan, 2022).

Beyond implicitly acknowledging that data work is imbued with a sense of human subjectivity by trying to control it, few attend directly to how the subjective nature of data work shapes computer vision.

Data workers, like all workers, occupy a specific position in the world as they conduct their work. Positionality—how values, experiences, social identities, politics, time, and space shape how one understands the world—tangibly affects all work. While there have been increasing calls for attending to how positionality shapes machine learning (e.g., Cambo & Gergle, 2022), work to date has largely focused on the marginalization of data workers and the power imbalances between data workers and their employers (Gray & Siddharth, 2019; Miceli & Posada, 2022). There has been little knowledge produced about how worker positionality operates in computer vision beyond an inconvenient and risky reality that should be carefully controlled. This work focuses on better understanding how the positionalities of data workers influence data work and the implications that influence has for computer vision.

The role of positionality in research has recently become more explicitly acknowledged in social computing, with researchers attempting to outline how their own positionality might have influenced their work (Liang et al., 2021). Understanding the role of the positionalities of *others* has been largely implicit in social computing, as we present findings from the perspectives of participants. Yet understanding the positional perspectives of data workers can illuminate how identity is shaped by human values, experiences, and beliefs. This chapter is focused acutely on how the positions data workers occupy influence data work for computer vision. More specifically, I attend to the following research questions:

1. How do data workers' approach to different types of data work for computer vision and what do these approaches communicate about their positionalities?
2. How do data worker positionalities influence the outcomes of data work?
3. What, if any, tensions arise between data worker positionalities and the data work they are assigned?

In this work, I employed ethnographic observations and interviews to unearth more explicit understandings of worker positionalities. I conducted a year-long ethnographic study with a small business process outsourcing (BPO) company in Europe, where I observed and interviewed data workers and analyzed documentation. I also conducted interviews with freelance data workers providing services on the platform Upwork. Interviews were designed to understand how data workers reason through computer vision data work that is salient to human life—including both human-centric (e.g., images of people) and human-adjacent (e.g., images of clothing) data. I asked participants to describe how they approach annotating and collecting data, what is challenging about their work, and their perceptions about diversity in data work. I thematically analyzed data around the concept of positionality, identifying often implicit explanations of data worker positionality and how those perspectives influence data outcomes.

Through the Findings of this chapter, I showcase the various ways that data worker positionalities influence their approach to data work, including the factors data workers felt shaped their perspectives on the world, like cultural familiarity and media. I show that data workers rely on tacit knowledge—an implicit knowledge gained through one's personal life experiences—when conducting identity work for computer vision. I also describe some of the unintended negative outcomes that occur when data worker positionalities do not align with client expectations of data outcomes and are unable to be captured by simple bias control mechanisms. While positionality operates as a form of tacit knowledge in data work, data workers are also not entirely unaware that their own perspectives may be limited. As such, they advocate for diverse workforces who can provide different subjective perspectives and experiences to make up for gaps in knowledge or understanding.

I discuss how these Findings illuminate gaps in the positional knowledge of data workers and also their clients, who fail to realize data workers act on their tacit knowledge about identity. I highlight how positionality does not exist solely within the confines of an individual but is negotiated across a web of positional actors that influence data work—including clients, trainers, supervisors, other data workers, and data instances themselves. I then describe how current bias mitigation practices fail to account for positionality, implicitly adopting a positivist worldview that prioritizes "correct" versus "incorrect" classifications. Finally, I propose positional (il)legibility as an approach to data work that explicitly embraces positional perspectives. I argue that certain data is legible or illegible to positional actors based on their own positional worldviews. Positional (il)legibility offers promising opportunities for actively and explicitly accounting for positionality in data work.

Participants and Analysis

This chapter focuses specifically on the perspectives of data workers. As such, I only analyzed the data I collected from data worker participants (see **Table 13**). Much like I described in Chapter 7, I specifically analyzed data from data worker participants around themes of positionality—descriptions and expressions of worker perspectives and how they applied them to their data work.

Data Worker Participants

Alias	***Source***	***Employer***	***Role***	***Country***
Yasmin	EnVision Data	EnVision Data	Annotator, project supervisor	Bulgaria
Ghaliyah	EnVision Data	EnVision Data	Annotator, trainer	Bulgaria
Dinorah	EnVision Data	EnVision Data	Project supervisor	Bulgaria
Aakrama	EnVision Data	EnVision Data	Annotator	Bulgaria
Abyar	EnVision Data	EnVision Data	Annotator	Bulgaria
Wares	EnVision Data	EnVision Data	Annotator	Afghanistan
Sumbul	EnVision Data	EnVision Data	Annotator	Afghanistan
Shokouh	EnVision Data	EnVision Data	Annotator	Afghanistan
Sadham	EnVision Data	EnVision Data	Annotator	Lebanon
Raiha	EnVision Data	EnVision Data	Annotator	Lebanon
Makaarim	EnVision Data	EnVision Data	Annotator	Lebanon
Hijrat	EnVision Data	EnVision Data	Annotator	Lebanon
Baksish	EnVision Data	EnVision Data	Annotator	Lebanon
Azyan	EnVision Data	EnVision Data	Annotator	Lebanon
Jaako	EnVision Data	Freelance	Collector	Kenya
Rebecca	EnVision Data	Freelance	Collector	Philippines
Thanh	EnVision Data	Freelance	Collector	Vietnam
Manjola	EnVision Data	Freelance	Collector	Albania
Lyonis	Upwork	Freelance	Annotator, collector, trainer	Uganda
Pelumi	Upwork	Freelance	Annotator, collector	Uganda
Malik	Upwork	Freelance	Annotator, collector, supervisor	United States
Sadhil	Upwork	Freelance	Annotator	India
Nedeljko	Upwork	Freelance	Annotator	Serbia
Gemma	Upwork	Freelance	Annotator, collector	Kenya
Raines	Upwork	Freelance	Annotator	Russia
Bernardita	Upwork	Freelance	Annotator, trainer	El Salvador
Lucano	Upwork	Freelance	Annotator, collector	Venezuela

Table 13. A table describing data worker participants in this chapter. The source column refers to where the participants were recruited from. The employer column refers to where the participant is employed. Freelance participants were recruited from Upwork but often worked as freelancers on a variety of platforms. Some freelance participants were recruited as part of field work with EnVision Data. The role column describes the various roles the worker described doing.

Next, I present Findings on how data workers—both collectors and annotators—went about conducting data work for the projects presented in this Section. Specifically, I attend to how data worker positionalities influenced their work practices.

Findings

Throughout the varieties of data work participants conducted, they regularly relied on tacit knowledge about identity, a knowledge they implicitly gained through their personal life and experiences.

Participants' descriptions of their work provided insight into how the various positionalities they inhabit influence their approach to their work. Given the nature of tacit knowledge as innate, informal, and intuitive, participants had difficulty formally and explicitly identifying how they made decisions about identity.

In this section, I argue that data workers inscribe tacit knowledge about identity into their data work and how that knowledge reflects certain positionalities. Findings show that data workers embed their own positional perspectives in both annotation and collection tasks. Generally, referring to one's own tacit knowledge is an implicit act; data workers were not explicitly aware how they were using their own positional perspectives to attend to their work. How they relied on positional knowledge was generally contextual to the specific task and specific categories they were attending to.

In the Findings section, I first describe instances where data workers implicitly refer to their own experiences and knowledge to conduct their work. I then describe tensions and barriers that arise when clients do not actively consider the positional differences of data workers, and thus data work does not match client expectations. After, describe instances where data workers explicitly acknowledge some concepts which might be unfamiliar to them, and thus difficult to work on. Finally, I highlight that workers still view diversity as a net positive to data work in computer vision.

How Worker Positionality Influences Data Work

Data workers, whether conducting data annotation or data generation projects, regularly relied on the tacit knowledge developed by their own positionalities—the synthesis of experiences, values, beliefs, affinities, and sociocultural context. Participants regularly relied on contextual familiarity they personally built through life experiences, localized values and norms, personal affinities with identity categories and values, and other characteristics that make up one's positionality. Such knowledge was approached as "common sense," an approach also reflected in the lack of explicit instructions or engagement with positional perspectives on behalf of clients.

When the workers found their subjects to be familiar to them in some way, they had a much easier time with annotation. In the case of tagging faces with emotion concepts for the Emovos project, participants seemed to have an easier time with actors that they recognized. Both Sumbul and Shokouh described having an easier time with movie stars *“from Hollywood, from Bollywood, and from [their] own country”* (Sumbul). Shokouh described how familiarity with actors and media made understanding their expressions easier for her: *“Some people, like the movie stars, it was more easier cuz we have seen that exact movie or series or show so we knew how was he or she [feeling at] that time, it made it easier.”* Based on my analysis, those who have seen the media in the dataset may tag emotions differently than those who have not, simply because they are already familiar with the emotional context of the scenes they are tagging. It is possible that, for example, an emotion like “disgust” is perceived as “sadness” if one doesn’t know the narrative informing a facial expression.

Sadham similarly discussed how he knew the differences between different races because of what he saw online and on television. In this case, familiarity and access to certain media, informed by larger market trends and international media exchange, made certain faces more or less accessible to data workers annotating the Emovos dataset. For those public figures or actors they were unfamiliar with, they could not refer to a prior contextual knowledge informed by a specific media context. Those who do not have access to media may be less familiar with the emotional context of the faces being annotated or the way racial categories are described. Neither familiarity nor unfamiliarity with the original media context of the dataset indicates more correct or incorrect answers. Instead, familiarity can be viewed as a positional lens that colors how data workers view the data in front of them.

Beyond familiarity with media contexts, data workers showcased certain culturally contextual perspectives of identity categories. Ghaliyah worked as an annotation supervisor on the ChAI project for interview video interpretation. Annotators on the project were asked to tag each person with a gender: male, female, or undefined. When asked how she made decisions about determining each person’s gender, she largely relied on cultural cues she was familiar with as a Muslim woman: *“Most of the people in the videos, they were from Islamic countries. For example, for the women, most of them were wearing hijab. … in those countries, the cultural uniforms are divided between men and women.”* While data workers were not given any guidelines for determining gender, she had tacit knowledge of how religious garb is typically divided between binary gender categories. *“It wasn’t very challenging based on the gender norms that we have based on the society, so we just said male and female,”* she explained.

Ghaliyah viewed something like classifying gender as something so obvious within her own culture that she never questioned labeling gender in her work practices.

There weren't always singular reliable indicators for identity categories like gender. Gender presentation can be multifaceted and complex. Sadham and Raiha both described that they were also aware that common "indicators" of gender were not always reliable, because gender presentation is not always static. In particular, they both referred to how men might have long hair and women might have short hair, or that men might have a beard or a shaved face. Sadham explained: *"We know what is the difference between men and women … I will figure out that this is a woman, yeah, of course."* In this case, annotators described familiarity with diverse presentations and a knowledge there was no "standard" case for defining gender. Yet, at the same time, gender was still "obvious" to them. Manjola similarly declared that gender, in comparison to other categories like age, is *"obvious ... I mean, it's just that I know that they are somehow."* Obviousness and common sense are reflective of a tacit knowledge about visually classifying gender common across all cultures. Workers could not explain what about it was "obvious."

Gemma expressed that trying to annotate gender was difficult for her, because she was aware that some people are transgender or do not fit into clear boxes of male or female. *"Actually for gender I find it a bit tricky because some people are transgender and they don't perceive they are transgender as another type of face, so they just put female and male … they just put it as female, male, so that's what they consider most of the time in such projects."* In this case, Gemma expressed a personal knowledge of transgender identities that was not expressed by other data workers. Such personal knowledge reflects an exposure and awareness unique to her own life experience, that other data workers may not have experienced. Because she felt she could not make a decision about gender in these cases, she would relay all edge cases to her supervisor to make the decision instead. In such cases, she did not have to make a labeling decision; her supervisor applied the label, and the data never came back to her to verify it.

On the other hand, while Gemma was concerned with gender classification because of the existence of transgender people, she had never questioned racial categories. She states that the naturalization of human categories meant she had rarely questioned other groups*: "People are divided* *into many categories. My concern was about the people who do not consider the transgender, but about the racial grouping, I've never even thought about it."* Gemma, who did a number of freelance projects where she had to classify everything from attentiveness to social groups, worked on a project where she was tasked with tagging faces with racial classifications. Much like the ChAI project, she did not get any instructions beyond classifying each face; the client did not provide any examples. *"You know a white man is white, and you know that Hindus wear this type of clothes, Middle East people, they wear hijabs. So, it's not something you have to ask the client each and every time how a Caucasian person looks, you just use your knowledge of what you know about people."*

Ghaliyah similarly relied on her own cultural familiarity to determine ethnicity for the ChAI project. Because she stated that the majority of applicants appeared to wear Islamic garb, she said that ethnicity was relatively simple to select; she would label the majority of the people in the videos as "South Asian (Indian-Pakistani-...)" given her perception of their appearance, religious garb, and language. On the other hand, she would often label those with darker skin tones as "African," because they did not necessarily align with her view of what counted as "South Asian." For these projects, both Ghaliyah and Gemma relied on locally contextual and experiential knowledge about racial categories to make determinations about identity concepts. The ChAI project also involved multiple annotators labeling the same data. In these projects, disagreements between annotator worldviews began to surface, showcasing how not all workers interpret the data the same way. For example, Ghaliyah labeled a data instance as "South Asian," while Yasmin labeled that same instance as "Black African."

Relying on tacit knowledge about identity also extended beyond annotating identity concepts. Rebecca, a freelance data worker located in the Philippines, said that she often translated the instructions for EnVision Data's SensEyes data collection project from English into Filipino. Although most of the people she contacted for data read English, she said they'd feel more comfortable with their own language, and the message would feel "*more personalized*." English would be more likely to trigger distrustful responses, because English is often used in scams or surveys. In such cases, Rebecca made the conscious decision to translate her clients' instructions because she had experiential knowledge about how people in the Philippines perceived English-language requests as scams. She accessed that knowledge in order to build trust with potential data subjects.

Data workers relied on their positionality to make decisions about both data annotation and data collection, both human-centric and object-centric data. The tacit knowledge informing data work varied by each data worker's positionality. As showcased by the examples above, some data workers had differential positionalities that lent them unique perspectives, such as the difference between Sadham and Gemma's ideas about gender categories. Even when explicit instructions do provide examples, like with the ChAI project and emotions, data workers must rely on their own tacit knowledge to make sense of the data in front of them. Further, group projects highlighted that each data worker had different positional perspectives, given different labeling decisions when labeling the same data. In the next section, I highlight some of the tensions that arise between client expectations and data worker positionalities, particularly given clients are not attending to data worker positionalities in guidelines and training.

Unintended Outcomes When Positionality is Unattended To

Clients rarely provided explicit expectations about identity concepts; such expectations were often implicit, by providing labels (e.g., "male" or "female," "South Asian" or "Black African") that evoked tacit knowledge about those identities. Given that clients provided little information on identity-based and culturally contextual expectations for their projects, data workers relying on their own experiences and intuitions did not always perform as expected. As participants relied on their own positionally-situated tacit knowledge to make decisions about data, they also often ran up against misunderstandings, introduced implicit biases, and engaged in unexpected practices to complete their work. In this section, I present a number of tensions that arose because clients did not adequately address positionality in guidelines and training, including: (1) misunderstanding client perspectives; (2) implicitly introducing social biases into annotation; and (3) navigating culturally contextual barriers to data collection. This section highlights the types of unintended consequences shaping computer vision data when clients do not consider how data worker positionalities might differ from their own.

Mistakes and Misunderstandings

Sometimes, the positionally informed interpretations data workers made seemed to clash with the expectations of Western-centric clients. Many data workers submitted work that was deemed by the clients as incorrect, but data workers explained was simply different in their own cultural context.

For example, workers had a difficult time when certain objects, like garments, differed greatly in their own sociocultural context from the expected context of clients. Sadhil, an annotator who worked for two and half years at a computer vision company before switching to freelancing on Upwork, described an instance of cultural confusion between himself and his client. Sadhil is based in India, and his client, who was based in Japan, requested annotated data for a clothing classifier. He says that in India, a blouse is a type of women's clothing that goes underneath a sari, and so when he was asked to collect images of blouses, he *"collected the images according to blouse that [I know] in India."* However, the client said this was not a blouse, and so he had to research what a blouse looked like in the context of Japan and found that it was completely different and was a much longer shirt-like garment. He had to shift his own view and build an understanding of what a blouse looked like in Japan, so that he could annotate in a way that his client deemed correct.

Some misunderstandings also had to do with differing environments. Bernardita also had difficulty with annotating cars in North America. She described annotating large buses in Canada as trains. *"As I work with different cultures, I am not familiar with all of that … everything is really different, and when I saw the first time a bus in Canada, I thought it was a train, so it was really confusing for me when he [the client] said, 'No, this is a bus, this is not a train.'"* Malik explained a similar misunderstanding when it came to mailboxes in North America. Mailboxes were often incorrectly annotated because annotators he worked with in the Global South weren't used to novelty mailboxes, like fish-shaped mailboxes, in the United States. Annotation guidelines did not explain how objects and environments could differ depending on locale. Therefore, clients regularly approached data workers to explain their own cultural expectations in cases where they received labeled data that did not match their expectations. Such moments represent an unintentional and implicit exchange of cultural ideas, where data workers then adopt and change their own perspectives on the world to accommodate client expectations.

Misunderstandings and confusions also occurred when data workers had differential perceptions about identity categories than clients. For example, Gemma, who as already mentioned, was attentive to queer identities, described being unsure how to annotate gender in cases where she felt gender was ambiguous:

> *"Sometimes, it's really hard to tell someone's gender because of their sexuality, someone might apply makeup or is transgender, so then you can't know really the gender of that person, so in that situation, mostly I just ask the client [what they believe the gender should be], cuz that's something you can't just tell."*

Gemma also described being unable to choose labels between multiple racial categories:

> *"You cannot differentiate between a Caucasian person and Hispanic person, so it was sometimes a bit challenging … If I found an image I don't understand then I can just call the manager or the supervisor to ask them about it … If they can't get it, they forward it to the client."*

Given data work serves client needs, Gemma would shift hard decisions to the client, so they could ensure the data matched their own needs and perspectives. Instances like these also highlight that people perceive gender and race differently and may not always be aware of those perceptions when applying labels.

Many data workers referred to these instances of misunderstanding or differential interpretation as "mistakes," even in cases where they might be technically correct in their own local context or there may be no way to ascertain a correct answer, like with gender and race. Misunderstandings—framed as mistakes—highlight the shortcomings of guidelines and trainings that are presented as neutral or technical. Misunderstandings become cultural boundary objects from which clients begin to attend to data worker positionalities in the form of corrections. Mistakes may also highlight for clients the limitations of their own positionalities, revealing realities they had never thought to include in the instructions. However, misunderstandings often surface during client review processes; clients notice when large swaths of data do not match their expectations. There are other cases when misunderstandings are so implicit that they become invisible. In the next section, I show how data workers would often embed their own social biases unintentionally in their approach to data work.

Implicit Social Biases

Beyond instances where data workers described being uncertain about the cultural or identity-based labels they were applying, some data workers showcased implicit social biases when trying to describe annotation difficulties. They did not describe their perspectives as biases explicitly, but instead demonstrated having more difficulty interpreting certain gender, race, and age groups. In some cases, their explanations of these difficulties reflected common social biases that permeate across cultural contexts (e.g., anti-Blackness, anti-Asianness) and are commonly reflected in AI.

Most data workers did not express finding annotating expressions any more difficult for men or women, besides one: Sumbul. Sumbul described having more difficulty understanding the emotions of women in comparison to men when annotating the Emovos project. She said, *"Woman are mostly difficult to know in which mood they are than men. For example, we saw the disgust and fear expressions mostly in woman, not in men. In women, it was a bit difficult to know in which mood they are."* As an example of gender bias, it is interesting that Sumbul had a harder time identifying the emotions of people of the same gender identity; while she could not describe what made it more difficult in concrete terms, it is likely that internalized perspectives of women influenced her perspective that women's emotions are more difficult to read.

Sumbul, who is relatively young, also described having more difficulty understanding the emotions of younger people*. "Because we know that younger people are more excited and they show other expressions as well, but the younger was a bit more difficult than the olders."* In this case, she ascribes a wider range of diversity of emotions to younger people, which means she both has a harder time annotating them and is viewing older people as portraying more simplistic expressions. The age bias she demonstrated may impact annotations for both young and old faces. In collecting different selfies of people on the street, Manjola, though she claimed age was "easy" to tell because older people have "*wrinkles*," still could not identify specific age brackets just by looking at people's faces.

The majority of biases that data workers exhibited were racial biases. In particular, data workers living and working in homogenous cultural contexts expressed having difficulty annotating certain features on certain ethnic groups of people. Bernardita, Raiha, Sadham, Lyonis, Pelumi, Shokouh, and Wares all described having difficulty annotating certain racial groups, across multiple different project types. In particular, participants across multiple countries described having difficulties with Black and East Asian individuals specifically. Wares stated, *"Different people, for example, Chinese, Japanese, Korean, and African … the difference between them are a lot greater than between male and female."*

Bernardita described accidentally annotating different people as the same in a facial recognition project. She attributed her mistakes to an implicit racial bias, which made her unable to distinguish the difference between different Black individuals. Wares and Shokouh, were both Afghan-based workers hired by EnVision Data for the Emovos project focused on assigning emotion to different faces. Both expressed having a difficult time interpreting the emotions of certain groups. Shokouh described to me that "Africans" were particularly difficult for her, saying that:

> *"Tagging the expression of sad and angry was a little bit hard … Sometimes we thought that he or she was angry, but it was not like that. The type of their face was like that, so we had to look closely … [Africans] looked serious and also … the type of their eyebrows, it was a little hard to recognize."*

Lyonis portrayed these difficulties as if they were more technical than social, stating: *"There are some races, which definitely will be very challenging just from a general point of view."* Lyonis, Pelumi, and Raiha all had difficulty annotating East Asian faces, particularly emotion and keypoint annotations around the eyes. *"Like you go to like China, or like to maybe Korea, there is a little bit it's a little bit hard to figure out the corners [of the eyes]"* (Lyonis). Though her intent was not malicious, Raiha described this difficulty in ways that reflected common racist descriptors about East Asian eyes, saying she had a hard time because they are not *"completely open*." She felt that guidelines specific to East Asian faces would help her improve her annotations.

Positionality represents how a person fits into a metaphorical space; how one's position enables them to more closely relate to or understand those like them, while disabling their ability to relate to or understand those unlike them. The implicit biases held by annotators against East Asian populations might result in less accurate keypoint annotations, while the biases annotators had about both African and East Asian populations might result in harmful stereotypes around emotion. While popular media was seen as a resource for data workers to ground their interpretation in specific contexts, it was also seen as a potential source of biases. Sadham expressed concern that social media and television could "give wrong ideas" about specific groups of people. He said that media he has encountered has promoted racist ideologies such as "*Black people stealing stuff*" and Asian people being "*full of disease*," especially during the current political climate around COVID-19. He was concerned that such ideas might be insidiously shaping data work practices. In the absence of explicit guidelines, data workers fell back on things like media to act as training guidelines. Clients seem unlikely to consider the ideas that data workers are exposed to about identity in media, or in other aspects of life. Insidious and unintentional racial biases were not accounted for in how clients approached or assessed the projects.

Many of the instances of social biases are particularly hard to ascertain, because the task and the bias are not necessarily the same. For example, in annotating emotion, many data workers displayed racial biases. It is difficult to notice the intersection of social bias and misunderstandings. Clients may be unable to discern less accurate keypoint annotations for East Asian faces or attributing more negative emotions to Black individuals. This difficulty could arise because those annotation differences are not glaringly "incorrect" in the way that labeling a bus a train. It could also arise because both client and annotator share similar implicit biases, especially considering Sadham's observation that identity beliefs may be learned from Western or global media. Also, while the biases data workers held themselves were implicit, they explicitly named the groups of people they had difficulty with. It is likely that other biases remain entirely invisible, simply because workers do not find annotating them difficult.

Differential Expectations for Data Collection

Thus far, the majority of the unintended outcomes in this Section have focused on how data workers imbue data with unexpected or biased positional perspectives. However, the positionalities data workers inhabit also impacted *how* data is *collected*. Data collectors approached data collection tasks dependent on their understanding of how the potential data subjects they were recruiting might treat them or react to data requests.

Some participants described barriers to data collection that were specific to their local context and cultural beliefs. Their knowledge of these specific beliefs gave them the ability to navigate and circumvent them to complete their data collection tasks. More specifically, Gemma, Jaako, and Manjola all described issues that they faced when collecting selfies for the SensEyes project.

Gemma and Jaako, both based in Kenya, described similar culturally contingent barriers to collecting face data. Both describe common religious concerns about face photographs being used by data collectors for "*black magic*" (Gemma) and "*devil worship*" (Jaako). Gemma explained that there were times asking participants for their face data that they became violent with her, because they did not trust her intentions. *"Some become violent when you're taking their pictures … especially [when] there are some projects like you have to take pictures of children,"* Gemma explained. Her tactic for trying to make collection easier and to avoid potentially dangerous situations was to offer a portion of her payments to participants.

Jaako, on the other hand, said he would lie to participants about what the images would be used for, saying that they were for a personal project:

> *"I told them that is my project, that I'm coming up with, that I'm creating an identity management system that could be used in future for the bank, and I needed their support for the videos. That's how I made them believe me, because if I told them it is something international, they wouldn't believe and they would have thought it was a … some kind of, devil worshiping thing, because in Africa, people are very primitive still … Most of them are not finished primary school, and I think that's the reason. They are primitive."*

Jaako, who expressed a positive perspective of technology and AI, in particular, described the culture in Kenya as "*primitive*" multiple times. He stated that anything involving the internet and international communities was seen as a potential scam, and that being scammed was quite common, especially for older people. *"In Kenya, older people have so much beliefs in their cultural values and they believe anything westernized, anything foreign, is either a scam or is satanic."* As an aspiring software engineer, his views seem to indicate an internalized westernization around technology. *"I come from a very humble background and talking to someone international has never been a dream in our community,"* he told me, describing how his involvement with Upwork projects is benefitting his education and work portfolio*. "Maybe when I get another project, this time I'll maybe explain to them exactly what I'm doing and I'll enlighten some of them."* His words indicate that a positive view of technology is more reasonable and that people in the community should not be concerned with donating their face data.

Much like Jaako, Manjola also decided to lie about the commercial purpose of the data she was collecting. She did so in order to ease the situation and avoid uncomfortable encounters with strangers, who might otherwise question her intentions. Further, Manjola described being wary of approaching strangers as a woman and allowed her older woman neighbor to accompany her. *"I find it more comfortable with women, but I think men were really nice too, like really available to help, and some of them were with their wives, so it was easier for me to communicate openly."* Further reflective of a broader social concern about women interacting with strangers on the street, Manjola said she did not tell her father she was doing this type of work, because he would be troubled by the idea of her going house to house to collect data. Lyonis, on the other hand, said that women were much more reluctant to talk to him, given that he was a strange man.

Both Gemma's and Jaaako's stories provided interesting insights into barriers to accessing certain populations ethically which Western companies might otherwise overlook. Gemma, who was honest with participants, faced potential violent backlash from people in her community. In order to make it safer and easier to collect data, she gave participants a portion of her earnings. Jaako was put in a position where, in order to complete his task and maintain a positive professional profile on Upwork, he felt it was necessary to lie to navigate cultural taboos he was personally familiar with. On the other hand, Manjola faced barriers specific to being a woman approaching strangers for their data and being unsure how they might react; Lyonis was less able to collect images of women because of their perception of him as a strange man. Manjola's and Lyonis' differential gendered experiences also highlighted that collection of certain genders might be more or less difficult depending on the gender and presentation of the data worker.

Data workers had in depth knowledge of cultural beliefs and local behaviors that clients and EnVision Data were entirely unaware of, reflecting a specific positional familiarity and expertise that those based in the Global North or Western regions were unprepared to account for. Such localized positional knowledge allowed the data workers to complete their work, which was necessary for their financial wellbeing. However, these examples also indicate that not all data collectors are able to easily access all populations very easily, solely because they are in a certain location. That could mean that the data collected is biased towards others they are more comfortable or able to approach. Further, clients were unable to predict or understand the cultural context of their data workers and how it might implicate the collection of their datasets. The selfies Jaako and Manjola collected for EnVision Data were collected in a way that could be deemed unethical, despite the BPO's mission to provide ethical data to its clients.

When data is collected, workers act as the connective tissue between client expectations and the cultural worlds that data subjects inhabit. Data collectors become aware of the context of collection, the culture data subjects are situated in, and how they might navigate data collection to meet client demands. To meet project goals, data collectors attempt to hone their tacit knowledge about culture to ensure collection is successful. Clients were unlikely to understand the contextual barriers to collecting data and how data workers might navigate those barriers. EnVision Data was entirely unaware of the cultural beliefs Jaako was collecting data in, or that he was lying to participants about their data use to complete the project.

Awareness of the Limitations of Personal Positionality

While thus far, the Findings of this work suggests that positionality is always implicit to data workers, there were also instances where data workers were aware that they held differential positionalities, Participants acknowledged that some concepts were unfamiliar to them and thus presented more difficulty to their work. Participants acknowledged several different positional characteristics that might contribute to poorer data work outcomes, including language barriers, inexperience with certain groups of people, and the desire to avoid causing harm due to a lack of knowledge or familiarity.

Given that the majority of participants serviced clients in the Global North, commonly seen as "Western" countries, most clients relayed instructions to data workers in English. Maakarim noted that this might cause gaps in fully understanding or interpreting instructions. In discussing the required ethical AI training module in particular, he said that the majority of his colleagues have *an "under intermediate level of English"* and that *"the command of English for the course is higher than they have*." He stated that even when subtitles were available, many workers would push themselves to understand English because it is "*the language of the environment*" (data work).

It became apparent in discussing the course with data workers that many of the concepts were being interpreted differently than intended. The course was aimed at teaching workers how to collect and annotate data "without bias," but the majority of workers viewed the term "bias" as directly related to the examples given in the course. For example, Azyan, Hijrat, Baksish, and Sadham all believed that the term "bias" was specifically about ensuring equal representation between men and women, as that was the example given in the course. Raiha was unable to explain her understanding of the term “bias.” Maakarim was the only participant who discussed the ethical AI training module who extrapolated the term “bias” beyond this singular example to more broadly discuss it as equity or fairness as a concept. It is likely that the term “bias” and how it is explained through examples in English did not translate conceptually for those workers whose strengths were in their first language, and not English. Beyond training, many concepts in implementing data collection and annotation tasks might be misinterpreted or confusing for annotators, because they are either underexplained, narrowly scoped, or conceptually untenable between languages.

However, despite ambiguity around the term “bias,” some data workers were aware of the potential for different biases to impact data work. They were concerned about annotators accidentally embedding their biases in the data. For example, Aakrama discussed his fear that bias would permeate how Arab people were assessed in the ChAI project.

> *“I think the culture is important, because … maybe you know, in Arab culture, people speak loud, and most of this video is about Arabic culture. When you don’t know, thinking this man is angry, but some people in Arabic country speak very loud … and when you don’t know this, maybe you make a mistake about this man, [like he] is angry or have a problem … And sometimes, in this project, because this project is in English language, sometimes people, because they don't have knowledge about language, cannot speak it. But if speaking native language, [they] don’t have a problem. We can’t decide … is it about language or about nervous[ness] or stress … because the person is trying to find some word to say but you might be thinking it is about nervous[ness] or stress. … In this project, we have this problem because we don't know about [the] Arabic language, but most of the people speak Arabic, and when you don’t have sense about what this man say, can’t decide correctly what the passion of the person is.” (Aakrama)*

Aakrama highlights concerns about potential implicit cultural biases leading to Arab speaking candidates in the ChAI project to be rated more poorly in desirable categories, like “*Would you invite this person for a job interview?*” It is possible that those annotators who do not understand speaking patterns or facial expressions commonly presented in Arabic cultures view individuals negatively. Further, Aakrama feels that crucial context for properly rating videos is lost when the annotator is unable to understand the language in the videos, or when the candidate is asked to speak in their non-dominant language. Gemma spoke to how collecting data was much more accessible to her due to shared language. “*So, I mostly collect data for Africans, because language is something different, when you speak to someone in your language they can understand you [more] easily than going to speak to* *someone who doesn't understand your language*," she explained. Concerns about annotator biases also highlight the interaction between annotator worldviews and data subjects. Certain characteristics may only be legible to annotators who identify a shared positionality with how they perceive the data subjects. Aakrama might find confidence in the speaking patterns of Arabic data subjects, while other non-Arabic annotators might find this confidence illegible to them. Gemma also takes on the role of shaping what data subjects look like in her work, opting out of collecting subjects unfamiliar to her.

Bernardita similarly acknowledged the subjective role annotators play in interpreting data subjects' identities. She explained that the majority of mistakes she sees annotators make is with people of color:

> *"Sometimes it happens to me, that I think they are the same person and I have to message my client and say can you help me, I think this is the same person. I think they are really similar, but sometimes because they are, how to say this, the color of the skin … Black people, I think it's a little bit more difficult for annotators to identify them … [they] look almost the same. We are not used to see those kind of people."*

In this statement, she is openly acknowledging that she specifically has a difficult time distinguishing between different people in images or videos when they are Black. She attributes this difficulty to underexposure to Black individuals in the area that she lives in, evoking an other-race effect (see (Anzures et al., 2013)). Gemma stated that she only collects images of Africans, because she doesn't have *"access [to] a lot of Caucasian people … other people can do [that collection]."* Gemma actively limits her role as a data collector to data subjects who are accessible and familiar to her. Gemma provides an example of how a data worker might attempt to limit their own biases in their work. Yet, such personal responsibility fundamentally requires that a data worker is aware of their own positional limitations—and that clients are pliant to such decisions.

Maakarim felt that he could attempt to close positional gaps by *"putting [himself] in another's shoes"* and trying to do more research on certain identities:

> *"I do know the LGBT, there are multiple genders … I'm not quite knowledgeable in this area, so maybe I have to make my research before I speak about it. I do because maybe sometimes I do not have enough context or knowledge knowing which label will make this person anxious or feel bad if he saw it somewhere or affect his life. I might be concerned if I have doubt about the label, even if I was under supervision."*

In this example, Maakarim starts with a base knowledge of LGBT and gender expansive identities but is aware that his knowledge is minimal. He describes feeling "*concerned*" because he might accidentally label data in a way which is offensive or causes harm to LGBT populations. Research in this case is an attempt to close the gaps caused by his own lack of experience with LGBT populations, while also realizing he still might not feel comfortable annotating such data. Further, tensions arise between Maakarim's concerns about what a potential data subject might feel about his interpretation of their appearance, but his work is actually focused on providing data to an organization who might otherwise not be concerned with the data subject's feelings. Maakarim has chosen to prioritize, at least on an emotional level, the potential harm his individual interpretations could cause to a data subject, even if his work—and income—are predicated on a client's desires and beliefs.

The limitations of interpretation were not pressing for all participants, like they were for Maakarim, but they were pressing for some. Much like Maakarim acknowledged the limitations of his own interpretation of gender, Sadhil, Malik, Bernardita, and Sadham all acknowledged that while some concepts may be the same across countries, many concepts differ from country to country. Sadhil said, *"Same in any country, if you are working on human fall detection, it's the same in all countries. But when the thing is for the clothing, it can be different in different countries."* Malik, a freelancer based in the United States, similarly commented on how certain clients might seek data workers from specific cultures due to these differences:

> *"I had a meeting with someone who worked at LabelBox, so she used to manage the labeling there, and she told me that … some tasks that require some knowledge of fashion, like clothes and this kind of stuff, they look for Americans as well because, like the culture, like, it's important to look at it from an American perspective."*

Bernardita described how cultural differences can cause confusion for data workers. *"When we are talking about culture, we can say that it's a little bit confusing to do our work, because they have different things, … and we can consider both things, can be the same, but we know them as different things."* Sadhil, Malik, and Bernardita—as well as the woman at LabelBox Malik is referring to—actively acknowledge that culture plays a role in how certain concepts are labeled for computer vision tasks.

Wares summarizes how annotation is not an objective process and that "ground truth" is imbued with specific values. He described an instance on the Emovos project where he had disagreed with his supervisor on what was the "correct" label, specifically because he believed that there was no "correct" answer. *"It's about recognition from every person differently. When we look to someone, we select the* *things that we get from their face … but maybe, another person gets something else. … we select the expression we think [is] right the most."* Wares, who himself described difficulty annotating certain groups of people, was explicitly aware that each person saw each image differently, and thus each annotation was more a reflection of the annotators' own worldview. Sadham felt that this diversity of worldviews was the most difficult aspect of data work. He explained*, "Everyone, he will have his own idea, so we have to be aware about it, and our supervisor also teach us about … wrong ideas. So, we have to contact each other and report about any problem."* He believed that clearer communication between data workers and also with supervisors would mitigate issues caused by disagreements about data work.

As demonstrated throughout the Findings, the gulfs between client and data worker positionalities were a major source of tension, misunderstanding, and unnoticed bias. Gaps between different data workers also led to differential approaches to data annotation and collection and to concerns about limited worldviews and implicit bias shaping the data. Homogeneity in data work teams also led to more limited worldviews about concepts like identity or cultural context. Therefore, as I present in the next section, participants still felt that diverse worldviews were beneficial to data work.

The Benefits of Diverse Positionalities in Data Work

Despite some of the concerns in Awareness of the Limitations of Personal Positionality, many participants expressed a belief that diverse perspectives and experiences can improve the outcomes of data work. Data workers believed that the diverse positionalities of their peers helped to educate them and expand their own worldviews, and also helped to close knowledge gaps in the annotation process. As Abyar commented, *"The more we work, the more we know."* Work experience, especially with diverse colleagues, can make data workers more effective and thoughtful in their approach to their work. As workers are exposed to different perspectives and ways of viewing the world, they develop new tacit knowledge that influences their approach to their work.

Raiha expressed that having a diverse team made datasets not only more accurate, but also helped to avoid biases: *"When we collect the dataset we need people from different nationalities, different races, … maybe gender also… in order to collect a good and accurate dataset."* Raiha attributed the diversity of worldviews from the types of diverse positions a data worker might occupy as a boon. Despite the different opinions workers might bring to the table, Raiha felt the more diverse a workforce is, the more accurate the annotation would be. She also felt that the more voices there were in the room, the more likely it was that a dataset would avoid undesirable outcomes, like racist views.

It is possible that some might view more diversity as potentially negative, as well. Maakarim commented directly on the notion that diversity could be negative or cause less accurate data work. He stated: *"I do not think diversity is a bad factor or corruptive but is completely the opposite … the cooperation between those different backgrounds will help us solve many problems."* Much like Raiha, he believed that differing experiences or perspectives can bolster teams' weaker areas. "Especially if we have like, some of us have shortage in language or not really understand the context in which the guidelines or the client is having the AI model work for."

Finally, Ghaliyah applauded EnVision Data for hiring data workers from different parts of the world. She said:

> *"The most important thing for AI is that we have different opinions, different ways of thinking. Of course, it depends, I think. For example, some annotators from one country, they will annotate some way. And the other people from different countries, they annotate another way. I think it depends on the society … I think the background of the annotator depends on the decision he or she makes for the annotation too."*

Through this statement, Ghaliyah acknowledges that people with different positional experiences will view the world in different ways.

None of the participants in this study expressed negative views of diversity. Instead, participants' commentary insinuates that having diverse teams working on the same project might provide a richer and fuller vision of the world. At least, having diverse data work teams can highlight the different perspectives which are currently being overlooked or made invisible by current processes. Much like standpoint theorists, participant viewpoints on diversity recognize that people approach problems from different positional perspectives. Most importantly, these different perspectives may fill gaps which may otherwise be missed—and eventually introduce negative and harmful outcomes once a model is deployed.

Discussion

I have built on the prior work of Miceli and Posada (Miceli & Posada, 2021, 2022) to center data workers and their practices, specifically in the development of computer vision datasets. I have documented the ways that data worker positionalities influence their work, as well as lead to unintended outcomes in scenarios where positionality is not attended to. I have revealed the limitations of personal positionality that data workers are aware of. At the same time, I have documented that workers still believe diversity is a boon for computer vision.

I will now discuss these findings, further contextualizing the role of positionality in data work. I begin by breaking down what a lens of positionality applied to data work awards us. I describe the benefits of attending to positionality. I highlight two insights which offer opportunities for computer vision practice. First, I describe how understanding positional perspectives in data work allows us to then attend to positional gaps that would otherwise be invisible. Second, I illustrate how attending to positionality from the vantage point of data workers reveals a larger web of positional actors involved in data work. Approaching the positionality of data workers as part of this web also reveals gaps and misalignments in the relationships governing data work.

Influence of Positionality on Computer Vision Data

As demonstrated by the Findings in this study, the positionality of data workers is salient to data work—in both data collection and data annotation, and across human-centric and object-centric data types. Concretely defining the way data workers referenced and exercised their subjective positions is difficult, if not impossible, given workers referred to a tacit knowledge influenced by a number of interlocking identities and relationships in conducting their work. Much like Rose argues (G. Rose, 1997), I found that fully accounting for positionality in research contexts is an ideal; despite reflexive practice, much of positionality is inaccessible and uncertain, especially when analyzing the practices of research subjects through one's own positional lens. Rather than attempt to define a taxonomy of data worker positionality, I instead revealed the multitudes of positionalities participants expressed in discussing their approach to work. Understanding how data workers express their positionality through their work allows us to do two things.

First, it paints a more contextual and rich image of how identity concepts are applied to data concepts for computer vision. It reveals how workers refer to culturally situated notions about identity, rather than idealized universal ones—like religious garb and what it communicates about gender. Identity is designed through a positional lens. Given that identity is interpreted through the eyes of data workers, workers have differential standpoints. They hold knowledge about some identities, but not others, which reflects their exposure to certain ideals and whether they believe them (e.g., understanding of trans people, exposure to certain media portrayals). Such knowledge is deeply innate, based on experiences, values, local context, and economic conditions. Thus, beliefs about identities are largely obvious to workers and difficult to unearth. Nonetheless, attending to positionality and attempting to piece apart how it manifests in data work reveals the gaps in current practices. Gaps in worker knowledge, worker/client communications, and client expectations begin to emerge. We can then more acutely attend to closing such gaps.

Second, it reveals a web of positionalities present in the process of data work. **Figure 14** showcases the web of relationships revealed through examining only data workers as the central point, but one could imagine centering clients or supervisors would expand this map of positionalities into further networks. Data workers not only interpret data instances, like individual images, through a positional perspective. They also negotiate positionality with other human actors during the process of their work. For example, Jaako relied on his own positional familiarity with cultural beliefs in Kenya to navigate asking data subjects for their face data. Gemma decided to offload the responsibility of making a decision about gender classifications to her supervisor when she was unsure. While data workers are central to this specific study, the others that data workers interact with also have their own positionalities, which influence the practices of data work. While not documented in the scope of this study, Gemma's supervisor then had to make their own decision about gender, as influenced by their particular standpoint. Even data instances themselves demonstrate a specific worldview, documenting a place, time, and perspective as they are captured. Positionality has a different impact depending on what the data options are (e.g., data workers having trouble when data has people of certain races). Such realities are demonstrated through the distrust data subjects had of English calls for participation, causing Rebecca to translate them into Filipino.

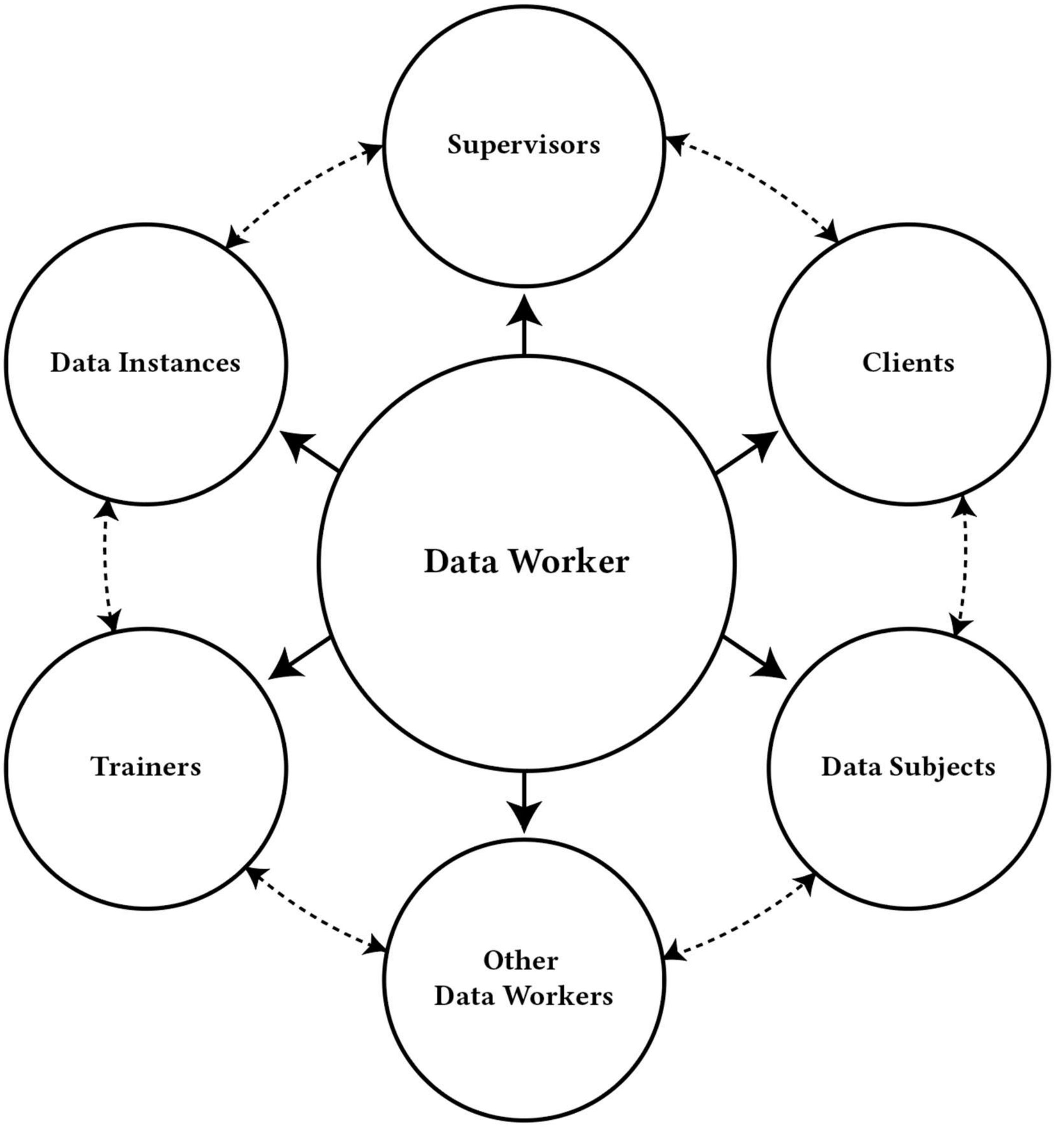


Figure 14. A diagram illustrating the other positional actors that data workers interact with. Data workers interact with data instances, but also data subjects, trainers, supervisors, clients, and other data workers. Further, all of these positional actors may interact with each other outside of the context of centering data workers. Data subjects may also interact with data instances, and supervisors interact with trainers. Each actor in this network may influence and negotiate with one another.

As argued by Miceli and Posada (Miceli & Posada, 2022), instructions begin to offer glimpses into the worldviews of clients, which data workers must attempt to adopt. When Sadhil interpreted the definition of a "blouse" as something different than the clients' expectations, it revealed that the client also had their own worldview which was otherwise left implicit in the term "blouse." While such definitional worldviews are often easily reconciled with communication and examples, many other worldviews are much more difficult to address. For example, trying to attend to differential perspectives on keypoint annotations for certain racial categories is more difficult, because “accuracy” is dependent on pixel-level annotations. Further, many worldviews are so deeply embedded that workers likely couldn’t describe them; trying to build shared understanding about such deeply implicit perspectives is extremely difficult. Jaako’s perspective on his own country’s culture as “primitive” also begs questions about the relational exchange of positional perspectives between data workers and their clients in the Global North. Did Jaako have this perspective due to his regular interactions with more technocratic clients, or was it something else? Attending to questions raised by the web of positional actors involved in the process of data work, and how data workers relate to this web, reveals a reality that data workers aren’t the sole source of “bias” in computer vision. There is ample opportunity to further understand how relational positionalities shape data work—and thus intervene at these points of interaction.

The Failures of Bias Mitigation Approaches

Positionality is complex. Much of the positional standpoints that data workers are operating from are invisible. To make matters more difficult, workers also have a difficult time articulating the role of positionality in their decisions. Further, positionality is not confined to the individual. Positionality is relational. Data worker perspectives are shaped through interactions with trainers, supervisors, clients, data subjects, other data workers, and the data itself. Exploring positionality reveals the limitations of viewing data in computer vision from a positivist episteme—as something objective, neutral, and containing an inherent ground truth.

Current approaches in bias mitigation (e.g., Gong et al., 2019; Harrison & Pan, 2020) attempt to debias and align data with some universal truth. Bias discussions often imply that bias can be easily identified, measured, and mitigated. Bias mitigation frames bias as discrete categories to be attended to and measured, presenting group parity or performance parity as fair (Kong, 2022). From the perspective that data is not biased but instead laden with positional worldviews, bias mitigation approaches can be seen as a failure to attend to positionality. Rather than attend to bias as the implementation of a specific worldview on identity, bias mitigation largely assumes one reality is correct. For example, “debiasing” gender classification technologies might showcase parity between the categories of female and male (e.g., Das et al., 2019). However, this parity might mask the reality of different perspectives and realities of gender. First, it fails to account for other lived experiences like trans experiences, other gender identities outside this binary worldview like non-binary genders, or other intersecting identity categories like race or skin color. But further, it fails to acknowledge what gender means in context. Data workers embed cultural perspectives about race, gender, age, emotion, clothing, etc. For what culture, time, or politic does gender parity operate?

Further, bias mitigation approaches fail to account for the context of data generation and annotation. In data work, the tools that data workers have are: instructions (which they do not always get), their positionality, and their exposure to the categories they are expected to collect or annotate. Clients imbue guidelines with their own worldviews, yet fail to acknowledge, document, or explain these worldviews. They also fail to account for data workers as people with their own positional perspectives. Clients in the Global North regularly fail to realize that data workers might be unfamiliar with certain Western-centric categories, like Hispanic vs. Caucasian (e.g., Gemma). They do not account for how exposure to certain categories might occur for data workers. Often, this is through daily life, or exposure to media, which may or may not reflect the expectations of a Western context. Clients instead present barebones guidelines that do not engage with differing perspectives or provide examples that clearly articulate an expected worldview. Implicit in this failure to account for data worker positionalities is the expectation that workers automatically adopt a Western worldview.

In reality, bias mitigation can make “bias” invisible. Invisible are the ways that data workers, operating from specific positionalities, interpret the instructions given to them by clients, whose own positionalities influence these instructions. In reality, positionality very subtly shapes data and thus model outcomes. Model outcomes can differ between contexts based on data (Sen et al., 2015), showcasing there is no universally unbiased gold standard for computer vision. The types of subtle “biases” that arise when exploring the nuances of positionality reveal major limitations to a positivist, universalist approach to machine learning and fairness.

Positional (Il)legibility

Even in cases where a model does perform well in its intended context, applying the lens of positionality to data work can reveal how practices influence data work, in desired or undesired ways. Given that positionality gives workers certain perspectives—and some workers might be epistemologically "closer" to certain subjects than others—I propose an approach that attends explicitly to epistemic standpoints. Rather than focusing on whether data is biased or unbiased, I propose attending to certain perspectives in data work as either legible or illegible to workers. *Positional legibility* refers to perspectives on data that are familiar, clear, and understandable to workers. On the other hand, *positional illegibility* means that data workers are unfamiliar with or do not understand the data they are working with.

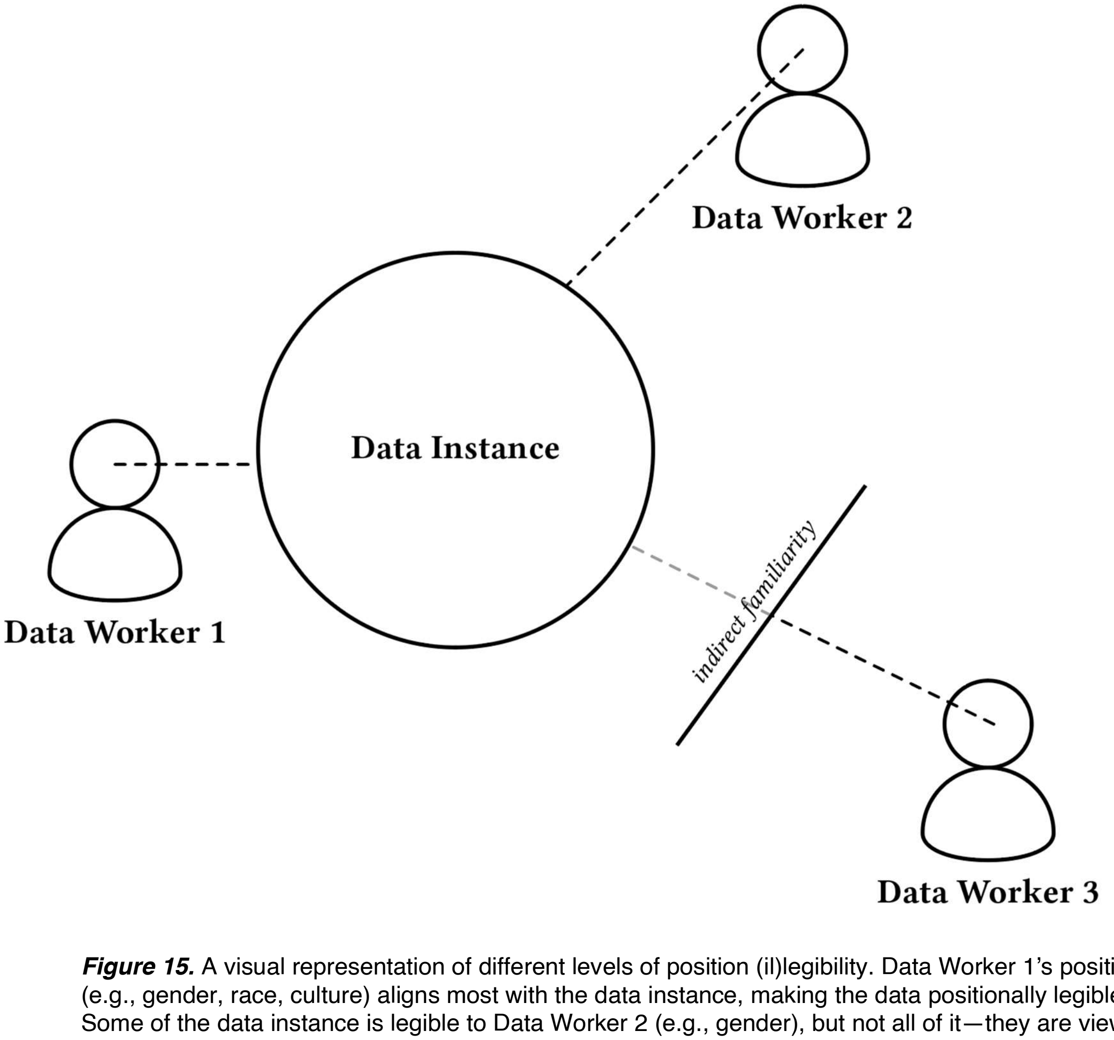


Figure 15. A visual representation of different levels of position (il)legibility. Data Worker 1's positionality (e.g., gender, race, culture) aligns most with the data instance, making the data positionally legible to them. Some of the data instance is legible to Data Worker 2 (e.g., gender), but not all of it—they are viewing it from a much different angle than Data Worker 1. The data instance is illegible to Data Worker 3, who has no experience with any aspects of the data instance; they instead view the data instance through the lens of

> others—like client instructions or media—but have no personal experience with the characteristics themselves.

Prior scholars have also proposed that worker identity influences the accuracy of their work. Those with identities shared by their data subjects are more accurate at annotation, for example (Patton et al., 2019). However, in a positional-first approach to annotation, accuracy is not the primary goal, as accuracy purports a universal ground truth. Unlike traditional debiasing approaches, the notion of positional (il)legibility would mean adopting an interpretivist epistemology that posits there is no objective reality, only subjective and situated interpretations.

Currently, context in the process of data work is lost to debiasing approaches. Workers tacitly refer to positionality as an interpretive resource—making sense of a western mindset as best as they can. They must translate the positionality implicit in guidelines through their own lens, then conduct the work based off of their interpretations. Misunderstandings reveal gulfs between positionalities of clients largely based in the Global North and data workers based in the Global South. For example, clients looking to collect data ethically might not realize when data workers have breached their own personal view on ethics. In such cases, data workers likely do not realize when they are engaging in “problematic” practices from the perspective of western ethics. They are not being purposefully unethical but are relying on tacit knowledge about the context of collection to get their jobs done. Similarly, clients assume that data workers in different contexts, where racial and gender categories might differ greatly from client countries, will innately understand these categories without examples. They not only fail to consider these categories may have different meanings (or no meaning) to data workers, but that any knowledge of these categories that workers have might come from portrayals in the media, rather than from exposure to people in their daily lives. Such exposure might result in annotations Wares was concerned about, that all Arabic men in hiring videos are “angry.” Clients need to make their own positions on data work legible to workers explicitly, outlining their own worldviews, their expectations based on those worldviews, and why it matters. Clients also need more open channels of communication, so that workers are not solely relaying confusions or mistakes, but also negotiating their own perspectives. Some data might be more legible to data workers than they are to clients. Gemma might have more knowledge about gender identities beyond cisgender male or cisgender female than her clients. However, current practices prioritize the worldviews of clients, regardless of whether the data is actually legible to them.

Adopting positional (il)legibility as an approach to classification, rather than assuming that objective classifications are possible regardless of data worker positionality, embraces positionality by the horns. Positional (il)legibility assumes positionality is always present in data work, given the range of positional actors involved in the process. Given that positionality is always present in data work, and thus data work is never neutral, objective, or "unbiased," we might design guidelines, select data workers, question assumptions, and attend to open questions surrounding issues of positionality by mapping out our relationships with data instances (much like **Figure 15**). Wares' point about contacting one another when confusions arise points to areas to begin closing these different gulfs: better training, better guidelines, and better communication. While there are already many resources available for hiring and training data workers (e.g., Lee et al., 2022), creating guidelines (e.g., Kornilova, 2022; Oshodi, 2022), and documenting decisions (e.g., Gebru et al., 2021; Miceli et al., 2021; Mitchell et al., 2018), there is still ample opportunity to appropriate such resources for a positional approach to classification, rather than a positivist one.

Conclusion

Computer vision is premised on its training data, the data used to train a model what the world looks like. The necessity of training data has resulted in a whole new industry of tech work, called data work. Data workers collect and label the images needed to train computer vision. Emerging research on data work for machine learning has revealed issues of economic precarity (Altenried, 2020; Anwar & Graham, 2021; M. Graham et al., 2017) and a lack of power over work conditions and procedures (Heeks, Graham, et al., 2020; Miceli & Posada, 2022; Williams et al., 2022). Nonetheless, data workers are regularly expected to provide "unbiased" and objective data in the pursuit of fair computer vision.

This study rejects the notion of "unbiased" and objective data to specifically explore how human positionality—the standpoint through which an individual views the world—influences the processes and outcomes of computer vision data work. I conducted 27 interviews with data workers (employed as freelancers or at a data BPO) about how they interpret identity concepts when doing collection and annotation work. I found that worker positionality influences decisions during data work through implicit tacit knowledge, which data workers had a difficult time articulating. I also found unintended and unexpected approaches to data work, like social biases and unexpected collection procedures. Such unintended outcomes occur when positionality is not explicitly attended to by the clients hiring them.

I discussed how attending to positionality in data work reveals both the gaps in worker perspectives, but also a range of positional actors that influence data work. I outline how current approaches to bias mitigation in computer vision actively fail to account for bias beyond a positivist view of “correct” versus “incorrect,” instead of attending to the reality that positionality is not black or white. I propose positional (il)legibility as a framework for capturing positionality in the data work process and actively attending to both the pros and cons of positionality in data work.

9
HOW WORK PRACTICES REFLECT POWER

In the previous two chapters (see Chapter 7 and 8), I uncovered how both traditional tech workers and data workers actively refer to tacit knowledge informed by their positionalities when conducting computer vision work. Both traditional tech workers and data workers rely on their own experiences and perspectives about the world to do their work—their individual positionalities (see Chapters 7 and 8). Even while the design of computer vision categories made at the traditional tech worker level are often portrayed as neutral or objective (see Chapter 5), every design is a choice made via negotiation with other workers and the broader context of artifact development (see Chapter 7). Similarly, data workers rely on their own worldviews to interpret how to collect or annotate data (see Chapter 8).

Positionality is salient to both types of workers, yet has largely been implicitly acknowledged in work on data worker bias (e.g., Davani et al., 2021; Kutlu et al., 2020; Sap et al., 2022). Few studies have been conducted on the role of traditional workers in defining identity concepts. Such an acute focus on data worker bias over an examination of the role of traditional worker positionality highlights one of the findings of this chapter: traditional tech workers are viewed as making value tradeoffs about identity in computer vision, while data workers are seen as introducing bias. While scholars have examined the lack of power given to data workers and their resulting exploitation, there is opportunity to understand how data workers' positionalities are constrained by traditional tech worker beliefs and practices. Through engaging directly with this opportunity to understand positional power in computer vision, we can both (1) understand the relationship between data workers and their employers, traditional workers, and (2) actively reflect on how traditional workers instill their own "biases" (i.e., perspectives) as well.

In this chapter, I examine the relationships between traditional tech workers and data workers throughout the lifecycle of computer vision dataset development. Beyond data being viewed as the main source of bias, datasets are used to train computer vision models to classify the intended categories. Through an examination of the relationships between workers, I reveal the power dynamics present in development that limit individual workers' positional perspectives—especially data worker perspectives. In reality, not all workers involved in computer vision are equally empowered to define and negotiate identity concepts. Data worker positionalities, in particular, are viewed as risky and biased, while traditional tech worker positionalities are perceived as perspective, expertise, and value negotiation.

This work represents a culmination of studies focused on both traditional tech work and data work in industrial-level computer vision development. I combine findings across studies of these two populations. In focusing on data workers, I employed ethnographic observations with a data Business Process Outsourcing (BPO) company, EnVision Data, over the course of a year, in which I held regular meetings, examined documents, and interviewed employees. I also conducted interviews with nine freelance data workers on Upwork. I spoke with 27 data workers in total. Focusing on traditional tech work, I conducted interviews with 24 tech workers employed at small and large tech companies across a variety of roles. For both populations, I sought to understand how individuals' positionalities influenced their approach to their work, how they approached identity problems in computer vision, and how they interacted with other workers. More specifically, in this chapter, I attend to the following research questions:

1. How do traditional tech workers discuss their own positionality in relation to data workers?
2. How do contingent data workers discuss their own positionality in relation to their clients, traditional tech workers?
3. What does the relationship between traditional tech worker and contingent data worker positionality communicate about positional power in computer vision development?

Findings are organized across broad stages of the computer vision development pipeline: starting with defining data categories, moving into selection processes for data work, then the application of data categories, and concluding with the evaluation of data work. Through the findings, I showcase how identity is defined and negotiated—and eventually evaluated—at the level of traditional data work but applied at the level of data work. I demonstrate how traditional tech workers discuss data workers through the lens of bias mitigation and control, often unintentionally imparting their own specific white collar, techno-solutionist, and Western-centric positionalities onto economically disadvantaged data workers located in the Global South. Similarly, I show how data workers reflect on their own positions in relation to traditional tech workers, sometimes lamenting and sometimes internalizing Western-centric techno-solutionist perspectives.

Through a discussion of these findings, I outline how a specific form of power, positional power, operates in the field of computer vision. Positional power refers to the domination of certain positionalities in defining identity concepts in an artifact. Traditional tech workers' positionalities dominate the development of computer vision artifacts. Meanwhile, data workers are expected to put aside their own positional perspectives and embody traditional worker worldviews. In demonstrating how positional power operates in computer vision, I draw inspiration from Bourdieu’s theory of habitus and Hill Collins’ matrix of domination. I describe how both types of workers have internalized perspectives about conducting computer vision work that uphold current positional power structures. Traditional tech workers exhibit a habitus congruent with the values embedded in the field of computer vision and are thus awarded a positional power. On the other hand, data workers embody a habitus that is devalued in tech and internalize a belief that they are not qualified to be making conceptual decisions about identity. Further, positional power encompasses an interlocking matrix of worker identities, where even traditional workers with more marginalized social identities have less power within the field.

I argue that attending to the current state of positional power in computer vision—where traditional tech workers dominate how identity is conceptualized and data workers are to simply enact their perspectives—opens up new opportunities for reimagining the development of computer vision. I provide examples of how practitioners might rethink shifting power during the development process.

Related Work

I situate this chapter’s work within scholarship focused on power and positionality. First, I describe how power has been approached in prior sociotechnical work. I also describe the approach to power that I am taking in my analysis and presentation of this Chapter’s findings and discussion. I then focus on how power has been analyzed among my two worker types: traditional tech work and data work. I first describe prior scholarship on traditional tech workers, including notions of power embedded in that work, primarily focused on the power traditional tech workers have when implementing notions of fairness in machine learning. I conclude with scholarship on data work, which has primarily been aimed at understanding the exploitation of data workers. I expand work on these two types of workers by examining the relationships between them, particularly as it pertains to applying positional lenses to the development of identity concepts in computer vision.

Power and Positionality in Technology Design

Technology reflects the values of their designers and the broader social contexts they are embedded in (Winner, 1980). As Bowker and Star so eloquently stated, “to classify is human” (Bowker & Star, 2000). Computer vision is a technology reliant on the classification of data. Therefore, computer vision artifacts reflect specific worldviews about that data. How identity characteristics are embedded into computer vision models and datasets echo dominant ideologies about identity as entrenched within larger social structures. I have demonstrated in Chapters 3 and 4 how implementations of gender and race in computer vision, specifically, maintain oppressive and regressive categories which harm marginalized groups. Further, in Chapter 5, I showed how approaches to computer vision work reflect the values of specific knowledge disciplines. Those worldviews become increasingly embedded through the use of artifacts (Bender et al., 2021; Koch et al., 2021).

This chapter contends with two interlocking concepts: *positionality* and *power*. Positionality (as introduced in Chapter 2) posits that an individual’s experiences, values, beliefs, and identity impact how they view the world—and in turn, their positionality is mutually shaped by the world around them. The notion of positionality stems from feminist standpoint theory, the vantage point by which we as humans view the world (Harding, 2004). Standpoint theory argues that certain identities are closer to certain knowledges, lending a specific expertise. For example, Black women understand Black women’s experiences more than white women, and Indigenous cultural practices can be more fully understood by those Indigenous peoples who engage in them (Collins, 1998). While first- and second-wave standpoint theory has been critiqued for its essentialism—assuming that identity groups share the same outlooks on life—standpoint theory has since evolved as a more contextually situated theory (Rolin, 2009). Along the lines of being more contextually situated, Wylie suggests that since identities are not static and binary, we cannot assume a specific and fixed standpoint based on binary notions (Wylie, 2003).

Along these lines, Black feminist scholar, Patricia Hill Collins, pairs standpoint theory with the theory of intersectionality, “the ability of social phenomena such as race, class, and gender to mutually construct one another” (Collins, 1998). Collins argues that standpoint theory approaches positionality at the level of the individual, while intersectionality does so at the level of social groups—and a situated standpoint should address inequalities via “new understandings of social complexity” (Collins, 1998). Individuals *among* a group may operate from different positional standpoints, but the relationships *between* groups are also crucial to examinations of power. Collins focuses on notions of power in standpoint theory, arguing that power is central to our identity positions.

Power has been a major theme in much of social computing research. Scholars have examined power from a multitude of angles. For example, Walker and DeVito examined the power dynamics between different social identities under the LGBTQ+ umbrella (Walker & Devito, 2020). Kannabiran and Petersen proposed attending to power at the level of interface design (Kannabiran & Petersen, 2010). Kirabo et al. investigated the differential positions of power held by stakeholders in disability transportation services in Uganda (Kirabo et al., 2021). In the domain of machine learning, power has been especially central to discussions of data ethics and model imposition on data subjects. For example, Shilton et al. examine the powerlessness of the general public in the age of corporate datafication, arguing for increased engagement with the role of power in conducting research with digital data (Gilbert et al., 2021). Koopman proposed a framework of infopower to describe how algorithms fasten human subjects to data (Koopman, 2019). Finally, Miceli et al. specifically call for attending to power rather than notions of bias in machine learning, specifically centering the contexts of production in scholarly analyses (Miceli et al., 2022). They argue for explicitly attending to “historical inequities, labor conditions, and epistemological standpoints” rather than the notion of bias, because the frame of bias fails to account for the social realities of data production.

Power is always core to examinations of positionality (Merriam et al., 2001). In discussions of identity and positionality, power operates as a means of domination. Such power is not necessarily gained and maintained through physical force, but often through relational and symbolic interactions.

Standpoint theories are focused on understanding the perspectives of certain positionalities—along lines of race, gender, class, sexuality, ability, age, nationality, and culture. Embedded in this perspective is the reality that certain positions are privileged over others. Intersectionality is one theory for examining power relations between social groups. Collins also presents the matrix of domination as a tool for considering how some members of a group may be privileged in some ways, but dominant in others (Collins, 1990). But power between both individuals and groups manifests in various, intersecting ways. Pierre Bourdieu's symbolic power posits cultural roles as more influential than economic forces in determining relationships of power (Bourdieu, 1987). Cultural roles determine how social groups are bounded. For Bourdieu, power is created and maintained through *habitus*. Habitus "is a set of *dispositions* that which incline agents to act and react in certain ways. The dispositions generate practices, perceptions, and attitudes which are 'regular' without being consciously coordinated" (Bourdieu, 1991) (emphasis in original). Unlike Foucault who views power as ubiquitous and all encompassing, Bourdieu proposes a lens of power which is culturally created and situated. Much like Foucault, Bourdieu also proposes language as one of the main mechanisms for creating and maintaining power. Power exists within bounded fields where belief is produced and reproduced through the legitimacy of utterances, but also the legitimacy of those who utter them.

In my analysis of the relationship between traditional worker positionalities and data worker positionalities, I examine how power shapes practice in computer vision. Power is critical to defining identity in computer vision because it constrains and enables how different workers reference their positionalities. In the case of this chapter, I approach power as a system of domination that is embedded into the practices underlying computer vision development. Domination refers to how certain positionalities are privileged while others are minoritized, erased, and made invisible. As I will argue, certain positionalities are given the language, tools, and opportunities to define identity for computer vision, while other positionalities are treated as tools for embodying privileged positionalities. For example, traditional workers define the language of identity and then use data guidelines as a mechanism for ensuring data workers enact that language of identity. Having positional power in the context of computer vision describes how certain types of workers can override other workers' perspectives. For example, not only do traditional workers use data workers as tools to enact their positional perspectives, they also actively rewrite data workers' perspectives in the data when they disagree with it. Those with power have more capacity to define identity, select data services and clients, design data practices, and evaluate data work.

In my analysis of power, I am informed by the work of two theorists: Bourdieu and Hill Collins. Bourdieu provides a lens through which to view specific fields of practice as spaces of power, where certain people in that field are awarded capital and others are not. Further, Bourdieu proposes that power is created and maintained through *habitus*, a set of internalized dispositions that reflect deeply ingrained social norms within a field of practice. Much like the theory of positionality, the theory of habitus proposes that individuals are encultured into specific orientations within the world. Those orientations enable or disable individuals with different habitus to integrate into different "fields," or social spheres of activities (e.g., higher education, politics, or the tech industry). The habitus a person embodies is expressed through language, nonverbal communication, values, and modes of reasoning (Bourdieu, 1991). As Edgerton and Roberts summarize, habitus "shapes the parameters of people's sense of agency and possibility" (Edgerton & Roberts, 2014). While habitus is deeply ingrained, often operating at the subconscious level, one's habitus might change over time, as they encounter different fields and respond to them, or in instances of intentional reflection.

Bourdieu's perspective of power provides a tool through which to more deeply engage with the norms ingrained within the "field" (in the Bordieuan sense of a social arena) of computer vision. For example, it allows us to engage with work practices in computer vision as a field of power, where traditional workers inhabit a habitus that awards them positional power. Data workers similarly possess a habitus of lower positional power but uphold beliefs about the positional power of traditional workers. Thus, in examining how positional power plays out in the field of computer vision, I also attend to how power operates within traditional and data worker habitus.

Further, given that I am examining how power operates relationally as different workers define identity in computer vision, I also pull on Patricia Hill Collins' *matrix of domination*. While Bourdieu does not explicitly attend to identity classifications in his theory of power, Hill Collins examines power through the lens of social categories, like race and gender. Beyond examining computer vision as a field of power, I turn to Hill Collins to examine the social relations between individual types of workers, showing that positional power is further complicated by social classifications, like race and gender. She proposes that social classifications intersect to privilege and constrain both individuals and larger social groups. For example, she proposes that, in the case that two individuals share the same characteristics, but one has a higher level of education, the one with the higher level of education is awarded more power. Further, she complicates simplistic and one-dimensional views of social power as simply categorical by arguing that individuals may be privileged in one way but marginalized in another. I use Patricia Hill Collins' to attend to how the social positions that individual workers inhabit also influence the power they have in computer vision work, meaning that traditional workers do not have all encompassing power simply due to their classification as traditional workers.

In my analysis of power, I examine how different classes of workers—traditional workers and data workers—are given different positional power through the development practices of the field of computer vision. Specifically, I show how traditional workers are awarded positional power in defining identity in computer vision over data workers, which are treated as automated tools for which to further ingrain traditional worker perspectives into artifacts. I show how the habitus of these two classes of workers upholds the current status quo of positional power, that traditional workers are more qualified for the task of designing identity than data workers. However, I also further complicate the concept of positional power by arguing that certain individuals within worker classes are given more positional power than others, dependent on the social categories to which they belong.

Traditional Tech Worker Practices and Power Relations

HCI and CSCW communities have a long history of examining the work practices of white-collar workers. Scholars have examined everything from the distributed cognition of air traffic controllers (e.g., Bentley et al., 1992; MacKay, 1999; Qinghao & Dengkai, 2017) to music production (e.g., Benford et al., 2012; McGarry et al., 2017, 2021), employing ethnographic observations and interviews to understand worker practices and contexts. Given the longstanding relationship between HCI and the tech industry, there is a rich scholarship focused on studying tech workers in companies, large and small (e.g., Halverson et al., 2004; Olson & Olson, 2000; Seidelin et al., 2018). While work on computer vision practitioners, specifically, is non-existent, increasingly researchers are conducting work with machine learning practitioners. Research has focused on the current practices of machine learning practitioners and the challenges they face in conducting their work. For example, Muiruri et al. investigated the development of machine learning systems across sixteen Finnish organizations, identifying well-established approaches as well as a lack of desired tooling for monitoring ML systems (Muiruri et al., 2022). Zdanowska and Taylor conducted a study on how user experience practitioners design machine learning systems outside of big tech companies, and documented how UX practitioners negotiate UX approaches when designing ML systems with other workers in company contexts (Zdanowska & Taylor, 2022).

I described fairness efforts in industry contexts in Chapter 2. Yet work on fairness in industrial machine learning points out that such interventions are not without their challenges, such as limited resources for tooling (e.g., Holstein et al., 2019). Further, as demonstrated by the public scandal of Google firing prominent ML ethicist Timnit Gebru, management buy-in of fairness efforts is often limited (Simonite, 2021).

The limitations to implementing fairness efforts point towards issues of power—who and how individual tech workers are empowered to address ethical deficits in product and their organizations. Investigations of power in commercial contexts have largely been limited to the relationships between tech workers and the larger organizations in which they operate. For example, Wolf et al. conducted interviews with white collar workers, including technologists, focused on addressing wage theft in the United States, reflecting the relative organizational power tech workers have even under capitalism (Wolf et al., 2022). Su et al. explored how the “techlash” shaped the cultural environment in which tech workers were situated to explore how affect impacted knowledge work and organizational action (Su et al., 2021). They document how the emotional habitus of tech workers create cultures that enable or disable political action. Similarly, Amrute posits that the different positionalities—such as race, gender, and class—can open up collective organizing in tech organizations in new and fruitful ways (Amrute, 2019). Further, much like I described in Chapter 6, tech workers, in particular, often have economic and educational advantages over other forms of traditional white collar work (e.g., Binder et al., 2016; Meiling, 2021). Such economic and educational prestige awards tech workers relative power—or, in other words, Bordieuan cultural capital (Edgerton & Roberts, 2014)—to push back against their organizations in ways that are otherwise not present for other classes of workers.

Yet amongst traditional workers in different roles, there clearly exist different values about fairness and implementing identity concepts. Numerous scholars have documented conflicting values in computer vision artifacts, highlighting the opportunities to make different decisions than the status quo (e.g., Hanley et al., 2021). In prior work, I proposed that computer vision as reflective of colonialist ideologies about race and gender, often unintentionally embedding a harmful history in the present (Scheuerman et al., 2021).

Muller et al. interviewed data scientists at IBM about their interpretation and analysis of raw data, demonstrating how data scientists see themselves as ground truth (Muller et al., 2019). In other words, the traditional tech workers doing knowledge work are the source of “truth” in machine learning systems, rather than some universal or standardized reality. As argued by Crawford and Paglen, the process of categorizing data for the purposes of computer vision tasks is “itself a form of politics” in which representations in computer vision systems are solidified worldviews (Crawford & Paglen, 2019). Simply focusing on fairness as an ideal outcome for machine learning systems fails to account for how differential positional perspectives serve different interests and values. In her essay “Don’t ask if AI is good or fair, ask how it shifts power,” Kalluri questions commercial approaches to fairness, asking “fair and transparent according to whom?” (Kalluri, 2020). This work answers the call to attend to power, examining how traditional workers express their positionalities in shaping identity in computer vision. More specifically, I examine how traditional worker positionality is expressed in relation to data worker positionalities, showcasing how power manifests in the hands of traditional workers but blame falls on data workers for poor data outcomes.

Data Worker Practices and Dis(empowerment)

There has been a great deal of work documenting how the gig economy, broadly, and data work, specifically, is economically undervalued (e.g., Dunn, 2016; Gillis, 2001; Ruyter et al., 2018). These accounts on data work paint a damning picture. Graham et al. describe how globalized digital labor has degraded bargaining power for labor rights between employees and employers (Casilli & Posada, 2019). They highlight that the majority of requests come from the Global North, with much of the labor coming from the Global South, with wages reflecting this dichotomy; digital workers are paid less in the Global

South. A global digital economy has made it easy for client to engage in labor arbitrage, increasing competition between workers to undervalue their labor to secure jobs (Zheng, 2020).

Some of the most popular data work platforms have negative reputations for employee exploitation. Data workers on Appen have turned to online forums to warn of clients using the platform to deny pay for completed projects (Bogle, 2022). Sama, which markets itself as an “ethical AI” company and provides services to tech giants like Google and Meta, underpays employees in the Global South. Sama employees in Nairobi are paid the lowest in the world, despite providing traumatic labor for content moderation (Perrigo, 2022a), begging questions about the exploitation of specific regions of the world despite the necessity of content moderation labor. Sama is also being sued in Kenya for violating the Kenyan constitution on human rights (Perrigo, 2022b). Williams et al. write, “Unlike the “AI researchers” paid six-figure salaries in Silicon Valley corporations, these exploited workers are often recruited out of impoverished populations and paid as little as $1.46/hour after tax” (Williams et al., 2022). Their words highlight the vast chasms between traditional tech workers and contingent data workers.

Given workers are contingent and not traditionally employed, they are often unprotected by traditional labor laws (Garden, 2018). Scholars have argued that labor arbitrage practices have negative consequences on the health of local markets of the Global South targeted for outsourcing (Enwukwe, 2021; Heeks, Eskelund, et al., 2020; Lesala Khethisa et al., 2020). Researchers also point out the negative implications of platform work on gender and racial minorities (Arora, 2016; van Doorn, 2017).

Given economic exploitation and a lack of legal protections, it is unsurprising that scholars have focused on relationships of power between data workers and their employers. Miceli and Posada examine how power manifests between data workers and their clients, traditional tech workers, particularly through the form of instructions (Miceli & Posada, 2022). They identify how instructions act as mechanisms for controlling how data workers interpret data. Workers are also regularly surveilled and controlled, viewed with distrust and treated much like machines (L. C. Irani & Silberman, 2013; Shapiro, 2017). Unfortunately, the labor of data workers tends to be entirely invisible in the process of developing computer vision. Gray and Suri refer to these workers as ghost workers, given their invisibility in the development pipeline (Gray & Siddharth, 2019). Traditional tech workers are thus portrayed as valuable knowledge workers (L. C. Irani & Silberman, 2016) and machine learning is viewed as “magic” (L. Irani, 2016), as the underlying human labor powering machine learning is made invisible. As Altenried states, “the work of annotators is dictated by the interests, priorities and values of others above their station” (D. Wang et al., 2022).

Yet despite this invisibility, data work is central to the function of computer vision. Not only is data crucial to models working, it shapes the outcomes of models. Worker identity is influential in this process; workers positionalities influence how they approach data work (see Chapter 8). Goyal et al. found that annotator identities were significant to how workers rated the toxicity of content (Goyal et al., 2022). Barlas et al. created a dataset of people and collected tags from commercial systems and both Indian and U.S. annotators (Barlas et al., 2019). Their findings highlight cultural differences in annotation, as well as differences between commercial systems (whose annotators are unknown) and human annotators. These works emphasize that data worker’s positionalities—the worldview shaped by their identities—are highly relevant to data work outcomes. Data workers are not simply digital factory workers doing the mechanical work of applying universal truths to data but are experts conducting knowledge work about the data they work with.

I augment work on power, demonstrating that workers are not only disempowered economically and procedurally, but positionally. Data worker positionalities are devalued in comparison to traditional tech worker positionalities during the development of computer vision artifacts. Much like Parks says in (Dirty Data: Content Moderation, Regulatory Outsourcing, and The Cleaners), “the voices of Big Tech experts or Western digital activists reign supreme and tend to drown out the experiences of digital innovators and experts in the Global South.” I approach data workers as experts whose positionalities are fundamental to computer vision. Yet, as I will demonstrate in this chapter, data worker positionalities are undervalued in favor of tech worker positionalities, despite many unfavorable formulations of identity being developed at the tech worker level.

Analysis

In this chapter, as I stepped into the analysis, I specifically attended to issues of power. In attending to power in my analysis, I drew from theories from both Bourdieu and Hill Collins. To more deeply examine relations of power between traditional tech workers and data workers, I am inspired by Bourdieu's notion of *habitus* as a lens through which to engage with and examine my data. *Habitus* describes the socialized norms underlying behaviors and practices; it is "the way society becomes deposited in persons in the form of lasting dispositions, or trained capacities and structured propensities to think, feel and act in determinant ways, which then guide them" (Wacquant, 2006). It is unconsciously enacted and reinforced in individuals through their socialization to a specific context. The worldviews embedded in *habitus* reinforce beliefs about value, or capital—not only in an economic sense, but also in a symbolic and cultural sense. Given that I have previously demonstrated the role of tacit knowledge in conducting identity work in computer vision (see Chapter Two), *habitus* provides a mechanism with which to understand how the intuition underlying tacit knowledge also maintains specific power hierarchies between different types of workers. I also refer to Collins' matrix of domination, examining how certain positionalities—certain standpoints, certain knowledges, certain habituses—are privileged in the field of computer vision. I examine how those privileges reflect dominant identity hierarchies both *between* and *within* the two types of workers.

After collecting all data, I open-coded all interview transcripts to gain more intimate familiarity with the data and to identify emerging themes. For each recording, I wrote a memo describing the participants, the company they work for and the product(s) they work on, and their general perspective and experiences in relation to identity implementation. Given the vast and diverse data derived from this study, I regularly compared findings across participants. I began writing larger memos describing similar participants, similar products, and similar perspectives and experiences. Writing comparative and contrasting memos aided me in defining larger theories about the data. I wrote theoretical memos describing these theories and began populating them with examples from my data.

Once I had fully defined theoretical memos, I began the process of breaking them out into more expanded theories. I began memoing out themes that fit within the larger theory about the relationship of power to expressions of positionality in defining identity in computer vision. It was these final theme memos that became the findings of this chapter.

Findings

I describe how two different classes of workers—traditional tech workers and data workers—approach identity characteristics in the development of datasets for commercial computer vision systems. More specifically, I document how these workers' positionalities inform their approaches to defining, classifying, and evaluating data. I present findings on these positionally informed approaches across the data development pipeline—from defining data categories, selecting clients and services, conducting data work, and evaluating the completed datasets.

Power to Define Data Categories and Scope Projects

Traditional tech workers had the power to define identity at the start of a project, as well as scope how a project would go. Specifically, they could define how identity was meant to be represented for a specific task and they had the power to select or turn down clients whose visions for identity they disagreed with.

How Traditional Workers Define Identity

Before having data collected or annotated or having models created, clients first need to define what they want their model to classify and how. Two actors are generally involved in this process: traditional workers and customers. Data workers are not involved in defining identity for computer vision. As such, the power to define what identity should even be was negotiated between different traditional workers and, in some cases, their customers.

As demonstrated in Chapter Two, the way that identity concepts are defined varies greatly. Clients may request explicit identity classifications for either model outputs (e.g., demographic classification models) or for evaluation (e.g., bias testing). They also often request data for non-human objects which are imbued with sociocultural values and meaning (e.g., food or clothing). Deciding how categories should be defined isn't simplistic or objective but is a process of negotiation between multiple traditional tech workers with differing positional perspectives. For those categories to be useful to computer vision, data must be collected and, often, annotated. As Madison explained*, "working specifically on human-centric computer vision … that quickly, you know, hits against issues of defining* *identity groups and getting them annotated."* Traditional workers occupied roles that allowed them to define identity groups for computer vision products.

As demonstrated by Lynn's description of choosing demographic subgroups to evaluate model performance, defining the scope of identity categories is often overwhelming for traditional workers:

> *"The first thing that I needed to do, which was a terribly, terribly intimidating prospect, was to sit down and think about how are we even going to test for which of the population you know, underrepresented are being unfairly discriminated against in this category. And like, you know, the kind of resounding sentiment from the team was like, we have to constrain this problem because it's an impossibility criterion if we just allow ourselves to think about every single phenotype and every single appearance of human you know, facial features."*

Lynn's team had discussions about how to scope identity categories so that they were still possible to measure, given the diversity of humankind. They had to negotiate their own personal values about the reality of identity to settle on something measurable: *"So you think about, like, where is there maybe a standardized thing that I can steal from and then like, well, there's the US Census, which is highly problematic."*

Western culture also seems to drive how identity is defined in AI. Vasudha described how her company's computer vision models are trained primarily on celebrities, and due to this that it is *"unfortunately, [heavily] influenced by the West."* Identity is generally defined through a Western lens. The majority of traditional tech workers, who are actually defining the problem space and requirements for computer vision products, are based in the West.

Of course, not all traditional workers agreed on how best to scope identity, or what should or should not be classified. Elliot had to negotiate their own personal views on identity with the perceived views of their engineering colleagues:

> *"Basically, [I'm] telling machine learning researchers that if they're going to be in the business of making predictions about people and affecting things in the world, they very much need to adopt the … strategy of like, understanding racial dynamics, as opposed to, this is some fixed attribute of an individual and we're just going to control for it after the fact."*

They feel that identity categories are too often defined by "*engineers who don't think that much about identity.*" They express that they don't necessarily have the power to stop other workers at the company from building identity classification models, but instead must try to advocate for more incremental changes. They describe how they have pushed back on gender classifications and proposed reframing gender annotation as "*perceived gender … as given by this annotator*."

Nitesh described how unequal power dynamics could arise amongst traditional workers, stating that "*when you actually have a lot of experience, you can actually share some opinions*." Nitesh felt that experience in the workplace awarded one more leverage, where you could share opinions because they would be taken seriously by colleagues. Madison similarly described how issues like sexism disempowered women in comparison to men in her workplace:

> *"If you have a group of like women say … how they think the technical aspects should go, and it's what you're not used to hearing. There's like, no way they'll be taken [seriously] … But then if you have people who are white or Asian men, maybe more fitting the traditional personalities or understanding of who tech people are, putting forward ideas it does get a lot more traction."*

Traditional workers did not have equally distributed power in defining identity. Aspects of their positionalities mattered—their job titles, their experience levels, how their colleagues perceived them, and the societal effects of sexism and racism influenced how much power a worker had in defining identity. As described by Madison, product teams would often take suggestions more seriously from men—particularly white and Asian men who are seen as fulfilling the expected role of a tech worker. Women were only heard in cases where they mimicked the expected perspectives of their male colleagues. Similarly, those with more experience at a company, who had built relationships with colleagues and management, had more leverage to have their ideas heard.

Beyond the positional negotiations between traditional workers, negotiations occurred with customer representatives at other businesses. Many tech companies also serve their own clientele, providing models for specific business purposes, like marketing or content moderation. Kenny describes, through the application of labels to data, *"how [MultiplAI] entertain[s] a client's decision."* Kenny's choice of words implies that their B2B customers make subjective decisions about how they desire a model to perform, and data workers are the ones who apply those desired attributes to training data. In these cases, where traditional workers serve a customer, the customer largely drives expectations about how identity should be defined. There is not an inherently correct way to classify data, but rather a profitable way which a customer desires for a specific business purpose. As I'll describe in the next section, traditional workers do have more power in choosing their clients than data workers. However, they still need to serve client demands, especially when other traditional workers in higher management positions decide to take on customers.

The above examples showcase how traditional workers engage their own positional perspectives and negotiate them with their colleagues when defining categories for computer vision. Data workers, on the other hand, are not involved in the process of defining data categories. Gemma explained that, while she did have some opinions about how identity categories—gender, in her case—were defined, she did not feel empowered to speak up. She said:

> *"Most of the time when you are junior staff, there are some things you cannot just come and speak out. Some things are raised by the manager cuz when you try to show that you know a lot, it might cause some displeased [sic], so you just have to follow the instructions … you're just a junior staff, so you just do your work and then that's it."*

As demonstrated by Gemma's explanation, while data workers did demonstrate opinions about how categories were defined, they were not part of scoping conversations and had little opportunity to provide feedback on those categories.

Finally, some tech workers felt certain categories were less necessary than others. Often, this was not necessarily based on experience working with such categories, but instead due to other constraints, like budgets and lack of data. Sophie described how they did not test their facial biometrics on people with facial disabilities *"because they're less structural in terms of the physiological aspect of the face."* It was not easy to break down disability into discrete categories, like race or gender. Further, she described that there was a "*lack of data in comparison to, like, man versus woman*." Disability was rarely, if ever, a category traditional workers thought about—with the exception of Jeremy and Lydia. A lack of data and prioritization of disability as a category in computer vision models also reflects certain privileges that might lead to inequalities for people with disabilities in models.

To summarize, the first step in the development of computer vision is defining the data categories that drive classifications. Often, this meant defining what identity should be in the form of categories. Data workers were not involved in this process at all, despite having their own perspectives on how identity should be defined. The only positionalities present were those of traditional workers and their customers. Traditional workers referenced their own positional perspectives to define identity, determining which identities were deemed necessary to attend to and which were not. Given the traditional worker participants in this study were entirely based in the Global North, their approaches reflected a Western-centric positionality. Whoever fulfilled the role of the client—whether traditional workers who would hire data workers later or customers who hired traditional workers to provide modeling services—often held the most influence in how identity was defined. Prioritizing client needs indicated that economic forces largely drove defining identity in computer vision products.

In the remainder of the Findings, I describe the relationship between traditional tech workers, in their role as “clients” seeking data services, and data workers. In particular, I show how traditional tech workers view their own positional approaches to computer vision as creative, critical, negotiable, and nuanced, while implicitly perceiving data worker positionalities as risky. Traditional tech workers, in their capacity as clients, expect data workers to adopt and carefully reference *their* positional perspectives on identity. Thus, they attempt to carefully control data work processes.

How Traditional Workers Screen Potential Customers

While customers may approach certain companies for modeling service, representatives at those companies have the ability to turn them down or take them on. The power between traditional workers and their customers is relatively equitable during the selection phase, as the two parties determine what is and is not a good fit for a project. Traditional workers, in their roles as intermediaries between customers and data workers, have the power to select customers based on their own ethical assessments. The data workers providing data services for the customer downstream did not.

Nicholas describes instances where his company decided against working with certain customers based on their values as a United States company. He said*: “We've been asked to do some things in prediction, that wouldn't be good to do in the US, but would be okay to do in that country. And we've just said, … No, we won't do it.”* At other times, tech workers feel the need to negotiate their own values with that of their customers. Rather than outright rejecting customers, they try to work with them. For example, Lydia describes how *“different countries have different laws around bias and fairness … like, Japanese data is usually really sexist.”* From her own Western-centric view, working with Japanese customers is difficult because the data is “*sexist*” towards women. She values gender equality, and still feels the need to ensure that in models for Japanese customers. Rather than reject customers, she works to mitigate bias in the model despite her perspective that the customer doesn't care. She explicated further: *"They don't care about that in their country. So like, the customer doesn't care. And whereas we would want to mitigate … So, I guess for me, it's just kind of like having that conversation where there's cultural differences of where that balance should be."*

Nicholas described the process of negotiating with customers as tricky, because he wants to maintain a balance between company image and bringing in customers. *"We think what's best for your hiring practices and building your company."* At the same time, he doesn't want to necessarily embed models with their own beliefs over the contextual beliefs of their customers. Nicholas continued:

> *"Because we're an American based company, does that give us a right to sort of inject that attitude on another country and their hiring practices that don't follow that? … We have the right to not do business with them. But do we have the right to inject our belief on it and change algorithms for them? So that's sort of the philosophical discussion we've had internally with this a couple of times."*

Nicholas describes the tradeoffs he and his colleagues feel they need to make when selecting customers—whether they will take the customers on and inject their own values into the models, or whether they will reject them.

BPOs, like EnVision Data, have the ability to screen clients on behalf of their workers. However, it is not a common practice for BPOs to prioritize ethical projects, high pay, or fair treatment. Lynn, who often directly worked on selecting vendors for data outsourcing, described the careless attitudes of many of the BPOs she met with:

> *"I've talked to 40 or so different vendors over the course of my time there, and some of them were just horrible, horrible people. I would call it almost, you know, the way that they talk about their workforce, and they talk about the law enforcement [use cases] where they're annotating body cam videos, and things like that. And they're kind of like, 'Oh, yeah, we do that work, that's totally fine, we don't care if our workers are doing stuff like that.' That was the worst sales pitch I've ever sat through."*

Lynn, in her search for data workers to do content moderation labor, was also focused on being as ethical to the workers as possible and regularly encountered BPOs who had no concern for the wellbeing of their workers.

Social good companies like EnVision Data make an effort to select clients that they believe contribute ethical—or at least, not actively harmful—computer vision products. Samuel worked to determine whether clients requesting data services were ethical. They utilize an ethical scoring system that the C-level staff came up with: (0) very high risk of harm; (1) high risk of harm; (2) neutral; (3) indirect positive impact on society; (4) direct positive impact on society. While Samuel's role is assessing the ethics of clients, every decision to contract with a client goes through Irina, the CEO. Samuel and Irina both said it was difficult to actually assess the ethics of a project, because they had to rely on the clients' word.

While they refuse to work on projects they label 0s, such as weapons projects, they do take on 1s. Irina explained that their desire to avoid attributing to unethical computer vision products is often in tension with their mission to provide work to their data workers: *"Ideally, we'd be working primarily on deals with a score of 4, because that's more in line with our vision and who we are as a company … I don't think we can afford saying no to [1s]."* EnVision Data is not accountable for the product uses as a contractor; if they were to be accountable to end uses of the data, they would be more "picky" than they currently are. While they refuse to take on content moderation or weapons projects (e.g., autonomous aerial vehicles using computer vision to locate targets) now, they are aware that other BPOs take that work on. When rejecting a weapons project, they were aware their client went to iMerit for the project instead.

Irina described having a difficult time competing with cheaper annotation companies based in India and Africa, which do not operate as social good NGOs. She expressed that clients were generally not interested in ethical annotation labor or whether annotators were being treated fairly. In one project meeting, she described one of the clients as *"checking the box"* when asking about how annotators are paid, reflecting a cynicism about whether clients actually cared rather than it being a policy requirement. Given that one of her clients, Emovos, only paid half of what they had agreed to before cutting them off, such cynicism was unsurprising. Ghaliyah, who started off as an annotator at EnVision Data but now works as an annotation trainer, expressed a positive view of EnVision Data in comparison to other BPOs: *"We care about our annotators, this makes us better than other companies."*

Similarly, since EnVision Data's workforce is uniquely employed through NGOs in conflict affected areas, the NGOs can push back on EnVision Data's projects. In one example, the NGO EnVision Data works within Iraq refused to annotate a project that included lipstick, short skirts, and swimsuits. In particular, their women workforce based in Iraq was uncomfortable annotating these categories due to *"cultural sensitivities"* (Irina). In these cases, EnVision Data either filters out some content so that annotators are comfortable, or they reassign the project to another location. Though data workers at EnVision Data have the ability to push back on projects, they are not involved in actually screening clients. All of the data workers at EnVision Data had never actually interfaced with a client directly themselves.

Overall, data workers were not involved or considered when screening clients. Considering whether a client was ethical, or what their culture was, did not lead to traditional tech workers discussing the implications the chosen client might have for data work. However, traditional workers did occasionally consider the backgrounds of data workers on a project basis. In the next section, I detail how traditional workers discussed specific expertise for data work.

Power to Select Data Services

Traditional tech workers have the power to think critically about who they will hire to conduct data work. Generally, clients select one of three types of data workers: (1) they hire data workers through a BPO (often referred to as a vendor); (2) they hire freelance worker from websites like Upwork or crowd work platforms like Amazon Mechanical Turk; (3) they employ specified workers with expertise in specific tasks.

The reason that traditional workers decide to outsource data work is that the work is considered below the pay grade for traditional workers. The datasets which are outsourced to BPOs or freelance contractors are for work that is considered too tedious for full time traditional workers to be using their time for. Jeremy described how *"outsourcing is typically done for the tagging of the more precise, the more labor-intensive tagging."* Jeremy works on gesture recognition and so requires data of hands in different positions with annotated keypoints. *"For example, tagging the positions of all the fingertips on every frame of an image is very labor intense. … it's still tedious no matter what you do, when you're tagging all the fingertips. That's what we typically outsource the most."*

How Traditional Workers Select Data Workers

Once traditional workers decided to outsource data work, they had to decide who to outsource that work to. Not only did traditional workers have the decision-making power to choose to outsource work, they also had the power to decide who they would select to do that work. In particular, traditional workers seemed concerned about selecting data workers who would not introduce bias to the data or produce low quality work. Selecting data workers included an assessment of how the positionality of data workers might be undesirable to their vision of the product.

Concerns about the bias of data workers came up often with traditional worker participants. In particular, they expressed that data workers were likely to be biased, or to express undesirable cultural perspectives. Due to this, some clients were looking to hire workers they viewed as more "expert" for data services. Kaleigh explained, *"We are moving more and more towards full time employees doing data collection and labeling. The priority for that is especially sensitive or high-risk scenarios. So, think carefully."* Similarly, Lynn, in discussing the famous Gender Shades work expressed admiration for the use of dermatologists for labeling skin tone. She said, *"She basically had a social scientist come in and label each and this is a professional person, not like an annotator in India, or something like that."* Lynn's perspective insinuates that outsourced data workers are less qualified than more traditional workers. More specifically, Lynn's commentary reflects a distrust of outsourced annotators in the Global South to be doing high quality data work.

It was not always clear whether selecting data workers based on their expertise was an available resource for tech companies. For example, Kaleigh, a program manager in a larger company, described that it might be possible *to "specify if people should have a particular background in order to be a fit for the task. And I think it probably just depends on the task, whether the person creating the task decides to opt for that."* However, she was not actually sure whether it was an option, or how to utilize it if it was. Elliot said that determining which vendors to hire for data services is something they have little knowledge of, given they have never been in the room. *"I'm not sure who exactly would be in the room for those conversions,"* they admitted, though they speculated that policy professionals and lawyers would likely be involved. Nonetheless, there are areas pertinent to computer vision development that some traditional workers are not invited to participate in.

Contracts with BPOs are often written that do not allow workers to collect information from data workers, so understanding data worker perspectives is legally untenable. Elliot described further legal and policy barriers to collecting any sorts of information from data workers:

> *"So sometimes a researcher might be like okay I want to work with this vendor but I also want to understand some basic sociodemographic characteristics of my annotator pool and then they'll actually be hit with a legal or policy barrier where the institution will be like no you can't ask these people for race, gender, or other sensitive characteristics. That's not in the terms of their contract, so it's not possible … Then depending on the regional context it might also raise identity concerns as well depending on political climates and safety climates. It's kind of a million different things of what on the surface seems like a really simple question of understanding who your annotators are."*

Elliot describes law and policy as a "*barrier*" to engaging in practices that make understanding the positionalities of data workers accessible. In the context of large companies, legal teams often set the terms of what is prohibited and what is allowed to avoid litigation—not necessarily due to actual governmental law. In these cases, the legal team has power over product teams and research teams, like Elliot is part of. Legal decisions constrain the practices that other traditional workers can engage in.

Most traditional workers were not super aware of how outsourcing processes works and did not interface directly with vendors or with data workers. For example, Beiwen describes how data services are requested at Zeta, a large tech company that regularly outsources data: *"[It's] beyond my knowledge. Our team is more focusing on the machine learning part. When requesting data, there could be some request form us to also annotate the data, and then the dataset team will take care of the request. They may—I'm not sure if they will ask somebody else—but when they return to us, they provide the data with annotations."* Beiwen only knows the pipeline for requesting annotations, but not how annotations are actually obtained.

Such gaps in knowledge were common for traditional workers, especially those focused on engineering, like Beiwen and Nitesh. However, some traditional workers did have an understanding of pieces of the selection process. To start, for traditional tech worker clients working in large tech companies, there are often limitations on which vendors can be used for outsourcing data work. Elliot described how Maelstrom has an approved list of vendors and contracting with new vendors likely required an intensive approval process. They were unsure how vendors were initially approved, or what the conversation about new approvals looked like, saying: *"A lot of tech companies, it's a lot of patchwork*

… but I don’t know, I have to speculate.” Processes seemed to be especially opaque in large tech companies, where workers are attending to bits and pieces of the development process.

While traditional workers have the power to outsource necessary yet undesirable work, like data work, they did not always have the power to choose which vendors they could hire from. This was largely because of legal barriers at their companies, to protect the companies from litigation or malpractice. Traditional workers in legal or policy roles were able to engage in decisions about vendors, but the workers who directly needed data work were not. Even so, often, traditional workers did not seem concerned about vendors. They were not necessarily worried about wrestling over power to select vendor companies. They still had the power to choose to outsource data work and to select from a list of pre-approved vendors.

Elliot complicates the issue of selecting data workers based on their identity characteristics. Not only are contracts often legally limiting, the reason for contracts are often tied up in political barriers. BPOs might be seeking to protect their workers from being identified for safety or political reasons. They might also be seeking to protect themselves from any legal liability if issues arise around identity characteristics.

Not all traditional tech workers turn to BPOs for hiring data workers. Crowdsourcing might also be chosen by some clients as a means of collecting a lot of data quickly from many users. However, crowdsourcing did not seem to be particularly desirable in comparison to vendors. Elliot expressed that the use of an internal crowdsourcing platform served to *“avoid compensating people.”* Crowdsourcing makes selecting annotators based on any perceived expertise more difficult, because often tasks are posted and any users can apply to work on them. While it is possible to limit demographic criteria on some crowdsourcing platforms, participants did not mention those options. Instead, they focused on how crowdsourcing was lower quality than BPO work. Lynn expressed that crowdsourcing tools that allow you to gather labels from a *“non-consistent staff,”* but they are *“low quality, highly biased.”*

Concerns about data worker biases differed from the perspectives of data workers themselves. Data workers discussed the benefits of diversity to improving the accuracy of datasets and the resulting models. Shokouh said:

> *"In my perspective, it will be so good and it's so much interesting that if we get familiar with other cultures, new people, it will be more fun to get to know other people in different countries, about them, about their experiences, we share ideas. I think it's helpful."*

Shokouh values expanding her own positional horizons from learning from other cultures. Similarly, Dinorah felt that issues in data work did not stem from culture:

> *"It's definitely not from the culture I can say. I can see that they're intelligent, they have this passion for knowledge. This is what makes me love my job even more. They're really trying their best when it comes to work and to quality."*

Makaarim stated that it was not the responsibility to account for the diversity of annotators. He said, *"Mainly the client is the one responsible for setting the guidelines or the criteria they are seeking from the company."* However, he did feel that the client should be employing diverse annotators, even if it is not their responsibility. He felt that choosing not to use diverse annotators for projects was *"directly contributing to creating problems for people"* in the form of biased datasets.

To summarize, traditional workers had the power to outsource data work. When they chose to do so, they sought to select data workers who they felt would not be highly biased or produce low quality work. Traditional worker views of what made a data worker highly biased or low quality was tied up in their perspectives on data worker positionalities. Traditional workers expressed that the data workers they usually hired in the Global South were not ideal, because they seemed to believe that they had lower expertise than in-house annotators, or annotators trained in specific areas (e.g., dermatology). Yet, due to legal barriers with vendors, many traditional workers could not specify the identity characteristics they hoped for in a data worker staff. Data workers, on the other hand, thought that a diverse workforce was ideal for improving computer vision. Even if they did not feel it was the responsibility of their clients to hire more diverse data workforces, they thought it would create better datasets.

How Traditional Workers take Advantage of the Economic Precarity of Data Workers

The selection of data workers gives clients the power to consider which types of data worker positionalities are appropriate for the task at hand. Given clients also hire based on profit models, they also have the power to set market standards for data work costs. As a heavily outsourced form of labor, prices for data work are competitive across the globe.

Rather than culture, Dinorah felt the true issue in terms of quality stemmed from economic issues and that the availability of work is always in English:

> *"Unfortunately, due to the[ir] status, … most of them … cannot even afford a laptop when they're refugees, which is understandable. So, yes, when they get the computer, we give them multiple courses, like… sometimes we give them courses in English because most of them speak only Arabic, only Farsi, because most of our annotators are Arabs and are from Iran."*

Rather than their culture being inherently prone to bias, Dinorah's perspective insinuates that their lack of access to reliable technologies is what makes data workers unreliable. Further, EnVision Data provides English language courses because data work is largely in English, and not having a grasp on English might therefore lead data workers to receive negative reviews from clients who find their work to be low quality. Most data workers do not have access to such courses, as provided by EnVision Data as part of their company mission.

Data workers would even take on projects they are personally opposed to due to economic need. Ghaliyah explained how annotators have to do their jobs, regardless of their beliefs:

> *"Some annotators, they are from countries… for example, drinking is forbidden there, you cannot drink alcohol. But I think as the annotator, you have to accept it is your job and it is not something illegal or something very bad, it is like the client's point of view."*

Even Ghaliyah disagreed with projects as a supervisor. She described disagreeing with the premise of the ChAI project. She said, *"Personally, I didn't want to spend time like, choosing, … to care about the person's appearance a lot."*

Irina described an early project EnVision Data took on focused on content moderation, and how it negatively impacted the only worker who agreed to do it. *"In the very early days, we actually did … dataset collection for content moderation. We had something like 5 people back then and only 1 agreed to work on it."* This data worker worked full time to collect partially nude and fully nude images. Irina expressed a sense of guilt for taking on the project, though EnVision Data was a new company with very little economic power. She described how this project in particular shaped her perspective on declining future projects:

> *"It was so sad, just because we were really pressed for money and we agreed to do that project. But it was very uncomfortable for him and also for his team. And I think it was also kind of damaging for him. He did receive good compensation for it so I think it made it worthwhile for him but it really wasn't an ideal situation, especially when you're working with a vulnerable group. One of the biggest tradeoffs is the position of the*

> *annotator and the effect that the annotation has on them and all of the content they are exposed to. Ever since then we try to limit the amount of [not safe for work content] people are exposed to."*

Irina explained, *"As a social enterprise, we're telling people we want to make a good impact on their life and we're all about ethics, [making people do that work] isn't consistent with our advertising."*

Globalization and low pay not only pits BPOs against one another, but individual data workers against each other. Nedeljko, a freelancer worker in Serbia, described his frustration with trying to compete with data workers in cheaper countries:

> *"The semantic segmentation was paid $5 per hour and the bounding boxes was $3 per hour, pre-COVID. After COVID, they laid off all of the, from my country, from Serbia, they laid off all workers. And now they are only paying Philippines and they're paying, I think $1, per hour. The same company … All the managers were from Serbia and also they laid off."*

Lucano described how competitive it was to even get paying jobs for data work online: *"It was a first come first serve basis, so if I get there and do the job I get paid, and the ones that didn't have the opportunity or were asleep at the time, they didn't got any money."* Individual workers compete for jobs so that they can make an income at all, and those who are unable to commit to staying up long hours to quickly try to sign up for jobs simply do not get paid. Further, besides already paying low wages, clients would sometimes scam data workers out of their earned pay. Lucano describes how he had been taken advantage of by a client in the past, who did not pay him fairly for the work that he did. In particular, the client tried to develop an exploitative relationship with Lucano simply because he was from the same country as him.

Workers also often had to take on extra costs themselves. For example, Rebecca did a project for EnVision Data that involved printing documents. She described being afraid she would not make a profit due to the cost of printing.

Economic precarity also took a toll on data workers' bodies. Lucano explained how doing long hours at the computer led him to having back and neck issues. *"I had a lot of pain. That's the cost of working a lot of time on my computer,"* he explained. He described how he and colleagues would work ten to twelve hours or more, staying up through the night until 5 AM in the morning. He discusses the negative effects of the work on his body as a "*cost*" to doing the work, which is already low paid. Lucano, in particular, expressed complicated feelings about data work. While he felt he could not blame clients for worker conditions, he also felt it was their responsibility for structuring data work in such an exploitative manner:

> *“Maybe [the data workers] had no choice [but to work long hours]. It's very sad because—well I don't want to get philosophical. If a person wants to sacrifice some of their health because of that … I can’t judge the company, because they are not obligating the people to work like that, but it's a result of how they implemented, how they set up the platform … At least give the qualified annotator a stable income. I know that this is hard, because most of them are far away, they crowdsource this task for people in India or Africa or South America, Third World countries that they don't have to pay thousands of dollars monthly in a single person. They spend like 200 plus for the person to work diligently on the project. I wouldn’t say it is slavery, because they aren’t obligating the person … I don't know if it's a group of people decide to start work for those companies, there are a lot more people that don't have an income or that don't have stable job, and wanna make more money and this they look at this like a good opportunity. the companies are always going to find those people who need this kind of job and I don't think there's nothing to stop them from doing it.”*

Lucano believed that data workers had some level of autonomy, in that they did not have to take on jobs, but at the same time, expressed that they might not have a choice because of economic conditions. Therefore, there would always be workers willing to take on low paying work that damages their health. He implicitly expressed that such work conditions were similar to “*slavery*,” except that companies aren’t forcing people to work in the traditional sense of slavery. Instead, companies take advantage of the poor economic conditions of specific countries and drive down prices further.

Poor economic conditions also had a gendered component, as well as a geographic one. Many women data workers described how difficult it was for women to find work in their respective locations. Sumbul and Shokouh both described the current political conditions for women in Afghanistan as preventative of them being able to work outside of the home, meaning online work was the only option for them. Shokouh explained, *“Whenever you go out from your home the stress begin until you come back to your home. It is a very normal thing for every Afghan woman and girl.”* Sumbul similarly described the state of work for women: *“Especially for women and girls here in Afghanistan, that we cannot work outside home and it’s very difficult for us, and it was online work, that’s why, we were also jobless.”*

Sumbul said, because of her own experience as a woman, she thus enjoyed working with fellow women. *“Because I am also a woman, I know which difficulties, which challenges have in our society, in our community, in Kabul, especially in rural provinces. I know which challenges do a woman face with that, that’s why I’m really interested to work for them, and also to serve them,”* she explained. She also described helping widow women with job training so that they could continue to bring in income and help their families.

Makaarim felt that clients should acutely hire groups struggling with economic and political conditions. *“We know there a very underrepresented communities because of their conditions whether its economic or political so they will be deprived of such opportunity, and when you empower them to have access to this information it shows sympathy and strength, humanity.”* He felt that attending to worker needs is a sign of “*humanity*.” However, economic precarity for data workers was also fueled by tech companies devaluing necessary data work. As Kaleigh explained, her team had asked for *“double [their] previous budget.”* The reason was *“because that's what it's going to cost to actually do fairness, at least doubling data budgets, if not more than that. And that doesn't even include all of the extra headcount that's needed to be able to manage data collections in a much tighter way.”* She describes that she doesn’t want to approach data work as *“just toss it over the fence to a vendor and get it back*” but that teams are constantly *“compet[ing] for resources”* at Aqueous. Zephyr recognized that data work conditions would likely not improve over time, and that *“for most people, annotation isn’t a career*.” That is why she helps develop vocational training programs for workers at EnVision Data.

While data workers were paid poorly, they also often didn’t have many other options due to the economic conditions of their country. Given economic precarity, data workers generally do not have much power in selecting clients. Many data workers could not get other work, due to the economic or political status of their country or because of their visas as immigrants or refugees. Outsourcing data work reflects not only the decision-making power of traditional workers, who don’t want to take on the work themselves, but also the economic power of workers in the Global North. Those data workers in the Global South often did not have other viable options for work. Rarely did conducting data work offer a sustainable career path. Thus, social good companies like EnVision Data, while a data BPO itself, offered vocational training to help its workers find better career opportunities. Meanwhile, most companies had set norms for outsourcing in ways that take advantage of cheap data labor, and traditional workers, encultured in that environment, regularly rely on those norms.

Power to Design Data Practices

The main mechanism through which clients impart their perspectives about data categories onto data workers is through *data guidelines*. Guidelines, also known as instruction manuals, are documents where clients formalize their expectations about data categories. Siddhartha presents a sort of overview of how guidelines are used as instruction documents: *"If you want to train models with the same data, [the data team makes] sure that the [data workers], they are following the guidelines that they were told … they know everything that they're doing."* Clients often described guidelines from this perspective—as a set of instructions which outline the rules and expectations for data collection or annotation processes. Yet, guidelines also acted as mechanisms for controlling data worker positionalities.

How Traditional Workers Design Guidelines and Training to Control Data Worker Positionalities

Guidelines serve as documents meant to communicate traditional worker expectations about identity to data workers. Guidelines are reflective of traditional worker positionalities and are meant to enculture data workers to their positional worldviews.

Both Lynn and Vasudha worked on image classification models that included wedding concepts. Wedding concepts proved difficult specifically because of interpretations around same gender couples. Lynn explained her first introduction to dealing with guiding data workers to be "*unbiased*" was in training this wedding classifier for a B2B customer. She described the concepts she had to get across to data workers in the guidelines: *"It could be two women, it could be a man and a woman, they can have different skin tones … there are different races and different garb for weddings."* Much like Lynn, Vasuda similarly had difficulties with queer concepts when data workers were in cultures where queerness was taboo—or where current norms were changing. She explains: *"Like gay marriage became legal in India very recently … So, we can guide them to say, if two people are getting married, and both seem to look like women, tag them both as bride."* Vasudha speaks about guiding data workers to see two women as brides. While she does so in the context of applying data labels, training also works to shape the perspectives and cultural outlooks of data workers.

Differing cultural conceptions could often clash, making this goal difficult. One of the company's customers requested a moderation dataset for classifying explicit versus non-explicit content. The customer, based in the United States, expected nudity and pornography to be labeled as explicit, while other content would be labeled as non-explicit. For the explicit photos, she expected *"there would be nudity, right, like the American concept of explicit … you know, X-rated stuff."* However, they found that the annotators based in India labeled concepts outside of nudity and sex as explicit:

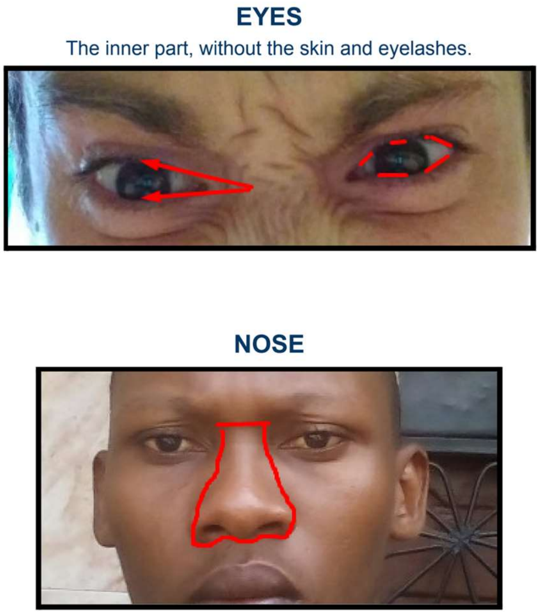
> *"So, we get our data set back, and we start combing through it. And all of a sudden, we start noticing that man on man, whether they were fully clothed, or kissing, or even just holding hands came back [labeled explicit] … And so we had to train this overseas workforce to say, `We don't consider homosexuality to be explicit or taboo.'"*

Here, training data workers to see data in a specific way—for example, that two men could kiss or two women could be brides—extends beyond the data. Traditional worker positionality is meant to shape the positionalities data workers inhabit. Training and guidelines cultivate the positional values of those in the Global North in individuals who might otherwise not be exposed to such values.

Given that many traditional worker participants didn't directly discuss how guidelines are created or written, examining examples of guidelines illustrates how they can be designed to reflect traditional worker positionalities. Often, this is because guidelines were not specific—they simply reflected back implicit assumptions about identity. As discussed in Chapter One, instructions and examples were not given for annotating demographics like gender or racial categories. Instead, annotators were given a list of options to choose from, such as "male / female / unsure" (ChAI project). Vasudha said that it was difficult to define in guidelines how to annotate identity categories like gender, which reflects data workers being equally unable to explain how they make categorical decisions about identity categories (e.g*., "I mean, it's just that I know that they are somehow"* (Manjola)). Like most clients, Vasudha describes telling data workers "*what you perceive is what you perceive*," yet she also explains that they are constrained to the categories of gender that they are given.

In many cases, guidelines are actually designed in ways which are so vague they fail to account for possible interpretations beyond traditional worker worldviews. Such "straightforward" guidelines suggest that clients may fail to predict when something is not actually "obvious" but is instead highly interpretive. For example, **Figure 16** is a snippet of the instructions provided to Nedeljko for a facial segmentation project. The instructions are simply a series of examples for selecting different facial features. Yet, facial feature annotation was one of the most variable project types among data worker participants. Bernardita, Wares, and Shokouh all expressed difficulty discerning different facial features on Black faces, while Lyonis, Pelumi, and Raiha expressed having a hard time selecting eye regions for East Asian faces. These workers came from geographic areas where seeing such faces was rare, and so they were unfamiliar with them. The guidelines given to Nedeljko—and the previously mentioned annotators—did not provide multiple examples across racial dimensions, indicating that perhaps clients did not expect data workers to find the task difficult. Beyond racial difficulties, even examining the outline of the nose in the bottom image of **Figure 16**, one could imagine another annotator interpreting the visible shadows of the nose as the bounds rather than outside of the shadows. Traditional workers, in designing the guidelines, were not thinking about whether data workers were familiar with different types of faces. Embedded in the guidelines is the implicit assumption that different types of faces were equally interpretable for data workers.

Facial Segmentation Example



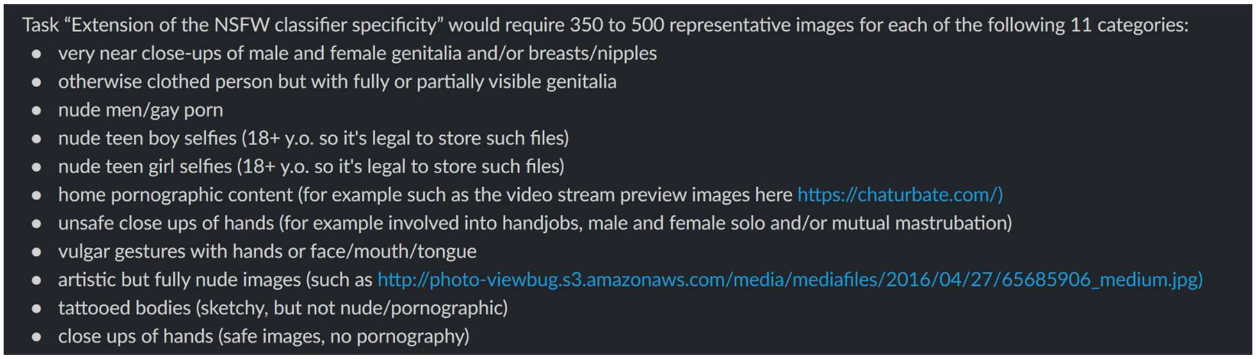
EYES

The inner part, without the skin and eyelashes.

NOSE

Figure 16. Facial segmentation project from Nedeljko.

CodeGuard Example



Task "Extension of the NSFW classifier specificity" would require 350 to 500 representative images for each of the following 11 categories:

- very near close-ups of male and female genitalia and/or breasts/nipples
- otherwise clothed person but with fully or partially visible genitalia
- nude men/gay porn
- nude teen boy selfies (18+ y.o. so it's legal to store such files)
- nude teen girl selfies (18+ y.o. so it's legal to store such files)
- home pornographic content (for example such as the video stream preview images here https://chaturbate.com/)
- unsafe close ups of hands (for example involved into handjobs, male and female solo and/or mutual mastrubation)
- vulgar gestures with hands or face/mouth/tongue
- artistic but fully nude images (such as http://photo-viewbug.s3.amazonaws.com/media/mediafiles/2016/04/27/65685906_medium.jpg)
- tattooed bodies (sketchy, but not nude/pornographic)
- close ups of hands (safe images, no pornography)

Figure 17. NSFW moderation classification (Codeguard) from EnVision Data

Similarly, the instructions for the Codeguard project, which involved EnVision Data workers collecting examples of NSFW content for a moderation classifier, was simply a list of content they desired images for. The list itself was non-specific, making instructions highly interpretable. For example, "vulgar gestures" are not only subjective from an individual perspective, but from a cultural perspective, as "vulgar gestures" vary widely by country. Similarly, determining whether a nude image is of a "teen" or 18 and up is highly subjective. It was also risky for the annotator who might be searching for such images, as they might be exposed to illegal underage content or culturally inappropriate content. Vasudha described the issue of cultural interpretation in her own experience with content moderation:

> *"So, when you think about the different aspects here, the one thing that we struggled often is sort of to draw the line between what is racy and what is adult. And in cultures like even wearing like, swimwear is still very offensive, like in Indian culture … And this is it all, not knowing what countries would have certain limitations has definitely been a challenge."*

Here, Vasudha is actively acknowledging the limitations of traditional workers' knowledge of other cultures. These limitations come through in the vague and non-specific nature of identity categories in guidelines.

Yet differing cultural interpretations seemed to be an area of concern for traditional tech workers. Both Lynn and Vasudha described scenarios where they dealt with different cultural interpretations of the data, which led to undesirable outcomes in a US context where the product would be deployed. The knowledge that cultural biases around identity categories could infiltrate datasets outside of the explicit annotations themselves led Lynn to be more careful with training annotators outside of the customer's cultural context in the future. Lynn said:

> *"I can tell you just having worked with overseas teams for so long, it's like whack a mole, like, as soon as you find a sense of cultural bias, and you just kind of have to peel away the layers of the onion ... But it's not to say that they are biased, right? And I'm not saying that their bias is problematic. They're a different culture from us. And they're, you know, they're coming at it from the way that they, when I say the word bias, I mean, you know, I'm meaning in the, in the statistical sense in that, you know, as I mentioned, every human is biased."*

Lynn acknowledges that identity work is subjective in that *"every human is biased*." At the same time, she maintains that there is still a desirable outcome for identity in models, in that she is still seeking a statistically correct outcome. As such, she seeks to control how data workers apply labels, to avoid this sense of statistical bias. Yet, since statistics are dependent on how a problem is framed, Lynn is overlooking that the desirable output is still reflective of her own cultural bias. Examining this statement reveals that traditional workers often view data workers as introducing statistical bias, in the sense that traditional worker expectations are statistically fair.

At EnVision Data, clients occasionally came with pre-written guidelines. However, oftentimes, clients came only with requirements or desired outcomes for the project, and the annotation trainers, like Yasmin and Ghaliyah, translated those documents into guidelines for the data workers. Larger tech companies which regularly outsourced data work also often wrote their own guidelines and worked with fleshed out pipelines for writing project guidelines. Vasuda discussed how at Aqueous, they intentionally ensure the guidelines are "*understood by the global audience*" of data workers they use all over the world. "*Most of our AI vendors come from China, from India, from Brazil, because our AI teams sit there*." They have a pipeline for "*all of the logistical work of now getting like translating it into language that is easily understood by the vendors*." In many cases, for clients who are not located in large tech companies, instructions are given only in English. Given that the majority of data workers I interviewed would often need to translate pieces of our conversations, it is likely they often had to translate English-only instructions. At EnVision Data, the annotation trainers purposefully translate instructions into Arabic given the majority of their workforce's native language. Freelancers hired for data collection projects, like Thahn, also translated English instructions to build better rapport with data subjects. Translation endeavors add another layer of interpretability that may lead to differing data outcomes. For freelance workers, like Thahn and Dinorah, the labor of translation and interpretation also fell on them, as clients did not provide translation services.

Of course, defining guidelines was not always viewed as a simple task. Some traditional workers acknowledged that guidelines were documents laden with perspectives. Vasudha explained that defining guidelines is difficult, because people have such different perspectives on their application: "*I don't think we are in a world where all of us are aligned on what those guidelines should be, just because global population perceptions are different perspectives are different."*

Traditional workers are tasked with determining, to the best of their ability, what should be communicated in the guidelines so that the data can be used in expected ways. Guidelines are meant to be used to ensure data consistency, objectivity, and quality. They are designed to enculture data workers to traditional worker positionalities, to train them to see the world through traditional worker eyes. However, in practice, communicating clearly and strictly through guidelines is difficult. As such, many guidelines are extremely vague, leading data workers to attempt to synthesize their own positional worldviews with the expectations of their clients.

How Data Workers Apply Data Categories

Data workers do not have the power to define what identity looks like in the overall task. By the time the project gets to the data workers, identity has already been defined and translated into technical requirements. The training material tells data workers what to do to implement this vision of identity. As Pelumi said: *"I use the training material as what is truth."*

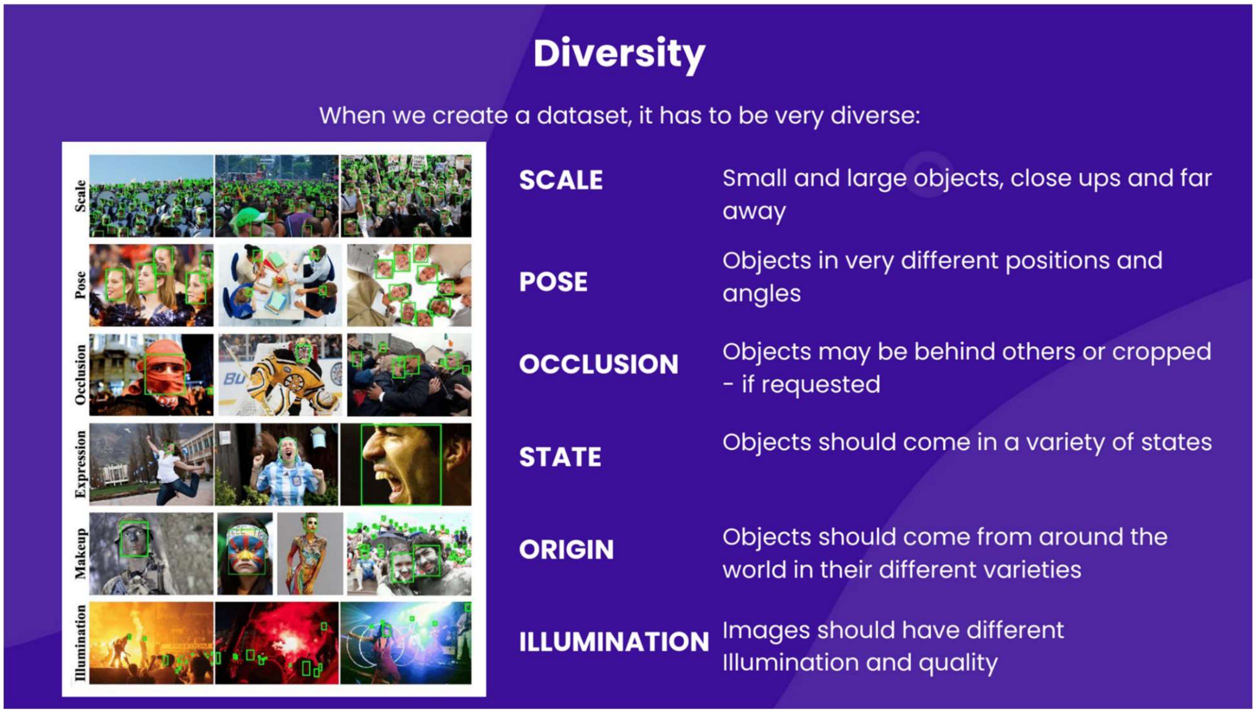
Ghaliyah, who works as an annotation trainer for EnVision Data, described how for each project, a supervisor walks the annotation team through the labeling guidelines. She acknowledged that the annotator's point of view may differ from the guidelines. Especially for some projects, which ask the annotator to make decisions about subjective or imprecise concepts, like how fast a person is speaking, it is difficult to prescribe a correct answer: *"At the end, it depends on the annotator's point of view. So, there were no 100% rule for some questions. For example, some questions were like, 'Is she or he speaking fast or slow?' … We cannot say 100% if this is the correct answer or not."* In actuality, what was considered "correct" was whether the client agreed with the data they received from data workers.

Largely, traditional workers didn't discuss data worker positionalities in the application of instructions; rather, they often talked about controlling for bias at a high level. For example, Nicholas explained, *“We know that AI can be biased if you have bias in your data, that we want to make sure that we want to remove the bias in the data … the biases that people inject into the process that they're not even aware of, you heard the term unconscious bias. Right, so we start to look at what those are.”* He discussed the concept of biased data at length, and how Resoom worked to control bias in the data, but he did not discuss data workers directly. Instead, both Nicholas and his colleague, Lydia, regularly discuss the positive role of having diverse teams of traditional tech workers to keep bias in check. In one example, Nathan explains that Amazon’s biased resume parsing model *“probably wouldn't have gone as far down the road as they did without checking for biases”* if they had a team with diverse disciplinary backgrounds.

As Kenny describes about the annotation process, *“[data workers] make a logical human decision. So, here's the thing about computer vision, if you can't figure it out with your own eyes, then the computer vision will never be able to figure it out. That's the base level.”* While Kenny explains that data workers are making “*logical*” human decisions, he is more so implying that computer vision can only reflect the positionally-informed decisions of human beings.

EnVision Data implemented a new required ethical AI training module in 2021 to help better keep “*the bias out as much as possible*.” The module was aimed at teaching their data workers to think critically about data possibilities, thinking outside of their limited positionality to try to imagine other ways of classifying the world. As seen in **Figure 18**, this might include how different objects—like food items, in another slide—differ across the world.

EnVision Data's Ethical AI training



Diversity

When we create a dataset, it has to be very diverse:

SCALE	Small and large objects, close ups and far away
POSE	Objects in very different positions and angles
OCCLUSION	Objects may be behind others or cropped - if requested
STATE	Objects should come in a variety of states
ORIGIN	Objects should come from around the world in their different varieties
ILLUMINATION	Images should have different Illumination and quality

Figure 18. A screenshot of EnVision Data's ethical AI training module.

Sadhil felt that annotators learn how to better annotate and recognize different concepts across cultures through experience annotating. He explained:

> *"These things come with experience. In simple terms, we can identify any object, like I can identify this is the thing called a chair, this is in front of me. So, who taught me this? So our parents taught us this, that this is chair, so in our brains, there are neurons. So my parents told to me five to ten times or fifty times, 'This is chair, this is chair, this is chair.' So I can identify this is chair."*

Sadhil compared himself—and other data workers—to how computer vision is taught to see patterns, reflective of Kenny's perspective that computer vision can only reflect what humans teach it. Yet Sadhil also acknowledged that annotators "*make mistakes*." However, he blamed these mistakes on annotators forgetting the goal of the model, rather than for either poor instruction or differential perspectives interpreting guidelines.

But guidelines, though they act as the "truth" which data workers are aiming to apply to data instances, often fail to account for all of the "edge cases" that show up in data. Pelumi, a freelance annotator that also oversees some projects as an annotation supervisor, explains his approach to identifying confusions in the guidelines. When he notices certain patterns, he turns to the guidelines to see if the annotations match the instructions. If they do not, he will talk with the annotators to identify the core issue. Sometimes, this leads to the annotation instructions being too confusing or too vague to cover the issue*. "Any slight change in the instruction, the client may be affected by it. So, when I agree [to changing the instructions], I quickly inform the client that there was this case, the training instructions didn't quite explain it, so we have decided to consider it this way. Do you agree or will this work out for you?"* If the client approves, then the annotators re-annotate those Attributes in accordance with the updated guidelines.

Like Pelumi's case, at EnVision Data, the supervisor may provide clarification themselves, reiterating or reinforcing their own interpretation of the guidelines. The supervisor has the role of mediating what annotation should look like based on their knowledge and interaction with the client. Yasmin, who supervised the ChAI project, explained how she would have to address confusions about specific expressions. "*Like the most edge cases that we faced were about the expressions of the people in the video, for example, if they were enthusiastic or not, or if they, for example, if they smiled more than once in this two minute video."* While she acknowledged that such expressions were subjective, she still had to guide annotators to understand the expectations of clients and correct "mistakes."

After all, guidelines regularly fail to account for different interpretations of the same "truth." Sadhil described a mix-up about what the item of clothing, a "blouse," was, because it looked different in India than in Japan.

> *"In India, a blouse is called a woman's dress … like half sleeve t-shirt, like, thing is called a blouse in India. And in Japan, a blouse is like a full quarter gown, so this was a different thing. In Japan, different clothing type is called Blouse [than in India]. So first, when we prepared the data, we collected the images according to blouse that we [know] in India, because I don't know that in Japan the blouse is the other thing. Then he [the client] said this was not the blouse. Then after searching on the internet about the Japan and their clothing, the blouse in Japan, then I understood that the blouse in Japan is the other thing, then I collected the Japanese blouse images … And many other things, same in any country, if you are working on human fall detection, it's the same in all countries. But when the thing is for the clothing, it can be different in different countries."*

Sadhil did quite a bit of work based on his intuition of what a blouse looked like in his own country of India before the client corrected him. He then had to conduct research to understand what a blouse looked like in the context of another culture. Gemma similarly described being stuck between multiple categorical options because she did not understand how to apply the labels. "You cannot differentiate between a Caucasian person and Hispanic person, so it was sometimes a bit challenging," she explained, implying she did not find the categories meaningful. Dinorah had a similar reaction to ethnicity categories: *"We have to be superficial. We have to see the color of the skin, we have to see, let's say, if we see someone super dark, we should write African, we cannot write Caucasian, even if he is from this origin, we cannot know this. We should just answer whatever we see, there is no other indication."* Such categories could be reflective of U.S. perspectives on ethnicity that are not common in her home country of Kenya.

Gemma also had to classify gender in another project, and she ran into some issues where she was unable to tell the gender of the person. *"Mostly it's their physical appearance, and then sometimes, it's really hard to tell someone's gender because of their sexuality, someone might apply makeup or is transgender, so then you can't know really the gender of that person,"* she explained. Gemma would offload the responsibility to make final decisions to her client. *"So, in that situation, mostly I just ask the client, cuz that's something you can't just tell."* In those cases, the client usually looks into the data themselves, and she no longer has to deal with it. Given Gemma had no power in defining how gender identity should be represented in the first place, the only options she had were to guess within the parameters of the categories she was given or simply let the client decide. She did not have the power to define the categories, and she did not agree that the categories fully accounted for gender identity as she perceived it. Letting the client decide implicitly acknowledged the power the client had over the entire process. Binary gender was an attribute designed to serve client needs, and so the client was the best person to make determinations when the data worker was unable to fulfill their role as inhabiting the client's positionality.

During the process of annotation, a number of workers expressed disagreement with the processes of annotation. In particular, they felt that annotations were often subjective, and thus it was unfair or unrealistic to expect a specific answer. For example, Abyar didn't necessarily agree with supervisor decisions. She felt that every person sees things differently, which didn't necessarily mean she was incorrect. However, in the end, she felt supervisors "know better" because they interface directly with the client. *"Sadly, yes [the supervisor has the final say] … but yeah, according to the supervisor, cuz they actually have direct contact with our client, so they know better."* Aakrama had even stronger feelings than Abyar. He explained that he felt the ChAI project, in particular, was "ridiculous" because there was no way to accurately rate the qualities of a person:

> *"[The] customer expects us to find the collective of the person in the short time of the interview. We must recognize if the man is stressful or very relaxed … many different fields that must be filled out. … many details must be suggested from the short movie. And it's very confusing because anybody can have a different idea about one person. Maybe I see this person and say he's nervous, maybe you don't say nervous … because the reality … because the rules is not certain … because anybody sees the very short, between the one and two minute interview, the customer expects us to find the personality of the [person] and make a number from 1 to 10 about this person."*

Aakrama felt the client's expectations were unreasonable, because there was no objective way to assess the qualities of a person in a short interview video. Further, he expressed concerns that people whose first language was not English might be seen as more "nervous" simply because they are focusing on speaking a second language. Much like with other categories, like gender, Aakrama was not given clear detailed instructions for how to assess categories in the project. Further, the categories were structured as assessments, asking data workers to rate the qualities of the person in the video. Therefore, the work was presented to data workers in a way that implicitly asked them to embody *their own* positionality, rather than the client's. This seemed to make data workers feel uncomfortable, because it more clearly exposed the subjectivity of their tasks. Data workers, encultured to embody the desires of their clients, were rarely asked to assess categories from their own perspectives in such a blatant manner. In the end, data categories are meant to reflect the desires—and positional perspectives—of the client. As summarized by Abyar, "We did whatever the client wants."

Power to Evaluate Data Work

The final step in the data work process is to evaluate the submitted datasets, checking whether collection or annotation quality is up to the expectations of the client. Data workers submit datasets to their clients for review, and then the clients ensure the dataset reflects their desired perspectives. In this section, I describe how traditional workers define the processes for evaluation datasets. I then describe how traditional workers largely shift the responsibility for any data bias onto data workers, given data workers were the ones conducting the data work.

How Traditional Workers Define Evaluation Procedures

Supervisors reviewing work before submitting it to the client was very important to EnVision Data, because negative client feedback could reflect poorly on the BPO as a whole. As Wares stated, *"Supervisor reviews and quality of work should be very good before submitting to clients."* However, not all data workers had a supervisor to submit work to. Sometimes, clients did the final review of the work. Clients doing final reviews was especially common for freelance data workers. Ghaliyah describes how the process of evaluation generally works at EnVision Data:

> *"Before submitting your works, we tell annotators, please check it multiple times … but after they submit their works, they cannot see their works anymore and the supervisor have access to the work only. The supervisor will go through all the details, all of their works. If they have problems, they will give back their works and they will work again on that. But in some cases, I think the supervisor will make corrections."*

The actual process of evaluating work looked different depending on the client. Some clients decided to use consensus models, assuming that multiple workers annotating the same data would lead to less biased outcomes. For example, because human beings have differing perspectives about the same concepts, Kenny *says "that's why we use three people … to keep the bias out as much as possible."* The purpose of using three people, rather than two, is to include a tie breaker. This meant that annotations were ruled by the majority. When two out of three data workers labeled one way, the final outlying data worker's perspective would be overruled. This arrangement minoritized outlying perspectives, assuming that disagreement with the majority indicated incorrectness.

Irina explained that they had tried automatic quality control techniques, but they did not work well. While they are more cumbersome and time consuming, Irina believes that human measures of quality are more effective than automatic ones. She said, *"What works better is having a supervisor do a double checking of the work."* She then believes that the supervisor should communicate with the team what the mistakes were, so they can learn from their mistakes. She explains that this process is "more human and addresses annotators as sentient and thinking beings instead of just this random crowdsource person."

As a supervisor, Dinorah had to view all the videos in the ChAI project where disagreements showed up. *"When there is a conflict … this is where I come and I look at the video and I say, okay, that is more likely to be correct."* However, she did not review data where all of the annotators agreed, indicating that a shared perspective was treated as truth.

As described above, some data workers disagreed with supervisors about their annotations being correct. In some cases, they managed to convince their supervisors that their perspectives were correct. For example, Sumbul described how when working on the Emovos project, she would try to prove her perspective to her supervisor:

> *"I just personally [told] her that this was my observation … for example, the disgust is the feeling that you dislike something or someone, and fear you are in … danger … These two expressions we had a small discussion between each other and sometimes [my supervisor] agreed with me. I keep the same [label] for that picture."*

Of course, supervisors did not always agree with data workers, and had the final say in whether a label would be updated or not. The model of EnVision Data, where data workers interfaced with supervisors and were valued as employees of the organization, allowed for data workers to express such disagreements. However, freelance data workers often operated much differently, rarely expressing disagreements with their clients (e.g., Gemma's example in How Data Workers Apply Data Categories).

Sometimes, clients actually decided to override the classifications that data workers assigned to data. In the SensEyes project, many of the ethnic categories for the data collected were reclassified by the client. The client, based in France, disagreed that the ethnic categories matched their expectations of appearance and would manually change the categories. Solange, the client representative, explained that they were less interested in the accuracy of the ethnicity of the individual and more interested in having diverse appearances. Traditional workers at SensEyes reviewed each piece of data and updated those they disagreed with. They most often reclassified Latin American data, because they did not want to classify white appearing people as Latin American for the purposes of their dataset.

Quality checks were also another area that more negatively affected data workers than traditional workers. In some cases, the quality checks were so opaque to data workers that they did not know which data was rejected. In the SensEyes project, this resulted in data workers needing to find a whole new batch of participants to collect selfies from, because they had no idea which selfies had been rejected in the first place. Manjola described having to continuously do data collection on the SensEyes project because so many people, especially older people, had a difficult time holding still. Her selfie data would then be rejected by the client.

The rating systems on platforms like Upwork might also negatively impact workers' abilities to get future work. Rebecca explained that because she could not complete all of the work within the expected time frame, she was given a negative review on Upwork. She explained that EnVision Data, who she was contracting for, told her they could not provide a five-star review for her incomplete work:

> *"They gave me a heads up that if you will not be able to complete the project, I will not be able to prove you that five-star review for this specific project. And knowing that I'm out of control with that situation I just had to say it was okay with me."*

The lower her reviews, the more negatively it impacts her job success score on Upwork, leading to less work. "It is really a huge factor." However, Rebecca did not try to convince EnVision Data to rate her more highly; she simply accepted the lower rating.

Much like in the other parts of the pipeline, evaluation procedures were entirely defined by traditional workers. The goal of evolution was to ensure that the submitted data reflected the visions of the traditional workers' initial requirements. In very few cases, data workers made the case that their interpretations were actually correct and not incorrect (e.g., Sumbul). Such examples reflected whether organizations were structured to empower data workers to speak up, like EnVision Data is. However, largely, data workers did not express their own opinions or attempt to convince clients their work was valuable. Even in cases where they got lower ratings, which could affect their ability to get future work, data workers largely accepted whatever outcomes the client decided was fair.

How Traditional Workers Shift the Responsibility for Data Bias

It was at the evaluation stage that notions of bias were often identified by clients. Vasudha had such experience with auditing models that she knew to expect potential bias for the populations doing the data work. She explained, *"We did notice wherever the model is trained, you're heavily biased for the population, our models predominantly were trained in Asian countries. So, it's tuned for that accuracy is amazing for that subset of the population."*

The responsibility for bias often fell on the data workers, given the notion that data is the source of bias rather than the data categories themselves. Perhaps some of this stemmed from traditional workers not wanting to take on the responsibility themselves. As Coleman said, *"It's a little bit like the history of the nuclear bomb. Everybody wanted to research on this stuff, and nobody wants to be* *responsible for having built it."* That annotators were responsible for bias often came through when traditional tech participants discussed the role of data workers. Elliot described a hypothetical image captioning system which was biased against women:

> *"Now imagine you have a system that takes in an image and captions that image for example, so it says, like, takes an image, maybe there's like a person in scrubs, you know, interacting with like some patient. And we find that if the person in scrubs has short hair and a beard, then the thing captions that there's a doctor talking to a patient. And if the person in scrubs has like, you know, long hair or some other like, you know, I don't know like other feminine signifiers that the dataset might be picking up on it'll caption it, there's a nurse talking to a patient. … it could also be that your annotators had a whole bunch of like biases. And so, the people who are annotating those images looked and were like, Oh, I think that this is like, I believe this is a man. And so, this is a doctor, I believe this is a woman, this is a nurse, and like, not consciously, but just like, right, like these things get filtered in."*

Responsibility for AI falling on data workers was internalized by those in the data workspace as well. Irina taught her data workers that this was the case during training modules:

> *"Obviously if they make mistakes, it will be really bad for AI, so we should be responsible. So one of the most important things for being an annotator is to be responsible for the things that you are doing so you have to be patient, you should care about your job, and you should check anything, because your behavior, your choices will have an effect on AI."*

Data workers internalized this. Dinorah believed every annotator must take responsibility for themselves and their positionalities:

> *"My vision is that everyone is responsible for himself … For example, see, my English is not that good. I am responsible for this; I cannot blame anyone that my English is on this level. I do not see here that culture affects that much, when it comes to development or giving something to a company."*

In her view, Dinorah believes that one's perspectives or experiences should never negatively affect the outcomes of their work.

Often, clients did not even check ratings when they received datasets. Yasmin said the client for the ChAI project never asked why annotators assigned certain ratings for each category. Yasmin described that the project was so subjective, that the client probably didn't see a reason to check each answer*. "It's just the human perspective on how the interview went, there's no right or wrong answer,"* she explained, indicating that the model which would be released was equally as subjective.

However, some traditional tech workers took on the responsibility of pushing back on managers before deployment. Coleman's job includes approving models. *"I have had situations where of course, I* *had to tell people in the way above my paygrade, no, I will not approve this. And then it's either meant sometimes they would basically follow my approval or non-approval note and say, okay, and sometimes they would override me,"* he explained, indicating that he did not always have enough power within his company to push back on model releases. However, he described building that power through seniority: *"The longer you are in the company, the more senior you get, the more people listen to you. And sometimes you just need to make sure that your management is on your side."* Traditional workers, unlike data workers, are able to build reputation and rapport with those above their stations.

Despite the fact that data workers were usually viewed as the core source of biased data or poor performance, most data workers had never even interacted with the types of models they were helping to train. Given that workers never get to interact with their end products, Irina and Yasmin both expressed a desire for data workers to be more included in the overall process of development. Yasmin said, *"As a community we need to think about to make annotators more involved about what will happen to the data after. On a core team level we are also aware of what will happen to the data most of the time but it may be very surface level."* Irina similarly had a goal to make human annotators more visible in the overall process of AI development: *"It's just acknowledgment of the fact there is a workforce behind this amazing AI model that's being created … And the clients are very much aware of that, our clients, … but when they are presenting that to their end clients, that's maybe where a bit of more acknowledgement [should] come through."*

Yasmin and Irina both felt that data workers were a core part of the computer vision development pipeline but were not acknowledged for their work. They want to see data work acknowledged by the end clients who use the data to train their models. Further, they hope to see data workers involved more in the overall process of development, from start to finish. They wanted data worker roles to be expended beyond solely collecting and annotating data in an almost automated fashion.

In sum, although the majority of decisions about identity concepts throughout the development pipeline are made by traditional workers, traditional workers largely shift the responsibility for any biases in the model to the data and thus the data workers who conducted the data work. The data workers themselves also internalize the view that data workers are responsible for bias, given the ingrained culture of perceiving data as the core source of machine learning bias.

Discussion

Given every individual has their own positionality—the way they view the world—then workers are always applying their own worldviews when developing computer vision. In this Chapter, I examined how different types of workers enacted their positional perspectives during the development of computer vision. Through the examination of how different workers referenced their positional perspectives, I identified how *power* manifested during the development of computer vision artifacts.

In the end, the entire development process of computer vision is controlled by traditional workers. Even while traditional workers had differential power dependent on their positionalities to shape identity, navigate legal barriers, and push back on managerial decisions, the culture of traditional tech work allowed for interpersonal negotiations and accruing decision making power. Further, computer vision work is structured so that all decisions are controlled by traditional workers—from defining categories, selecting data services and clients, training data workers, structuring guidelines, evaluating data, and even placing responsibility onto data workers.

On the other hand, data workers have very little power during the development process. They do not have the ability to define data categories, determine training or guidelines, or evaluate data. Those data workers at BPOs, like EnVision Data, aren't even involved in screening clients. Due to economic precarity, even freelance data workers are forced to choose between accepting undesirable projects and their incomes. The only place in the process data workers have any decision-making power at all in conducting data work—collecting or annotating data. However, even as data workers conduct data work, traditional workers maintain a tight leash through training and guidelines, and later evaluate data work to ensure it aligns with their expectations.

Traditional worker positionalities drive how identity is structured in computer vision. Data workers, as human beings with their own positional perspectives, are expected to embody and automate the positionalities of their clients, the traditional workers. Given data workers largely come from highly different sociocultural contexts than traditional workers, mechanisms are built into the pipeline for teaching data workers how to see like the traditional workers based in the Global North. Traditional workers ensure that data workers understand and internalize their positionalities through training and guidelines.

Data workers' positionalities are seen as a threat to quality data work; the belief that data workers are homogenous, lower class, less educated, and more likely to be from the Global South and thus less familiar with the ethical beliefs of the Global North, are seen as unreliable and more likely to produce poor quality data. On the other hand, traditional workers imagine their own positional perspectives—as informed by class, education, sociodemographic identity, and geographic location—as values to be negotiated with their fellow workers, and within the constraints of their organization. Data workers are given no processes, or sense of empowerment, to give input on how identity is implemented.

In the rest of the Discussion, I describe how positional power is present in computer vision work. I reference both Bourdieu's theory of habitus and Hill Collins' matrix of domination to describe how positional power manifests between traditional tech workers and data workers. I then conclude this section by describing ways we might reimagine the structure of positional power in computer vision development, to intentionally include the positional perspectives of data workers.

Positional Power in Computer Vision

Positionality plays a crucial role in how identity characteristics are implemented in computer vision. Workers reference the perspectives and values they gain through their life experiences. This is demonstrated clearly in Chapters 7 and 8, as well as. Yet not every type of worker's positional standpoint is treated as equal. In this chapter, I have attended to the power that arises between the positionalities of traditional tech workers and data workers presented in Chapters 7 and 8. Through documenting the positions workers occupy during the dataset development process, patterns of power emerge. Traditional workers have much more freedom in expressing their own positional standpoints during development and pushing back against those whose perspectives they disagree with. On the other hand, data workers have very little power during the entire process. Computer vision development is a field of power where habitus operates to influence the *positional power* of individual workers. Positional power is not necessarily intentionally enacted upon data workers by traditional workers but is latently reproduced through the earned set of norms and rules that implicitly shape how individuals view the world, much like

Bourdieu's habitus. Habitus "produces individual and collective practices - more history - in accordance with the schemes generated by history," it "guarantee[s] the 'correctness' of practices and their constancy over time, more reliably than all formal rules and explicit norms" (Bourdieu, 1990).

Positional power, and the practices that uphold it, have become so deeply ingrained in computer vision development that it is upheld by both traditional workers and data workers. It is presented by traditional workers and data workers alike as entirely natural, expected. Traditional workers openly discuss data worker positionality as something risky that introduces biases into the data. Data workers are seen as introducing cultural biases into the data. On the other hand, while data work is treated as the source of bias, traditional worker positionality is never treated as biased. Largely, traditional workers never questioned how their own positionality might introduce bias into the development process. Even the role traditional workers play in scoping identity is not viewed as biased, from either traditional workers themselves or by data workers, but instead as a necessarily valued tradeoff. The "engineers who don't think that much about identity" are not necessarily biased, but value different things than the traditional workers who do think about identity. This propensity to assume data workers are biased, unreliable, and should not be making key decisions about identity underlies the field of power of computer vision. Edgerton and Roberts explain that "dispositions of the habitus generate practices in fields which in turn can affect those dispositions by (de)valuing them" (Edgerton & Roberts, 2014). In the context of this world, the structure of worker positions, or habitus, in the field leads to the (de)valuing of their positional perspectives.

Traditional workers are awarded with positional power—the power to dominate how identity concepts should be defined and implemented in computer vision artifacts. From the start, traditional workers define what identity should look like in computer vision, have the capability to select and deny clients they disagree with, and choose which data workers to hire and what to pay them. They have the power to decide what is and is not ethical, dependent on their own positional perspectives. Their positional power is mutually constructed and upheld through other forms of power, such as economic capital (high salaries and the overall economic power their companies have to outsource undesirable work), social capital (their opinions, perspectives, and even disagreements are valued when discussing how to implement identity in computer vision), cultural capital (in that traditional workers often speak a similar language in discussing computer vision, a language that was largely inaccessible to data workers, who, even working on AI, never came into contact with actual AI systems) and symbolic capital (in that they are viewed as prestigious, as elites).

Even the way that prior work engages with identity in computer vision reinforces the underlying power structures currently present in computer vision. After all, all interventions, from designing better models to mitigating biases in data, are aimed at "designers"—often a broad description that encompasses traditional workers as a class. All of the implications sections found in prior literature on fairness and ethics presume that traditional tech workers are, and should be, the people in the positions to improve machine learning.

Data workers, on the other hand, have no positional power. They do not have the agency to define identity or to screen clients, even when they have opinions about both. Data workers express their positional standpoints only when conducting data work, like collecting data and annotating it. The habitus that data workers embody is one which is incongruent with the values of computer vision as a field of power. As demonstrated in my prior analyses of values in the computer vision and AI space (see Chapter 5), computer vision as a field of power largely prioritizes the habitus of workers in the Global North with specific technical backgrounds. As a result of their undesirable habitus, data workers are awarded no economic capital (in that they are underpaid and experience economic precarity) and little social capital (they are neither encouraged to develop opinions about or able to communicate their perspectives). They also have no cultural capital, in that they are positioned as outsiders to development, treated mainly as automated labor, they have not developed a shared language around computer vision or AI, and have little understanding of its actual use. Further, the educational status many data workers have gives them no power or prestige—given their home countries or status as refugees, their degrees are meaningless to change their situations. All of this lends to data workers having no symbolic capital, in that their role is not viewed as essential and is made invisible in the process of development.

Even as the disempowered group, data workers demonstrated internalized beliefs about their own positionalities as data workers. They expressed similar opinions about data workers introducing biases, while clients "know better." Data workers also felt responsibility for poor quality data work fell on data work shoulders, despite being underpaid and largely invisible. Even when they disagreed with client perspectives on identity, they never spoke up directly to clients about their opinions. Often, the mechanisms by which datasets are developed do not even allow them to speak up. Many data workers simply submitted data to their supervisors once done but did not interface directly with them. As demonstrated by the available literature on contingent online work, data work is often a highly invisible form of labor. In comparison to traditional tech work, the role of data workers in creating computer vision products is portrayed as automated, or "magical" (L. Irani, 2016). The labor of data workers is often the most crucial aspect of machine learning, but the credit for such labor is usually assigned to researchers and engineers. Beyond the visibility of the labor itself, the visibility of worker positionality is different depending on the type of work.

In cases where data workers are able to express their positional perspectives, when conducting data work, these expressions are seen as highly undesirable, and attempts to control them occur both before and after data work. Before data work occurs, traditional workers, in their capacity as clients, create data guidelines focused on ensuring specific data outcomes. Data workers often go through training so that they internalize the worldviews present in the guidelines as they conduct data work. In reality, these guidelines are often vague and fail to account for positionality; they do not attend to differential perspectives on identity concepts. While workers then resource their own positional perspectives to make decisions about data—such as referencing their cultural familiarity with clothing items—clients use their power to veto decisions they disagree with. During the evaluation stage, clients who disagree with data worker interpretations either send the data back to be corrected or simply override the annotations they received. Data workers are expected to be objective, neutral, and apply a "view from nowhere" to data (Haraway, 1988). Any trace of their own positional outlooks should be absent from data. Meanwhile, traditional workers openly discuss and negotiate different approaches to identity in their work. Though they may not explicitly discuss their positionality and its influence, it was always apparent that traditional workers are operating from their own specific habitus.

As a result of deeply ingrained dispositions about the roles of traditional workers and data workers, all participants expressed the attitude that traditional workers are qualified and educated enough to make identity decisions, and data workers are biased by their non-Western cultures. Therefore, traditional workers hold excess positional power over how identity is defined in computer vision. More specifically, the positionalities of traditional workers are more valued than those of data workers. Traditional workers have a *positional power* that data workers do not. Positional power is broader than the titles traditional workers hold. It extends to the value their own positional perspectives are given in developing identity in computer vision.

In the next section, I describe how the social identities inherent in worker positions also influence the level of positional power they are awarded in the field of computer vision. The paradigm proposed in Collins' matrix of domination can be used to examine both how data workers are disempowered, but how traditional workers hold different levels of positional power as well.

A Matrix of Positional Domination

As a result of the positional power traditional workers hold, data workers are actually treated as extensions of traditional workers. They are expected to embody the positionality of traditional workers when conducting their work. They are habituated to traditional tech worker perspectives through their training, the guidelines they are given to conduct data work, and through the evaluation methods used to strip data of any trace of data worker perspectives. In examining how data workers are expected to embody traditional workers, social hierarchies emerge in ways that extend beyond the perspective of positional power I just outlined. In particular, positional power between workers' different social identities becomes salient as certain workers are marginalized or made invisible through computer vision development. First, I attend to how the social identities held by data workers are erased and exploited by traditional tech workers seeking "unbiased" data work. Then I attend to how traditional workers also hold differing levels of positional power depending on their own intersections of social identities.

Traditional tech workers are primarily based in the Global North; participants were largely based in the United States and Western Europe (see **Table 10**). If we revisit **Figure 7**, we can see that the majority of the clients that data workers have served come from countries in the Global North, as well. In contrast, the majority of data workers were based in the Global South. Those based in conflict-affected regions are particularly disempowered, often with no other opportunities for work outside of contingent data work. As Dinorah stated, clients expected work to be conducted in English, making many projects all the more inaccessible for many data workers. Such conditions acutely impacted women participants, who often live in regions with far stricter laws about women's movement, affecting their abilities to get a job and an education. In Chapter One, women participants even discussed how collecting data put them at differential risk of harm than men. Meanwhile, their clients failed to account for the lived realities of data workers in the Global South. Their expectations for how high-quality data work was conducted aligned with their own experiences of work in the Global North. They did not account for differences in collection processes and safety, nor did they consider factors like internet access, ability to leave the home, family life, or working conditions. Beyond the differences in conditions, data workers were also expected to easily and naturally interpret data through the lens of the Global North. Seeing swimwear as racy or explicit would be considered a "cultural bias" for clients in the Global North.

The difficulties and realities of data workers outlined above showcase more than simply the different conditions underlying positional power in computer vision. Clients failing to account for the positions data workers inhabit can be interpreted as a direct extension of the desire for data workers to reflect and embody traditional worker positionalities. The field of computer vision employs what Wendy Chun calls "user amplification," taking one's own subject position and attempting to amplify it through technology. Crucial to user amplification is also the hierarchical component of erasure, making the complexity underlying computer systems invisible. Chun writes, "Such erasure is key to the professionalization of programming—a compensatory mastery built on hiding the machine." Traditional workers did not discuss how their own culture influenced their views on identity but seemed acutely attuned to how the culture of data workers, occupying positions in the Global South, influenced data. As such, they sought to make the positions of data workers invisible. Implicit in the field of computer vision work is the presumption that, if the traditional worker could clone themselves, they could do the job better than the data worker, since they already embody the positionality needed for the ideal outcome. In doing so, they make invisible the subject positions that data workers inhabit—including their identities as economically disadvantaged, as women, as refugees, as located in the Global South. In erasing the positionalities that data workers inhabit, traditional workers fail to account for the intersecting forms of oppression underlying the reliance on data work. For example, how gender and religious beliefs intersect to make annotating certain types of content harmful, or how gender and space intersect to make collecting data more dangerous or risky. In an attempt to produce universal representations of the world, rather than contextual ones (see Chapter 5), traditional tech workers implicitly presume a universal positionality.

Further, the current structures of domination underlying computer vision practices uphold interlocking systems of oppression. As other scholars have pointed out, data work relies on the economic exploitation of individuals largely located in the Global South. The notion that traditional workers are valuable stakeholders while data workers provide risky and mechanistic labor mutually reinforces the economic conditions underlying computer vision. Clients, in high paying and prestigious positions in wealthy economic countries, take advantage of data workers disempowered positions. Labor arbitrage pits workers from different countries against one another, driving prices down across the globe and further limiting diverse perspectives. Poor reviews fail to account for the economic conditions which might have led data workers not to finish tasks. At its worst, some clients even scammed data worker participants out of their earned pay, trying to build cultural rapport with data workers solely to exploit them. Beyond the individual impacts of exploitative economic practices, computer vision as a field of power relies on the economic exploitation of data workers. Many computer vision projects—even in their desire to improve the ethics of computer vision, through practices like developing highly diverse datasets of human faces—are only viable due to the economic configurations underlying computer vision. After all, EnVision Data could not undertake Xavient's project because it was too labor intensive and too expensive for their client.

Of course, power is not a clearly delineated top-down structure from traditional workers to data workers. Just like with data workers, some traditional workers have more positional power than others, as well. As demonstrated in Chapter Two, women and people of color often relied on their white male colleagues for leverage. As insinuated by Elliot, engineers often had the most power in defining identity in early project stages. Engineers having more positional power to define identity led to less nuanced and more discrete categories of identity that other traditional workers disagreed with or viewed as harmful. Only through collective action did the transgender and non-binary workers at Lynn's company, MultiplAI, manage to push back on decisions made by engineers to frame gender as sex. Identity classifications play a large role in determining the positional power available to different individual workers.

The field of computer vision is further complicated by the social positions each worker holds, such as race, gender, and class. To shift positional power in computer vision to be more equitable between workers, it is also necessary to attend to how positional power is awarded or denied given the matrix of social identities present in the field. In the next section, I describe potential approaches to reimagining positional power in computer vision. Not only do I encourage practitioners to think about shifting power *between* types of workers, but they should also consider the role of identity-based oppression *within* types of workers.

Reimagining Positional Power in the Field of Computer Vision

Teasing apart the way that habitus reinforces specific power hierarchies in computer vision—with traditional workers on top and data workers on the bottom—can also aid us in reimagining alternative structures of power. In this section, I engage with an imagined world where the underlying habitus of computer vision prioritized the positional perspectives of data workers across the development pipeline, rather than the positional perspectives of traditional workers. Further, as I have just discussed, those with the most marginalized identities, even among traditional workers, are currently disempowered. Given standpoint theory also posits that some people are "closer" to certain knowledges than others, I consider the tactical advantages of making the most marginalized identities dominant in the development of computer vision.

Shifting Defining Power

In the current structure of the field of computer vision, the power to define identity is largely in the hands of the traditional worker—and their customers, for whom they are serving. Data workers are entirely absent from this process and are simply handed identity categories to collect for and annotate during the data work stage. By the time identity categories get to data workers, data workers are expected to learn and embody the positionalities of traditional workers, to ensure data aligns with identity categories through traditional worker eyes. But what would it mean for data workers to have defining power?

One method of shifting defining power to include data workers would be to consider leveraging data workers' culturally situated positionalities in defining data identity categories. Rather than relying on traditional tech workers to attempt to create universal data standards—that largely reflect the views of the Global North anyway—data workers could develop culturally contingent definitions of identity. One could imagine how data workers in Kenya would develop very different models for identities like gender than the traditional workers based in the United States. After all, someone like Gemma might intentionally account for transgender identities in ways that are not currently attended to by traditional tech workers. Similarly, racial or ethnic categories might differ significantly if defined by data workers situated in, for example, India, where historically racial categories differ significantly than in the U.S. (Morning, 2008). Other categories of identity might also be more salient in other contexts which otherwise do not even show up in computer vision in the Global North.

Defining identity from a culturally contingent perspective, rather than an idealized universal one, might lead to more contextual modeling practices. Perhaps building computer vision models based on the positions of data workers, rather than the desires of customers, would lead to more accurate classifications in specific locales. Of course, there are certainly barriers in current approaches to computer vision to creating more culturally situated artifacts. Universalist models are largely the approach to computer vision because the positions which many traditional workers occupy, particularly in engineering, prioritize abstraction for the sake of universality. Further, it is cheaper and more efficient for the companies building them. However, it is worth questioning why traditional workers with no positional knowledge of the regions that they are attempting to gather data from or release products in should be the ones empowered to define identity categories. While identity categories always reflect the subjective positions of specific cultures and locales, including in the current status quo of computer vision development, harnessing the positional perspectives of data workers in defining categories can reveal new lenses through which to view identity, and new considerations for which to assess concepts like bias. Leveraging data worker positionalities more in defining data categories could lead to new insights, richer data labels, and more imaginative approaches than always relying on traditional tech workers in the Global North.

Shifting Selecting Power

Current selection practices give traditional workers the power to both select data workers to outsource to—within legal limitations—and to select the clients they develop computer vision for. Even when traditional workers select data workers, the data workers themselves have little leverage. Not only are they not able to determine which projects they want to work on or which clients they want to work with, the economic precarity of data work pushes them to take on jobs simply out of necessity. They have little bargaining power when it comes to choosing projects or negotiating payments. Here, I imagine what it would look like for data workers to be given selecting power.

Selecting power was the ability to select clients. One might imagine what it might mean for traditional workers and data workers to have mutual selection power, much like traditional workers have with their customers. For data workers, selecting power was heavily tied to economic precarity. If data workers had more economic capital, they would be able to set stricter boundaries about what types of work they would engage in. After all, numerous data workers expressed finding projects they were working on to be “ridiculous” or even unethical. Yet they would often still work on those projects because not doing so meant they would lose out on necessary income, a reality which was not present for traditional workers when assessing their clients. Traditional workers only had to consider whether the company they worked for would profit, but not whether they themselves would suffer economic losses.

While many have proposed paying data workers more given their centrality to the development of AI (e.g., M. Díaz et al., 2022), they are also experts in how AI systems might impact people. They have positional perspectives which allow them to assess the ethics of a project in ways traditional workers might not. After all, Maakarim clearly saw issues with how a hiring-based computing vision model might discriminate against people whose first language is not English. Not only could being able to turn down projects data workers find harmful benefit them individually—in that they would not be exposed to the harm—it might also allow them to communicate to clients what aspects of the projects are problematic. While realistically, many clients might not change anything and still attempt to find new workers for their project, data worker input might also help to reshape and improve certain projects to be less unethical.

Shifting Designing Power

The only place where data workers currently have any power in identity development for computer vision is in conducting data work. They are able to access tacit knowledge from their own positional worldviews as they collect and annotate data (see Chapter One). However, this power is heavily constrained. Traditional workers use their own power to ensure their positional perspectives are privileged. They train data workers to see data instances in the same way they do (e.g., two women as brides) and/or reinforce their worldviews through guidelines (e.g., breasts/nipples are to be considered explicit content). Data workers, largely based in entirely different cultures from traditional workers, are asked to put their own worldviews aside and embody the same positionalities as their clients, the traditional workers. In this section, I imagine what designing power might look like if data workers were allowed to embody their own positionalities in designing data practices, rather than those of their clients.

After all, data workers have expertise in conducting data work, while traditional workers do not. One benefit of shifting designing power would be data workers being able to design data practices could lead to more accessible and detailed guidelines. Often, the way that guidelines are designed only implicitly reflects traditional worker positionalities. The vagueness of expectations in guidelines means that data workers are already applying their own interpretations to the data. If data workers were involved in designing guidelines, they could discuss how they do or do not meet project requirements directly with their clients before conducting data work. It would allow data workers to construct mutual understanding with their clients, rather than promoting a one-way flow of information from client to data worker. Involving data workers in designing guidelines and conducting training would give data workers opportunities to negotiate and clarify with their guidelines before collection, avoiding culturally contingent confusions about what a blouse looks like in certain countries.

Ideally, involving data workers in the design of training and guidelines would also allow clients to learn more about the contexts that data workers are situated in, and prompt reconsiderations for how identity is being defined in guidelines before collection or annotation occurs. While clients might still seek to create computer vision models that classify nipples as obscene, they might consider how their own views of obscenity are biased by their positions.

Shifting Evaluating Power

The final stage in defining identity in computer vision is broadly the evaluation stage. Traditional workers evaluate the data given to them by data workers. The status quo of this stage means that traditional workers can determine whether data is biased, low quality, or does not meet their expectations. They then have the ability to veto data worker decisions, either sending the data back to be recollected or reannotated, or simply overriding the decision themselves (e.g., relabeling Latin American selfie videos as white). In this final section, I present possibilities for data workers to evaluate their own data.

Rather than deploying authoritarian quality assurance checks from traditional workers, who can simply override data worker perspectives they deem to be incorrect, one can imagine more democratic modes of assessing the quality of data work. Instead of relying on majority rules models of evaluation, data workers might attempt to reach consensus about data instances. For example, data workers might be encouraged to discuss the application of identity categories amongst themselves. Such discussions would allow data workers to form deeper understandings about the data, to see it from new perspectives, and ideally lead to higher quality datasets as workers collaboratively apply data labels. Involving data workers in evaluation procedures would also necessitate involving them in scoping the project, so that they understand how the data would go on to be used—an aspect of the development pipeline they are currently not involved in.

Discussing agreements and disagreements might prompt reconsiderations about the categories altogether. In the status quo approach that privileges the positional power of traditional workers, certain identity categories might be a misrepresentation of lived experiences. For example, relabeling Latin American people as Caucasian can be considered erasing the actual identities of the people in the data and replacing them with Western interpretations. If data workers were able to communicate disagreements about identity categories to their clients more effectively, identity in the data might also be more democratically reshaped before product deployment.

Conclusion

Positionality is crucial to implementing identity concepts for computer vision. Both traditional tech workers—like engineers and researchers—and data workers—like data collectors and annotators—define and interpret identity concepts differently given their own worldviews. Through interviews and observations, I examined the relationship between traditional tech workers and data workers during the development of computer vision.

I found that these two types of workers are given very different power to engage their positionalities. Traditional workers are given opportunities to enact their own positionalities and negotiate disagreements with their colleagues. The only place data workers are able to reference their positional viewpoints is during the conducting of data work. Yet, the development of computer vision is set up so that data worker positionalities are carefully curated and controlled. Traditional workers view data worker positionalities as threatening to the quality of the data and are particularly concerned that they will introduce biases into the data and subsequent models.

I discuss how practices in computer vision reflect a specific kind of power, positional power. Positional power refers to the agency workers have to engage their own perspectives and experiences when implementing identity for computer vision artifacts. In discussing how positional power manifests in computer vision, I draw inspiration from Bourdieu's theory of habitus and Collins' matrix of domination. I conclude with potential ways of reimagining positional power in computer vision to better enable data workers, as well as to account for different identity-based oppressions which are marginalized in computer vision development.

10
HOW IDENTITY MOVES FROM OPEN TO CLOSED

In this dissertation, I have explored how identity is constructed in computer vision. First, I presented a set of work focused on analyzing how identity is currently represented in computer vision artifacts—both models and datasets. Computer vision artifacts are the finalized output of the development pipeline. This work showcased that identity in computer vision artifacts is rigid, calcified, and closed. I showed that race and gender in datasets are portrayed as static, as well as neutral and apolitical, ignoring the social realities of race and gender. I also showed that the binary gender in computer vision models erases and marginalized transgender and non-binary people. How race and gender are historically embedded into these artifacts leads to a further calcification of marginalization of certain identity groups, like people of color and non-binary people. As such artifacts are deployed, they actively harm those groups in different ways, whether through stereotyping or erasure.

Further, I showcased how computer vision artifacts communicate implicitly the underlying values of their creators. Solely through analyzing datasets, models, and documentation, we can see how computer vision practitioners value closed models of identity in computer vision. Dataset creators valued efficiency, universality, impartiality, and model work, reflecting the careless and underexplained approaches to identity showcased in Chapters 3 and 4. Thus, a lack of human-centeredness seemed to be entrenched in the practices of dataset creators and the field of computer vision more broadly. Yet, there were still open questions about how practitioners actually approach embedding their values in computer vision artifacts.

Then, I presented a set of work focused on how human workers implement identity for computer vision artifacts. I revealed how both traditional tech workers and data workers embed their own positionalities in how they implement identity concepts in computer vision. I showed the different contexts traditional tech worker positionalities are enabled and constrained by, as well as how they negotiate their perspectives with their colleagues. I also showed how data workers reference their own positionalities when collecting and annotating data. For both traditional and data workers, I revealed how unexpected and undesirable outcomes of identity end up in computer vision systems due to positional gaps among workers. Positional gaps help to explain why workers, like the dataset creators of Chapter 5, approach identity in artifacts the way that they do. They also suggest a need to attend directly to the positions that different workers occupy during the development process, to mitigate and prevent the harm caused by positional gaps.

In comparing how both traditional and data workers can access their positionalities to conduct identity work in computer vision, I also revealed ingrained power differentials in how computer vision is developed. I showed that traditional tech worker positionalities are privileged and empowered over data worker positionalities. Traditional tech workers largely drive identity towards a specific vision, using data workers to enact that vision. Data workers are expected to put aside their own positionalities so that they can embody the positionalities of traditional tech workers. While traditional tech workers largely drive how identity is implemented in computer vision, data workers are seen as the source of undesirable characteristics, like bias, that end up in finalized artifacts. However, many data workers might have contextual and localized knowledge that would benefit the development of identity concepts that traditional workers cannot even foresee. This work suggests a promising approach to identity in computer vision that centers and empowers data workers as knowledgeable resources during development.

The above work shows that, while computer vision artifacts represent a rigid and closed model of identity, workers still ingrain their own positionalities within those artifacts. The process of defining identity for computer vision is inherently subjective, and workers—even when constrained by the power embedded in development practices—always access their implicit positional perspectives to make decisions about identity categories. Thus, somehow, through the highly social process of producing identity for computer vision artifacts, concepts of identity become increasingly rigid. Identity moves from something that is negotiated among workers with different experiences and perspectives, to something that is calcified in datasets and models that enact very specific classifications of concepts like race and gender.

In this concluding chapter, I synthesize this work on both artifacts and work practices to argue that identity in computer vision is a result of a very specific narrow model of identity—one which moves from something open to something closed. I present a theoretical framework for how identity moves from an open and intangible concept to something that reflects the positional perspectives of various workers during development to, finally, a concrete and closed attribute to be used in a technical system.

In the next section, I detail how identity is transformed in the development of computer vision. I describe five steps in developing computer vision which transform identity. I will elucidate the transformation processes commonly employed when developing computer vision artifacts. I apply the open-to-closed model of identity to the development of computer vision by mapping moments of transformation—how "Identity" becomes an Attribute, how an Attribute is subtly changed throughout the development of the product, and how the Attribute becomes solidified into a Technical Attribute.

I then conclude this chapter by describing the open-to-closed model of identity in technology development. I outline the three phases and what each means: "Identity," Attribute, and Technical Attribute. First, I discuss "Identity" as a nebulous and intangible meta concept which exists in the world, prior to any forms of technology development. "Identity" as a concept exists regardless of human intervention, though many different scholars have attempted to theorize about what "Identity" means and where it comes from. I then discuss how this meta concept of "Identity" serves as a resource from which different human actors attempt to pull from to help concretize identity into something tangible. In the case of technology development, workers access their own positional perspectives about "Identity" so that they can ground decisions in those perspectives. They thus create Attributes, malleable categories of identity that they work with during the process of technology development. The goal of this Attribute is to slowly turn it into something fixed and unmalleable, the Technical Attribute. The Technical Attribute represents the narrowest closed perspective of identity, which has been implemented within a technology. The technology can then enact this view of identity on the world around it, ascribing a singular and highly specific view of identity concepts.

Through my theory of open-to-closed identity in technology development, I reveal how something as complex and messy as "identity" as a concept is made into something technical. Understanding how identity is transformed for the purposes of technology development gives us a new approach for interrogating the rigid categories often presented as neutral and objective in artifacts like datasets. While often assessments of technology occur at the artifact level, when identity has already been solidified into a Technical Attribute, we can use an open-to-closed model to reverse engineer Technical Attributes back to their “Identity” roots.

Five Step Transformation Process in Computer Vision

Let me begin by reiterating in more detail the development of “human group” categories in the SensEyes project (as briefly described in Chapters 8 and 9). In 2020, EnVision Data was contacted by a small technology company, SensEyes. SensEyes provides identity verification for mobile applications, and they desired a new dataset of 5000 selfie videos to train a model for identity verification and spoof detection. The goal behind this new dataset was to ensure their models would work equally well across different sub-groups; to ensure fairness. At this stage, SensEyes had already taken a fuzzy and amorphous concept of identity and constrained it to something more technical—defining human subgroups to measure classification parity.

Thus, SensEyes came to EnVision Data with a list of pre-established data requirements. SensEyes wanted the new dataset to be diverse across various human attributes, so that their product worked more robustly on different faces. In envisioning what “race” should be, SensEyes’ research and development team had already established the attribute “human groups.” They came to EnVision Data with a list of “human group” categories; they wanted an equal distribution across the five categories “Caucasian,” “Asian,” “African,” “Latin American,” and “Middle Eastern.” As demonstrated by Chapter 9, these human groups were laden with the positional perspectives of workers at SensEyes. They were colored by their own perceptions about racial categories.

EnVision Data took on the project. Due to the scale of the project, EnVision Data decided to hire freelance data workers from Upwork from around the world. EnVision Data had mapped regions of the world to the categories that they were given. Data workers were then hired based on their region, such as Southeast Asia. Each data worker was asked to collect videos from their contacts or those around them.

Data workers were not asked to target specific groups themselves. The hope in hiring freelancers based on their location was that their regions would map to the expected "human group" categories. Therefore, the data collected from data workers located in the Philippines would be assigned the category of "Asian," while the data from workers in Guatemala would be assigned the category "Latin American." Here, EnVision Data had to transform the five "human group" categories into something that could be actioned on. Thus, they layered their own interpretation of "human groups" onto each category, associating categories like "Latin American" with specific countries they assumed "Latin American" would be located. But both EnVision Data and SensEyes discovered that their definitions for the attribute of "human groups" did not align. EnVision Data took an approach to collecting data from the "human groups" categories through regional distribution. They assumed that the data collected from a certain region would map to the expected categories, even if imperfectly.

Yet during the quality assurance check that SensEyes researchers performed on the data, SensEyes reclassified many of the videos. They reclassified "Latin Americans" who they felt looked too white as "Caucasian" and those who looked too Black as "African." As their goal of their product was not to classify people into the five categories of "human groups," SensEyes did not care about the accuracy (in terms of geolocation) of the category that the person in the video was assigned to. They instead focused on getting an equal distribution of the *appearance* of people across the globe. The reclassifications were each done by the researcher and development team at SensEyes, using intuition about the appearance of the data subjects in each video. Once again, identity in the nebulous concept of "human groups" was transformed via the interpretation of workers. As it was transformed, it became more solidified, into something that was deemed usable for a specific task—fairness measurement.

EnVision Data adopted a location-based approach that reflected concepts of ethnicity, while SensEyes adopted a phenotypic approach that was more aligned with concepts of race. "Human groups" acted as a stand in for a conflated notion of race and ethnicity that EnVision Data and SensEyes had not coalesced on, despite the shared goals of creating a fairly distributed dataset. The categories themselves, which were meant to provide a more concrete shared understanding about what a "human group" was, instead acted as a mechanism for revealing different mental models for identities like "Latin American." In the end, the implicit mental models that each entity had of "human groups" and its associated categories led to different data outcomes. However, the client, SensEyes, had the power to redefine and solidify the categories, overriding EnVision Data and their data workers' perceptions. As the concept of "human groups" became more solidified, it also reflected narrower positional worldviews. In the process of identity development for computer vision, some actors always have more power over the process than others.

From the outset, the process of implementing human-centered computer vision artifacts may seem straightforward and mechanical: define relevant categories, explicate them in data requirements, collect and annotate data based on those requirements, train model. However, in reality, there are dozens of decisions that are being made throughout the process of implementation, involving a number of human actors. Many of the decisions about human identity are implicit; they are never explicitly discussed or concretely defined by the actors involved in development. Given that each person involved in the process carries their own understanding of human attributes, they often assume a shared mental model of what that attribute is. EnVision Data had assumed that data collected from "Latin American" data collectors would reflect the expected category of "Latin American." However, SensEyes was hoping for an otherwise undescribed and unspecified phenotype: a certain skin color and a certain appearance. Often, such disagreements can be so implicit in the data that different actors never realize they are perceiving identity in different ways. However, in this case, SensEyes was able to perceive those differences because the skin tone of subjects in the data did not match their expectations. Because they were able to perceive a differential interpretation of "human groups," they were also able to easily override it.

Yet, regardless of whether different perceptions of identity are discussed explicitly or revealed implicitly, each decision transforms the concept of human identity—what it is and what it looks like—in the data and the models trained on the data. This variation is evident in how SensEyes reclassified "human groups" that EnVision Data believed fit the criteria, an explicit recalibrating of what "Latin American" was made to look like in the data. There were also more implicit transformations of the attribute of "human groups," such as who the data collectors had access to when collecting the data, how they determined their attributes visually, and how the data subjects expressed themselves when the data was collected. Every step in solidifying the broad concept of "identity" into something technical for a classification task means "identity" becomes narrower and narrower.

I have demonstrated in my dissertation that human actors—in their capacity as traditional tech workers or data workers—reference their own positionalities to define identity in computer vision. I have also demonstrated that identity in computer vision artifacts is rigid, discrete, and calcified—it operates under a "closed" model of identity, where identity categories are defined by very specific data instances. Synthesizing insights from both the development processes of traditional work and data work and how identity is communicated through artifacts like datasets and models paints a much fuller and richer picture of how identity is constructed in computer vision. More specifically, I synthesize the work presented in this dissertation to ground a new theory about how identity in technology development occurs.

I propose a theory of how identity moves from "open"—laden with opportunity, nuance, and interpretation—to something that is "closed" through the processes of development. I argue that human identity in technology development is implemented through a three phase transformation process, from (1) "Identity" as an open, complex, and intangible concept, to (2) an Attribute where workers attempt to coalesce around different positional perspectives of "Identity," to (3) a finalized Technical Attribute that can enact its own closed model of "Identity" in the world. The open-to-closed process can be seen in **Figure 24**.

To demonstrate this open-to-closed transformation process, I detail how identity is transformed in the context of computer vision. I outline a five-step development process where the transformation of "Identity" to Technical Attribute occurs. I map the transformations between each stage of the development process as "moments of transformation" where workers actively transform identity categories through their social and technical decisions. Understanding the transformation process by which identity moves from something "open" to something "closed" reveals intervention points throughout the development process for revisiting and reinterpreting how identity is negotiated and defined. It also helps us operationalize the implicit characteristics of positional power present during the process of transforming identity. Thus, we can assess and take action on dynamics of positional power at each stage of identity transformation.

In this section, I describe how different actors enact their own positional power over the process of transforming "Identity" into a Technical Attribute for computer vision models. To demonstrate how identity moves from open to closed, I go in depth on how the three phases outlined above actually occur in the context of my data on computer vision. I will describe in detail how "Identity" becomes a Technical

Attribute during the process of developing a computer vision artifact. I outline five different steps of transformation that occur during the development process of computer vision, mapping them to specific practices that workers engage in to solidify a Technical Attribute. I detail the practices of workers through process maps, which highlight different key decisions that workers can make when transforming "Identity."

Step One: Transforming "Identity" Into an Attribute

The first step to creating an artifact is taking "Identity" and transforming it into an Attribute. "Identity" is not tangible, but conceptual. It exists as a vast and nebulous concept with the capacity for multiple intersecting or diverging interpretations. Different people have different mental models about "Identity." In fact, most people have many mental models of "Identity."

An Attribute is a feature derived from concepts about human identity. Given that the way human identity is imagined, experienced, and theorized as vast and contradictory, there are many types of Attributes and use cases for them. Taking concepts from "Identity" and attending to them as an Attribute, or a feature, is the first step in creating a human-centric computer vision product. The product may be an entirely new product, which has not been built yet, but it is also often an existing product which is being updated. When individuals work to define an Attribute so that a feature of a product can be implemented, they access their own positionally informed mental models associated with "Identity" concepts relevant to that Attribute.

Certain Attributes seem clearly derived from mental models of "Identity"—such as race, ethnicity, or gender. In order to build a product like a gender classification model, workers must determine that they desire the Attribute of "gender," and develop an understanding of what "gender" is. Nevertheless, that understanding or defining of "gender" can vary. While gender is often implemented as a binary (male versus female; man versus woman; masculine versus feminine), even the conceptualization of each binary reflects slightly different mental models. While "male" insinuates a sex-driven model of gender identity, man insinuates a social model, and masculine implies a self-presentation model. Of the companies represented in my study, many implemented "gender" in their artifacts, but often differently. The company, MultiplAI, implemented a "gendered appearance" Attribute, while others, like Exodia, used "gender" which was formed by mental models of sex identity.

Mental Model of "Identity"	Product Feature	Attribute	Categories
"Gender identity"	Gender classification	Gender	Male; female
"Racial identity"	Bias assessment	Race	White; Black or African American; American Indian or Alaska Native; Asian; Native Hawaiian or Other Pacific Islander
"Personal identity"	Facial recognition	Individual Identity	Is X individual; is not X individual
"Emotion"	Affective classification	Emotion	Happy; Sad; Angry; Neutral
"Personhood"	Human face detection	Face	Is person; is not person

Table 14. A table with examples of different potential models of the concept of "Identity" and how they might be reimagined as features. Features lead workers towards concrete definitions in which to ground Attributes for a product.

While traditional workers often associated "Identity" with "demographics", other Attributes are less obviously derived from "Identity," because they aren't necessarily envisioned as types of identities (in the same way "gender identity" or "racial identity" is theorized as types of identities). Yet Attributes like "emotion," "personhood," and "personal identity" are also tied to models of "Identity." Emotion is often used as an Attribute for affective or expression classifications and attempts to map human expressions to human feelings. Personhood is informed by models of identity that differentiate human beings from other beings and objects. Finally, personal identity is tied to theories about identity that assume each human is unique and can be identifiable by unique characteristics.

Table 14 shows some potential approaches to defining Attributes. Often, a product team is tasked with implementing some "Identity" concept in a product, whether that product is new or existing. From my data, I gleaned four different decision points which might occur during this process. A product team might decide to discuss how to define the Attribute amongst themselves, negotiating their differing positional perspectives within the confines of the product team. They might also decide to consult research, bringing in different perspectives from traditional workers outside of their direct team, who might have very different perspectives on identity. Researchers also have their own decisions to make, such as consulting existing literature on the “Identity” concept at hand or conducting empirical research to understand best practices around implementation. Whether the product team takes the research teams’ perspectives into account is dependent upon existing practices within the company and how decision-making power operates between different traditional workers.

Traditional Workers Distill “Identity” into Attribute

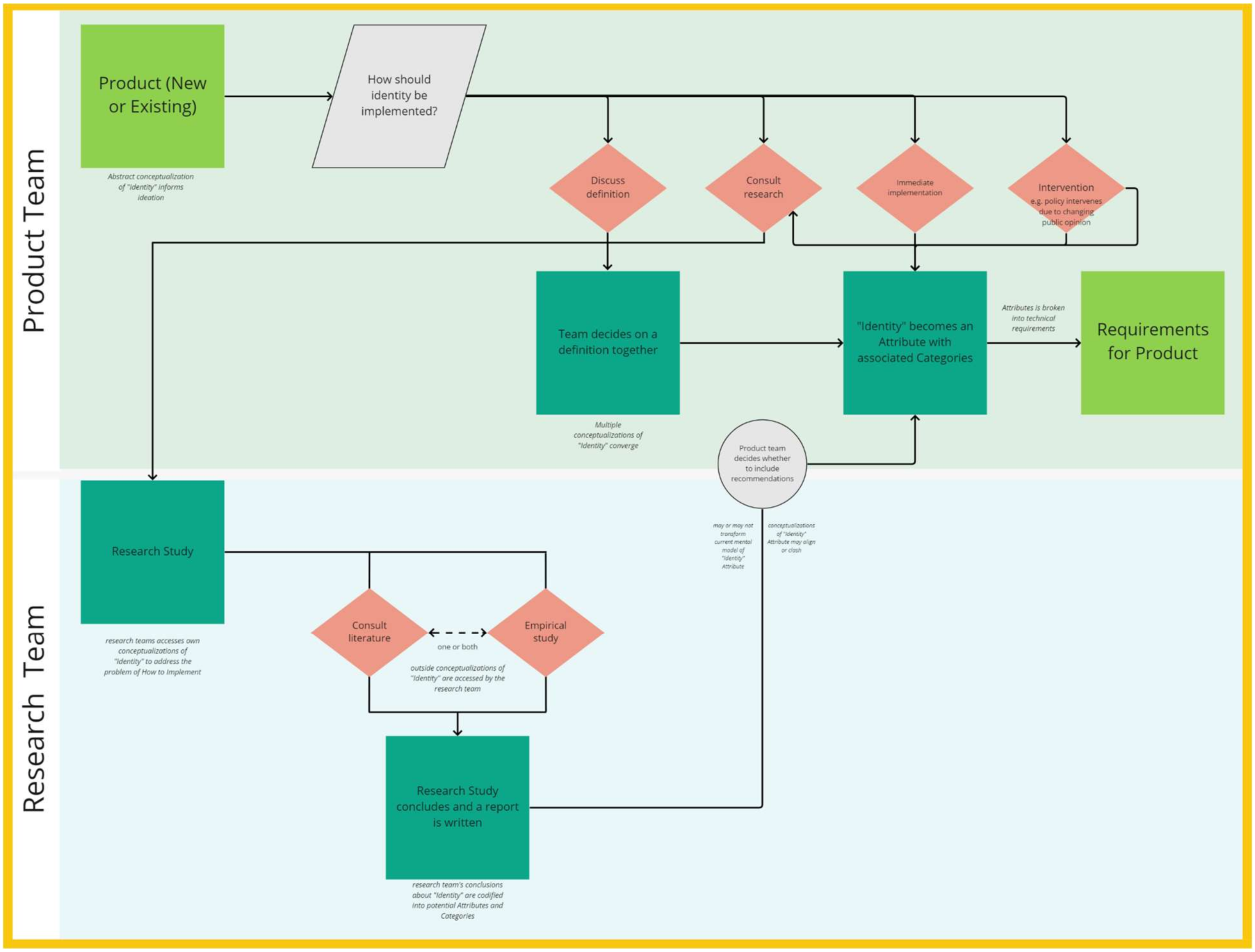

Product Team
Product (New or Existing)
Abstract conceptualization of "Identity" informs ideation
How should identity be implemented?
Discuss definition
Consult research
Immediate implementation
Intervention e.g. policy intervenes due to changing public opinion
Team decides on a definition together
Multiple conceptualizations of "Identity" converge
"Identity" becomes an Attribute with associated Categories
Attributes is broken into technical requirements
Requirements for Product
Product team decides whether to include recommendations
may or may not transform current mental model of "Identity" Attribute
conceptualizations of "Identity" Attribute may align or clash
Research Team
Research Study
research teams accesses own conceptualizations of "Identity" to address the problem of How to Implement
Consult literature
one or both
Empirical study
outside conceptualizations of "Identity" are accessed by the research team
Research Study concludes and a report is written
research team's conclusions about "Identity" are codified into potential Attributes and Categories

Figure 19. A process map showing approaches that traditional tech workers take when considering how to turn “Identity” into an Attribute. This initial ideation process may involve a product team and a research team, though research is not always involved. Each pink diamond represents a potential decision point that can be made, leading down different potential paths. In the end, “Identity” is stilled into an Attribute in the form of requirements.

More often than not, it seemed that product teams simply decided on immediate implementation of an Attribute, implicitly assuming that everyone had an agreed upon definition of how “Identity” should be concretized. According to traditional tech workers in Chapter 7, it was not common practice for traditional workers in technical roles to think deeply about how to implement Attributes. In some cases, the practice of simply implementing without thinking led to direct intervention from other workers in the company, like policy. For example, Kaleigh described in Chapter 7 how policy would step in to ensure product teams complied with updated approaches to “Identity.” When existing mental models about how “Identity” should be properly implemented, it can make it more difficult to settle on the best steps forward—especially when there is a legacy model already employing one group’s mental model of “Identity.” Often, decision making about Attributes are resolved through positional power within institutions (see Chapter 9), rather than through equivalent co-construction.

Regardless of the approach taken, by the end of Step One, the core team of traditional workers working on the product has coalesced around the ideas governing the Attribute. Step One reflects how multiple mental models have contributed to the way the team thinks the Attribute should be implemented. After determining how “Identity” should be represented in the form of an Attribute, workers also have to decide on what Categories should define that Attribute. For example, SensEyes defined the Attribute of “human” groups as the Categories of “Caucasian, Asian, African, Latin American, and Middle Eastern.” **Table 14** shows different potential Attributes that can be derived from different mental models of “Identity.” Each “Identity” generally leads to one Attribute; something like “gender identity” was never represented in the form of multiple Attributes in a single product. Though, Attributes might differ between models at the same company (e.g., facial classification versus image labeling (see Chapter 2)).

Three Points Where “Identity” is Transformed into an Attribute

The point at which “Identity” becomes an Attribute matters because the starting point for which “Identity” is derived might push traditional workers to define it in specific ways. Throughout my data, I saw three different patterns for translating “Identity” into an Attribute: (1) at the start of new product development; (2) without specific product goals; and (3) when changing an existing product.

At the Start of New Product Development

In some cases, Attributes are determined from the start of product development, and driven by the desire for a specific product feature. In the case of building products from scratch, a team generally determines the need for a product—for example, to build a suite of computer vision models clients can then purchase and use for their own unique needs, or it might be to build a model for auto-captioning photos the users of your social media website upload. In such cases, the features of the product often drive the selection of Attributes. “Identity” is generally viewed through the lens of utility, often informed by demographics. Attributes like “age,” “race,” and “gender” are imagined as values attached to broad social behaviors, like shopping choices.

For example, in Chapter 7, Kenny described a very utilitarian and market-driven approach to “Identity” when selecting and defining Attributes. He described how the Attribute for gender was largely driven by client demand. He explained that the vast majority of clients requesting the Attribute gender are doing advertising, associating behaviors with the concepts behind gender as an identity. Thus, they defined the Attribute “gender” with two Categories: “male” and “female,” based on advertiser beliefs and needs. In this case, there is a relatively simple method for selecting the Attribute: business. The simplicity of the approach makes any engagement with “Identity” in defining the Attribute all the more implicit. Underneath the market-driven goals of the client is also a historical set of practices imported from the field of marketing, which has historically used gender as a means of tying identity to products. Even while the stakeholders involved have a mental model of gender identity, there is little explanation or negotiation as to what the Attribute “gender” is beyond a presumably predictive variable.

When “Identity” is transformed for a totally new product, rather than for any existing one, defining the Attribute differs. Mental models about “Identity” are largely colored by client needs or market use cases, driving workers to think about defining the Attribute in specific ways. Given the history of gendered approaches in marketing, refining “gender identity” into an Attribute for a marketing classifier leads workers towards a more binary model of “gender.”

Without Specific Product Goals

Some Attributes are defined without specific product goals in mind. For example, both Kaleigh and Vasudha work at the same large tech company on computer vision products. Kaleigh works as a program manager who oversees fairness efforts and responsible AI approaches, while Vasudha works as a product manager who oversees implementation. The company that Kaleigh and Vasudha work at doesn’t ascribe a top-down hierarchy when it comes to product development like the smaller company Kenny works at. One of the models they worked with came about through open-ended research, rather than specific market-driven goals. As such, product ideas can often come to fruition through a number of avenues. Many larger tech companies take this “innovation comes from anywhere” (Chin Leong, 2013) approach.

While one might expect that those Attributes that are defined without specific product goals or clients driving business needs would involve a deeper engagement with “Identity,” “Identity” is still often implicit in the process of determining and defining Attributes. Instead, “Identity” is often approached as a means for technical achievement, and then later assessed for its utility. Technical research teams often approach classifying identity concepts for the sake of exploring new research areas or improving performance on existing problems (as seen in Chapter 7). How “Identity” is transformed into an Attribute is reflective of the priorities and interests of the teams developing them.

When Changing an Existing Product

Revisiting legacy systems which have implemented identity Attributes is increasingly common. A legacy product might have an embedded Technical Attribute that is being reconceptualized and revisited, so workers are starting from a new model of “Identity.” Often, the workers involved in the original creation of the model—and its associated identity categories—are no longer at the company, and there is often no legacy documentation explaining the choices that were made. Often, the combination of the huge number of employees and teams at large tech companies with this “innovation comes from anywhere” approach also makes it difficult to trace the genealogies of Attributes back to their ideation; often they are undocumented decisions that become lost over time with employee turnover.

As described in Chapter 7, Kaleigh is now overseeing an overhaul of gender in Aqueous’ core computer vision model, which would have downstream effects on all of the products currently employing it. To update “gender,” Kaleigh is returning to Step One of the development process, more aware and more critical of how certain approaches to gender identity influenced the way the Technical Attribute was implemented. In negotiating the changes to be made to the system, Kaleigh’s perspective on implementing gender is informed by a mental model of gender as an “Identity” that is not visual, but internal; therefore, the use of gender as an Attribute meant to indicate something about a person’s “gender identity” is logically incorrect and potentially harmful. She must negotiate her mental model of

“Identity” with those of the product teams she is working with, which largely view gender as something to be classified for specific tasks.

Step Two: Translating Attributes and Categories into Data Guidelines

After Step One is complete and the concept underlying the Attribute is loosely defined, workers must begin to work on actually implementing the Attribute. In Step Two, traditional tech workers turn to data workers to begin gathering and labeling examples of the Attribute—the data that will represent it. In order to ensure their vision for the Attribute remains intact, traditional tech workers design trainings and guidelines for data workers. Idealistically, these trainings and guidelines work to ensure that data workers understand traditional tech worker perspectives about the Attribute. Attributes are further refined into something solidified in Step Two because the ideas governing Attributes from Step One are put into specific requirements for data. By the end of Step Two, descriptions of the Attribute have been formalized in the form of documents like guidelines and trainings.

Legal Constraints to Outsourcing

The way that the Attribute becomes transformed in Step Two is also largely shaped through logistics and constraints about what data can be collected, how it can be collected, and who can collect it. In the context of larger companies, legal teams play a major role in what data can be collected and which data workers can be involved in that collection. Larger companies often have a process of legal review for projects, including what a team might be able to outsource. Legal review encompasses both, literally, the legality of a project, but also if the project complies with company policies. Not all companies have this process, and even large companies may have standardized contracts, where projects that meet the approved standards no longer have to go through a legal approval process. The legal review process might differ widely depending on the company and its policies, as well, and policies are constantly evolving, especially at large companies which face more critical press and many more lawsuits. However, when legal is involved in defining what is acceptable to outsource, they also have the ability to shape what an Attribute might look like and what kind of documentation on the project can be collected.

Legal Approval Process for Accessing Data Vending Services

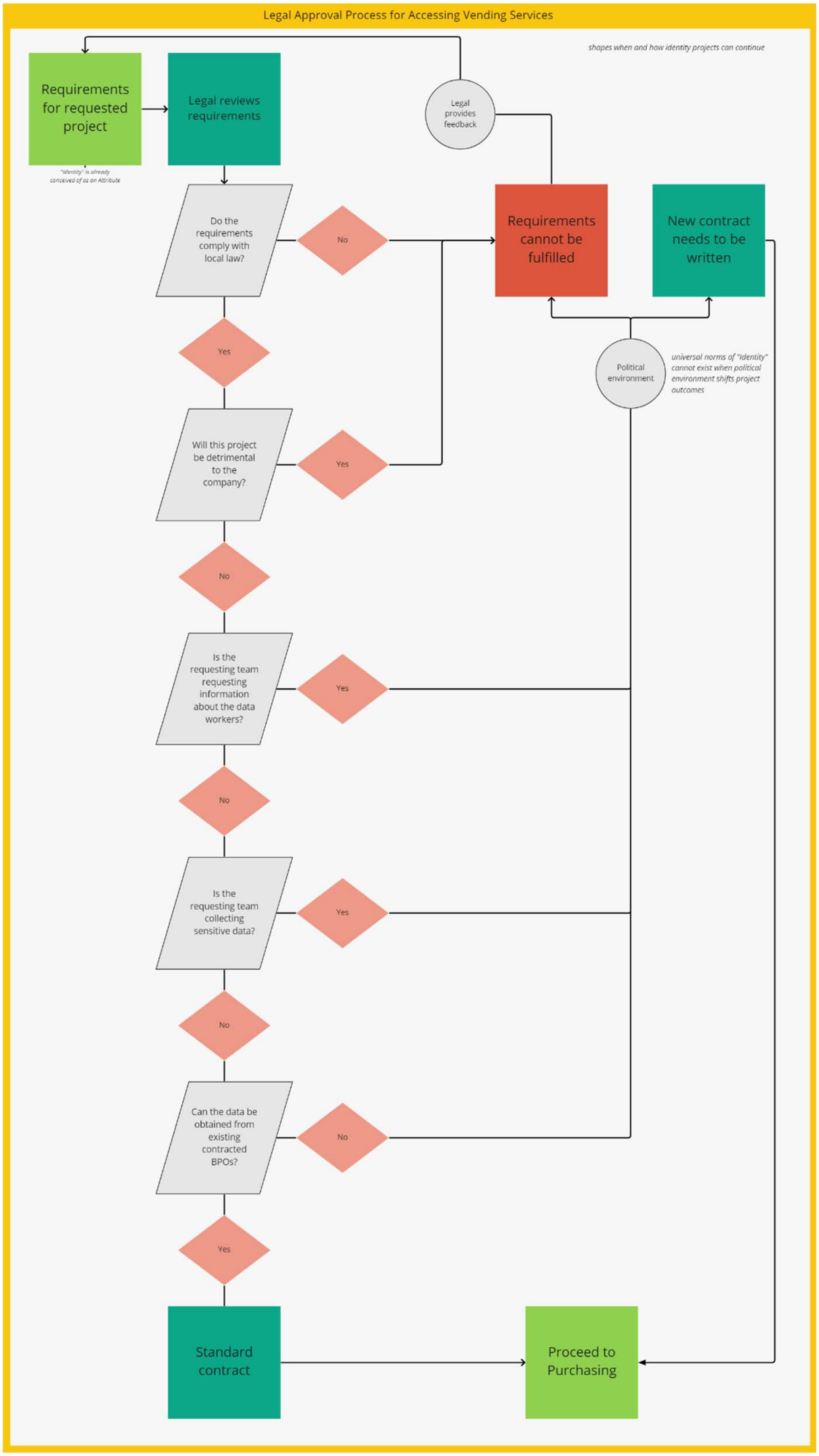

Legal Approval Process for Accessing Vending Services
shapes when and how identity projects can continue
Requirements for requested project
"Identity" is already conceived of as an Attribute
Legal reviews requirements
Legal provides feedback
Do the requirements comply with local law?
No
Requirements cannot be fulfilled
New contract needs to be written
Yes
Political environment
universal norms of "Identity" cannot exist when political environment shifts project outcomes
Will this project be detrimental to the company?
Yes
No
Is the requesting team requesting information about the data workers?
Yes
No
Is the requesting team collecting sensitive data?
Yes
No
Can the data be obtained from existing contracted BPOs?
No
Yes
Standard contract
Proceed to Purchasing

Figure 20. The above process map shows the legal approval process for traditional workers seeking data outsourcing vendors. Legal often has multiple checks before approving the purchasing of vendor services.

Some processes may also be put in place to avoid accidental violations. For example, a company may not allow anyone to collect facial data from Illinois, even if it does comply with the law, to avoid any accidental legal issues. Legal may have also decided that certain data is too risky to collect, certain countries/people are too risky to employ from, or certain things cannot be requested (e.g., recordings of workers doing their jobs). As described by Elliot in Chapter 8, companies may also have contracts with data vendors that do not allow traditional workers to collect information about data workers. Legal may also limit collecting sensitive data, such as biometric data for facial recognition, which may reshape how product teams can define an Attribute or if they can use the Attribute in their product at all. In cases where the legal team denies the proposed project, the product team must make changes to comply with legal and policy requirements.

Legal rejections, regardless of whether they are due to violations of legal law or to protect the parent company, particularly from downstream effects which might happen in Step Three, the data work step. Legal further constrains how data can be collected, where it can be collected from, and who can do the collecting. The way that the Attribute is shifted by legal is subtle. Limiting where data can be collected from constrains how an Attribute can be represented. If one cannot collect data in a certain setting, that setting will be absent from the final perspective of identity in the Technical Attribute. Similarly, limiting where annotation can occur, and limiting it to specific pre-approved companies, reflects the training and beliefs of those data workers.

Options for Outsourcing Data Work

Who is doing the data work influences how the Attribute becomes refined. If a team needs data collected or annotated by outside parties, they have several options. The first is to use third-party crowdworking platforms, such as Amazon Mechanical Turk, where workers across the globe can complete “micro tasks,” small individual tasks that individuals can complete quickly but eventually lead to a larger output—such as a fully annotated dataset. Elliot explains that, due to Maelstrom having strict policies about data protections, researchers and product teams cannot “just throw whatever they want onto AMT.” Largely, traditional tech workers felt that Attributes were more poorly refined when choosing crowdworking options, because they had less control over the workforce on crowdworking platforms. In other words, crowdworkers could influence Attributes more than desired (see Chapter 9).

A second option, most similar to the first option, is to use free crowdsourcing. Since Maelstrom is a large and well-known company, they have options for freely collecting gamified crowdsource data annotations and data validation. Most smaller companies, and even many large companies, do not have this option, so it is rather uncommon in the process of implementation. Once again, in using free crowdsourcing, traditional tech workers are concerned that Attributes might be shaped into something less closely aligned with their own visions.

A third option would be to employ contract workers. Contract workers are directly employed by the company. In the case of Maelstrom, they go through more bureaucratic channels like hiring firms to select contractors; however, many smaller companies employ contract workers directly from sites like Upwork. In using contract workers, traditional tech workers can directly interface with data workers, influencing their interpretations of data through guidelines and ensuring their work actually adheres to the guidelines.

The fourth and final option is to hire a vendor, or business process outsourcing (BPO) company, which specializes in data collection and/or annotation—such as Appen, Lionbridge, or EnVision Data. These companies act as intermediaries, supplying their own trained workforce but also managing that workforce, so that companies like Maelstrom need only supply instructions and approve the outputs. In the case of using vendors, the vendors themselves often influence how guidelines are structured, adding a layer of translation to the Attribute which would otherwise not occur in other options. For example, as documented in Chapter 8, EnVision Data's supervisors and trainers were involved in developing guidelines and overseeing data work activities.

As an example of how the choice in data worker type can influence the way an Attribute is refined, I describe the effects of outsourcing to a vendor by reviewing the overall processes at EnVision Data. **Figure 21** showcases the process of creating and refining guidelines for a project at EnVision Data. Once a client and EnVision Data agree to move forward with the proposed data project, then the implicit definition of what the Attribute is meant to be has to be translated into tangible directions for data workers to implement. At EnVision Data, a statement of work is written describing the product and its requirements. Two parties are involved in getting the project running: the operations team and the training manager.

Implementation of Description of Attribute in Data Guidelines

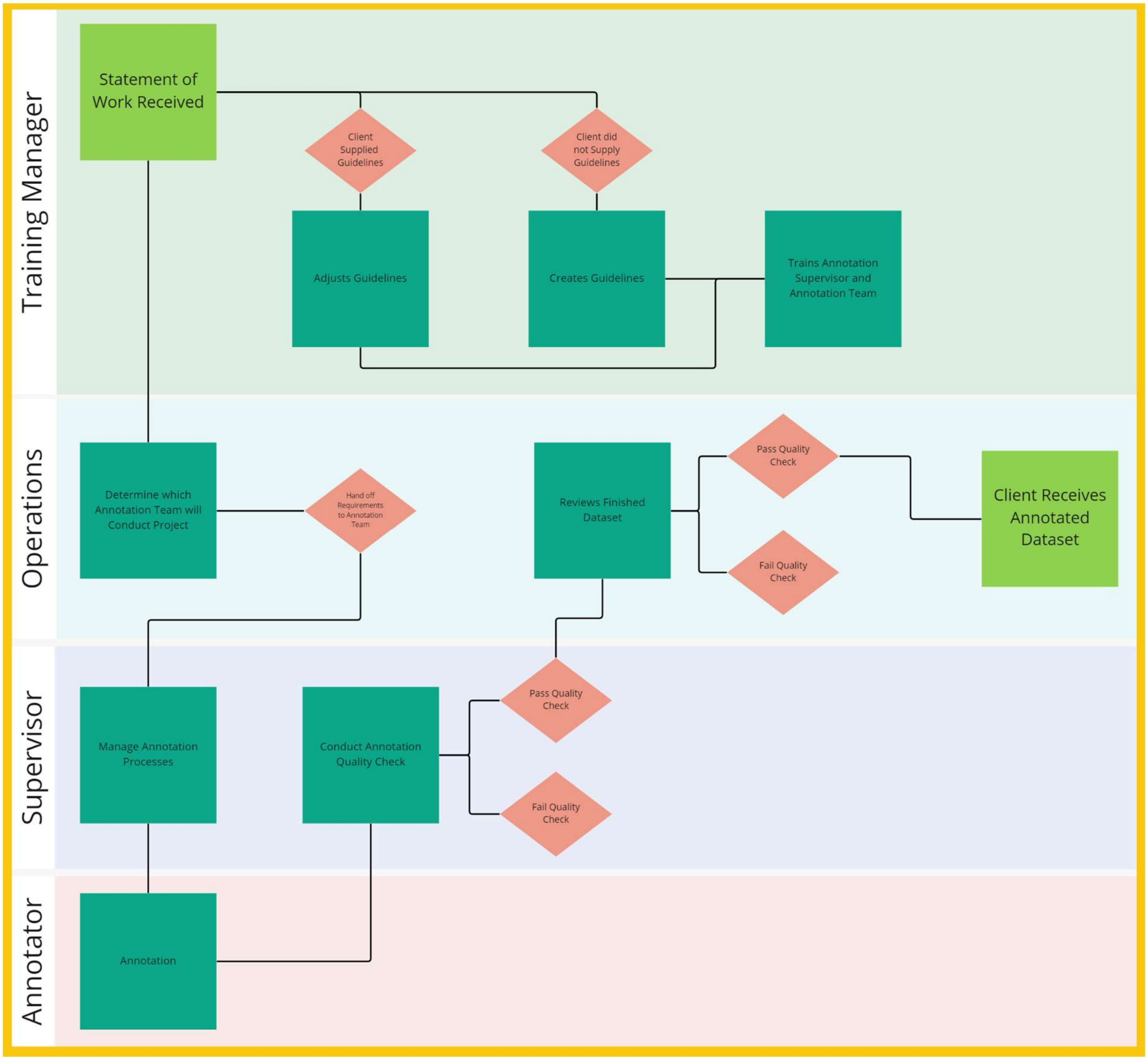

Training Manager
Operations
Supervisor
Annotator
Statement of Work Received
Client Supplied Guidelines
Client did not Supply Guidelines
Adjusts Guidelines
Creates Guidelines
Trains Annotation Supervisor and Annotation Team
Determine which Annotation Team will Conduct Project
Hand off Requirements to Annotation Team
Reviews Finished Dataset
Pass Quality Check
Fail Quality Check
Client Receives Annotated Dataset
Manage Annotation Processes
Conduct Annotation Quality Check
Pass Quality Check
Fail Quality Check
Annotation

Figure 21. **The** above process map shows the way that EnVision Data approached implementing Attributes in their clients' guidelines.

The operations team determines which local annotation team should conduct the project, based on their experience with past projects. Local annotation teams are situated in different countries, primarily in Central Asia. Certain teams may be more or less familiar with the types of Attributes and annotation of those Attributes than other teams, and familiarity is usually associated with expertise.

The training manager takes the documentation from the client for annotation guidelines. An important aspect of annotation guidelines is defining clearly the Categories which are meant to represent the Attribute. For example, the Attribute may be “emotion,” but there are many ways to represent “emotion.” Therefore, the training manager must list the potential categorical options for data workers, such as “angry” and “sad.” However, there are many Attributes treated as so obvious that the Categories are not clearly defined. “Gender” is an Attribute which often does not specify categories like “male” or “female,” but “male” and “female” naturally emerge as the expected categories. Defining categories for “gender” varied by project at EnVision Data, and usually only occurred when a client expected a certain distribution of “gender” Categories (e.g., the SensEyes project wanting an even split between “male” and “female”).

In some cases, the client supplies their own written annotation guidelines. In these cases, the training manager will simply adjust the guidelines to be more readable or understandable for the annotators, often translating them into Arabic at EnVision Data, given that the majority of their clients are in Western Europe and the majority of their workers are based in or are from Central Asia. Language translation necessitates the translation of the concepts of the Attribute, as many “identity” terms are described or discussed differently in different languages. Language is also known to shape how people view certain concepts.

In other cases, the training manager takes existing documentation to create guidelines from scratch. For example, they might be given a slide deck explaining the expected Attributes and then use that slide deck to create annotation instructions. This involves interpretation of the documentation to create actionable instructions. This approach may still involve language translation, given that the Attributes are largely named and explain in English.

Through the process of creating guidelines, the training manager acts as a translator of both ideas and language. They interpret the documentation given to them by clients to ensure annotators can apply the Attributes and Categories to data as expected. “As expected” is, however, different for different actors. The training manager also trains the annotation supervisor and annotators on the team, enforcing their vision of how the Attribute the client desires should be defined.

However, as documented in Chapter 8, workers often create vague and unclear guidelines. The lack of clarity of guidelines reflects two difficulties in capturing “Identity” for technical systems: (1) workers often don’t realize others do not share the same mental model of the Attribute as them; and (2) Attributes are always highly interpretable even when they are defined.

Step Three: Solidifying Attributes with Data Instances

By the start of Step Three, the Attribute and its Categories have been defined in the guidelines. These guidelines are then used to collect and annotate data instances—thus solidifying Attributes with actual examples. Beyond Step One, defining concepts about “Identity” into desirable Attributes, Step Three is the most impactful to shaping the representation of human identity in the final artifact. The Attribute changes from a theoretical representation to a concept represented by real data. What that data looks like and how it is annotated essentially defines the Attribute itself. Attributes are solidified in two ways: through data collection and through data annotation.

Data Collection

Data collection is the act of collecting image data representing the categories of a specific Attribute. During Step Two, the product team has determined where they want the data to come from. Generally, data is captured from real world settings or from online sources. Data from online sources might be collected via a company’s interface or taken from other websites. Which form of data collection is chosen also subtly shaped the Attribute.

Solidifying Attribute with Data Collection Processes

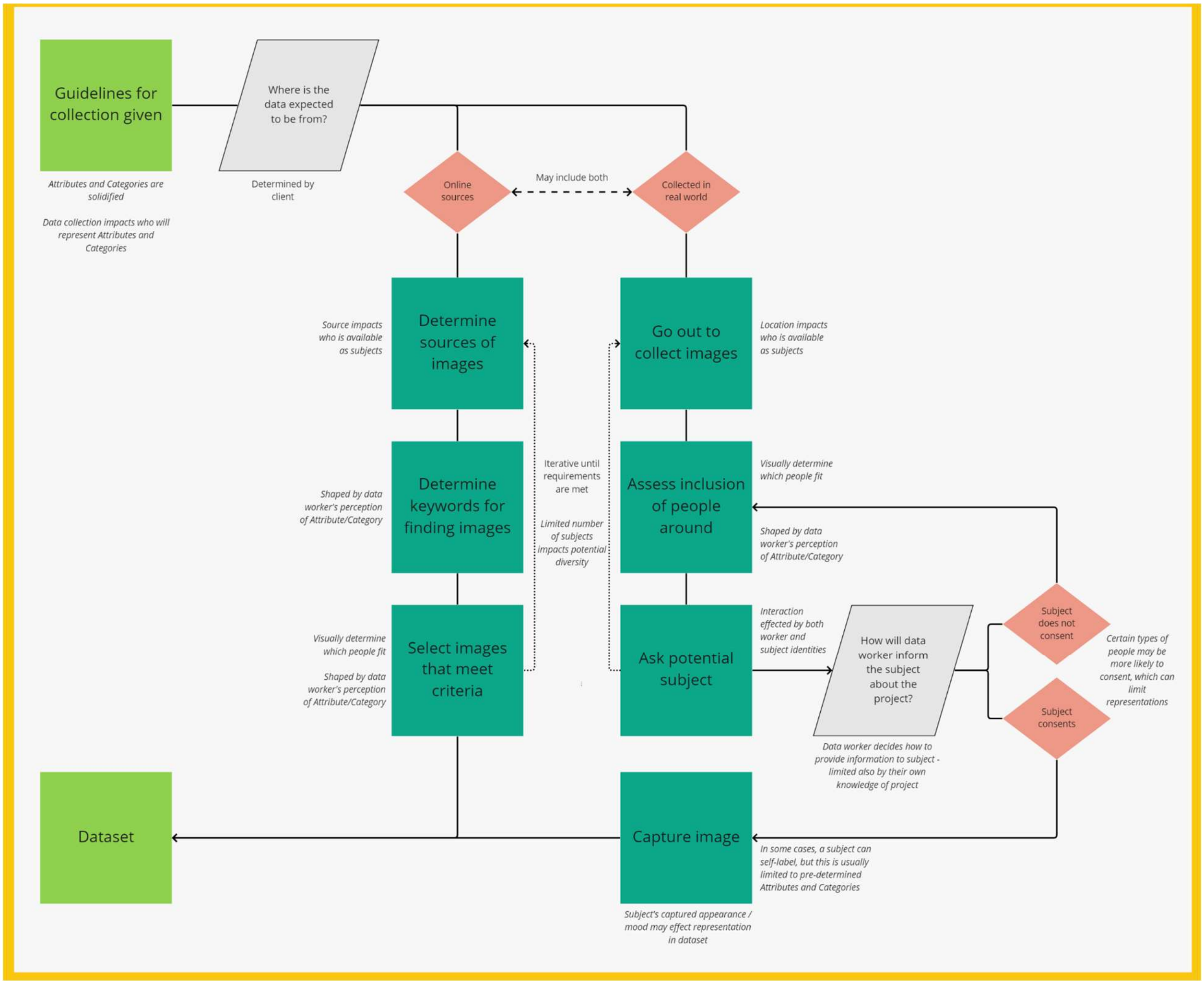

Guidelines for collection given
Attributes and Categories are solidified
Data collection impacts who will represent Attributes and Categories
Where is the data expected to be from?
Determined by client
Online sources
May include both
Collected in real world
Source impacts who is available as subjects
Determine sources of images
Go out to collect images
Location impacts who is available as subjects
Iterative until requirements are met
Limited number of subjects impacts potential diversity
Shaped by data worker's perception of Attribute/Category
Determine keywords for finding images
Assess inclusion of people around
Visually determine which people fit
Shaped by data worker's perception of Attribute/Category
Visually determine which people fit
Shaped by data worker's perception of Attribute/Category
Select images that meet criteria
Ask potential subject
Interaction effected by both worker and subject identities
How will data worker inform the subject about the project?
Data worker decides how to provide information to subject - limited also by their own knowledge of project
Subject does not consent
Subject consents
Certain types of people may be more likely to consent, which can limit representations
Dataset
Capture image
In some cases, a subject can self-label, but this is usually limited to pre-determined Attributes and Categories
Subject's captured appearance / mood may effect representation in dataset

Figure 22. A process map showing how an Attribute(s) in guidelines are solidified through the collection process.

When the data is collected in the real world, data collectors are tasked with going out and asking their connections or strangers for participation. A number of decisions rest on the data collectors' shoulders when taking this approach. Where the data is collected from, whether in a specific country, or in a city versus a rural area, or near a university, or in a specific neighborhood may all impact who is selected to represent the Attribute.

For example, Jeremy described that he needed images of hands for gesture recognition. Data of hands in different poses are collected by data workers. The data collected for Jeremy's product was broadly construed by the Attribute of "human diversity," where "age," "gender," and "skin tone" act as Categories with ideally as much variation as possible. He states that capturing such a range of diversity is difficult, because there is no way to control who decides to participate in data collection. Who is available to represent Categories under the concept of human variation shapes what the Attribute looks like—perhaps "human diversity" includes no images of children's hands. Therefore, children's hands, or hands of people under a certain age, are invisible to the Attribute "human diversity." There is also a level of how a subject's image is captured during a specific time and place, and so age, presentation, and mood are entangled with the Attributes collected.

Similarly, when data collectors are given specific Categories to collect for, they must visually assess the people around them to determine if they fit the Category. Visual assessment is not entirely accurate, and data collectors themselves acknowledge that it can be a guessing game. For example, how Manjola approached guessing "age" and how Gemma dealt with labeling race in Chapter 8. Whether someone is perceived as fitting a Category, and how they might fulfill that Category, is influenced by the data collectors' mental model of the Attribute and its Categories. A data collector might assume someone is female based on their presentation, or a child based on their height, or younger than they are based on their facial features.

Further, once the data collector identifies a potential subject, they have to approach them to capture their data. Data collectors not only have to ask permission to take their photo or for them to participate in uploading their own photo to an interface, they generally have to convince them that participating is worthwhile. We can see how certain data workers, like Jaako, navigated cultural barriers to collect data of people in his home country of Kenya in Chapter 8.

On the other hand, Nitesh's company used data uploaded by users on the company's platform. In this case, the Attributes used by the product team are constrained by what the platforms' users upload. The data is reflective of what the uploaders believe certain Attributes to be. Often, the data workers also decide the keywords for finding images, or they adapt suggested keywords from the guidelines to find more images. The keywords for finding data are shaped by the subjective perception an individual has of the Attribute, but the returned results are also shaped by the mental models of the original uploader. In selecting the images from results for each category, the data worker is visually determining which people fit with their vision of the Attribute but also aligning their vision of the Attribute with what they see in the search results.

The collection process is primarily shaping the Attribute through the lens of what *is* represented and visible, and what is not.

Data Annotation

Once data has been collected, it is also often annotated with the Categories traditional workers determined define the Attribute. Data workers look at each data instance and, following the guidelines created in Step Two, apply the Categories for each Attribute. The data worker labels these Categories based on their own assumptions about the Attribute.

Annotators are often expected to do different types of annotation, depending on the project. While the annotation process is the same regardless (assess image, apply annotation), the application of the Attribute is different depending on the type of annotation. The type of annotation is often connected to the application type. Categorical annotations involve selecting from predefined or “intuitive” categories (e.g., a list of 5 races, 2 genders, 3 clothing options, etc.). Spatial annotations focus on measurements, bounding boxes, polygons, etc. Finally, numerical annotations involve ratings or values (e.g., a scale of 1-10). The type of annotation means that data workers must approach, and think about, the Attribute from a specific perspective.

Solidifying Attribute with Data Collection Processes

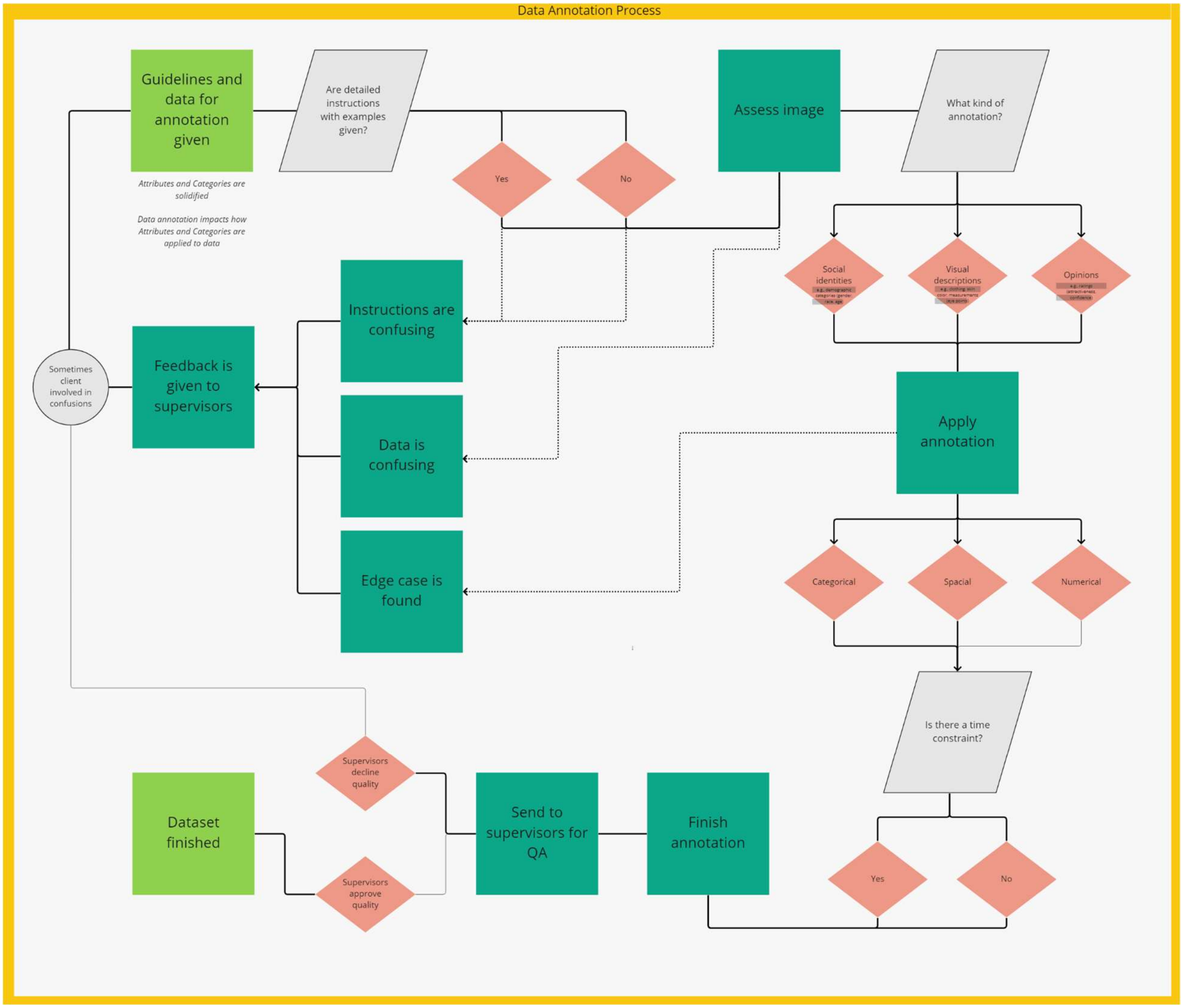

Data Annotation Process
Guidelines and data for annotation given
Attributes and Categories are solidified
Data annotation impacts how Attributes and Categories are applied to data
Are detailed instructions with examples given?
Yes
No
Assess image
What kind of annotation?
Social identities
e.g. demographic categories (gender, race, age
Visual descriptions
e.g. clothing, skin color, measurements, eye points
Opinions
e.g. ratings (attractiveness, confidence
Instructions are confusing
Data is confusing
Edge case is found
Sometimes client involved in confusions
Feedback is given to supervisors
Apply annotation
Categorical
Spacial
Numerical
Is there a time constraint?
Yes
No
Supervisors decline quality
Supervisors approve quality
Dataset finished
Send to supervisors for QA
Finish annotation

Figure 23. A process map showing how an Attribute(s) in guidelines are solidified through the annotation process.

As previously stated, guidelines may range from highly descriptive, with examples, or vague, expecting the annotators to simply know what the Attribute is. Certain Attributes tend to be much more vaguely defined. In particular, Attributes seen as highly tied to specific identity groups, like “gender” or “race” tend to be the least concretely defined, and rarely include visual examples. Traditional workers, in designing guidelines for specific identity Attributes, tend to assume their vision of them is obvious. Annotators are expected to intuitively “know” how to apply Categories. On the other hand, Attributes that may be perceived as less clearly tied to “Identity” like “emotion” or “culture” tend to be defined more clearly, often with visual examples. (Example) Using visual examples and more in-depth instructions for certain Attributes implies that workers view some Attributes as intuitive, clear, or obvious, in which everyone would share the same mental model.

Whether or not guidelines are detailed and provide examples also works to shape the outcome of the final Technical Attribute. While guidelines can provide annotators with an idea of what the client expects, a clearer image of the mental model the product team used to define the initial Attribute in Step One, they may also implicitly shape how annotators view all of the data. (Example). On the other hand, vague guidelines with no examples privileges the mental models of the annotators, so there may be more variation in how they interpret the Attribute.

Annotators may also be asked to provide their own opinions, or more subjective classifications, which inherently allow annotators to shape what the Attribute looks like. For example, they may be asked to rate the confidence of a person in an interview, like with the ChAI project. While the Categories are defined (a scale), the application is subjective. Usually subjective classifications are categorical (e.g., this person looks happy) or numerical (e.g., a rating of how happy a person looks).

Finally, some annotators may have a time constraint for annotating, expected to annotate x number of data instances in a certain time period or finish a whole project in a certain time period. Having a time constraint can lead annotators to move more quickly, engaging less critically with the data. On the other hand, without a time constraint, annotators may analyze data more carefully. The time constraint or lack thereof can shape the Attribute due to quickly labeled versus carefully labeled instances.

Step Four: Refining Attributes into Technical Attributes

After Attributes are concretized through the process of assigning data instances to their Categories, they are then further refined into finalized Technical Attributes. To refine an Attribute into a Technical Attribute, traditional workers review the data submitted to them by data workers. Quality checks are meant to ensure no mistakes are made during annotation. Mistakes indicate an Attribute that is implied "incorrectly," or not in line with the presumed shared mental model of what the Attribute should be. Here, traditional workers are able to regain control over the Attribute. They exercise their power in the computer vision development process to ensure that the data representing the Attribute meets their expectations.

Both data collection and data annotation are subject to revision. Refining—in the form of quality assurance—may be established in a number of ways. If traditional workers conducting reviews find the current state of the Attribute to be lacking in some way, the Attribute is then refined and subtly redefined through the process of revision.

Some traditional workers employ inter-annotator reliability to establish congruence between annotators (i.e., ensure the outcome of identity is homogenous). In such cases, Attributes are refined into Technical Attributes to reflect the perspectives of the majority. Those perspectives which are outliers are erased, so that the perspectives of the majority are represented. For example, if two data workers label an image "male" and one labels the same image "female," the image will be officially labeled with "male."

In other cases, traditional tech workers will check each data instance to determine whether they agree with the application of the annotation or not. When traditional workers find instances they disagree with, they then revise it. Revisions might occur by sending the data back to be re-collected or re-annotated by data workers, or it might involve traditional workers simply overriding the labels they disagree with, as occurred with the SensEyes project. Through this approach, traditional workers actively maintain their mental models for the Attribute as it is solidified into a Technical Attribute—though some implementations might be so vaguely implicitly different that they are not obvious during revision (e.g., the placement of keypoint annotations for different racial Attributes).

Step Four is the last step in filtering out undesirable perspectives on the Attribute. After review and revision are finished, the Attributes and their Categories, represented in the dataset, are given a stamp of approval. Once all of the undesirable perspectives have been filtered out of the Attribute, the Attribute is finalized. It has become narrowly defined. It is no longer malleable. It has become a Technical Attribute.

Step Five: Technical Attributes are Embedded in Artifacts

By Step Five, there is a solidified Technical Attribute. The Technical Attribute represents a very specific and narrow worldview of "Identity." Given the Technical Attribute is no longer malleable—it cannot be updated or changed—it is a completely closed model of "Identity." The Technical Attribute is solidified into artifacts, like datasets and the models trained on them. When the artifacts are deployed, they reinforce a very specific mental model of “Identity.”

The analyses of race and gender in Chapter 3 and gender in Chapter 4 are examples of how Technical Attributes are deployed. Gender in, for example, Microsoft Azure’s computer vision model has been entirely solidified by the data that Microsoft used to train their model. The Technical Attribute of “gender” (as defined in Microsoft Azure) is represented by the Technical Categories of “male” and “female” (as defined by the labeled data used to train the model). When the model is deployed on new subjects, they are classified on the basis of “gender” as represented in the data making up its Technical Categories. How “male” is defined is entirely based on what images are used to define “male” in the data the model was trained on. The model does not learn any new perspectives or worldviews about what “male” is—that is, unless workers return to prior steps to transform the Attribute again (e.g., by collecting new data, by adding new categories, etc.)

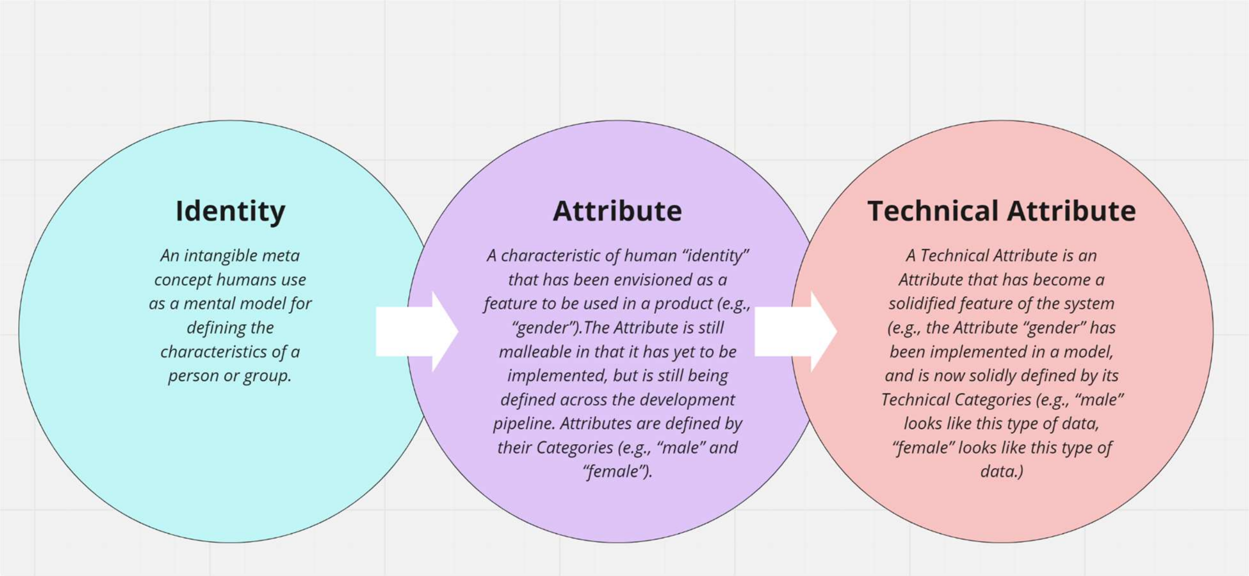
The consequence of Technical Attributes are that they implicitly, often invisibly enact very narrow and closed perspectives of identity on the world around them. These worldviews are also often portrayed as “objective” or “neutral” (see Chapters 3 and 5), despite being shaped by the positionalities of workers (see Chapters 7 and 8), who also hold different positional power (see Chapter 9). Untangling all of the positionalities embedded into Technical Attributes is an untenable task, because workers largely cannot identify how their own—or other’s—perspectives affect the transformation of identity. Further, despite these worldviews being incredibly narrow and value-laden, artifacts are used to make consequential decisions which negatively impact those who fall outside of the confines of the Technical Attributes embedded in them (see Chapter 4).

I have just shown how workers engage in a five-step transformation process during computer vision development practices. I documented how workers can make a variety of decisions during development that transform “Identity” into an Attribute and then a Technical Attribute. Next, I will describe at a high level how this five-step process is demonstrative of a model of identity present in technology development more broadly—what I call the open-to-closed model of identity.

The Open-to-Closed Model of Identity: Identity > Attribute > Technical Attribute

Identity is treated as an infrastructure on which to ground the development of a technical artifact. In building an artifact, human actors attempt to unearth underlying truths about identity that can be operationalized. As demonstrated in Chapter 3, the designers of artifacts like datasets present identity concepts as “ground truth.” This “ground truth” insinuates that the identity upon which the artifacts are built *is* reality, not simply a specific viewpoint of the world. However, the way that human actors actually “capture” identity through the development process showcases that there are numerous lenses for which to view identity. Therefore, grounding the development of a technical artifact in identity does not reflect any underlying “truth” about identity. Instead, it shows how workers reference their own positional vantage points to make sense of identity concepts. Workers pull threads from this nebulous and intangible concept of “identity” and attempt to weave together a clear, concise, and tight definition within a technical artifact. Though identity is viewed as an infrastructure from which to build technologies, in reality, technologies are grounded in the interpretations of the human actors who are designing them.

The Open-to-Closed Model of Identity



An intangible meta concept humans use as a mental model for defining the characteristics of a person or group.

A characteristic of human “identity” that has been envisioned as a feature to be used in a product (e.g., “gender”).The Attribute is still malleable in that it has yet to be implemented, but is still being defined across the development pipeline. Attributes are defined by their Categories (e.g., “male” and “female”).

A Technical Attribute is an Attribute that has become a solidified feature of the system (e.g., the Attribute “gender” has been implemented in a model, and is now solidly defined by its Technical Categories (e.g., “male” looks like this type of data, “female” looks like this type of data.)

Figure 24. A visualization of how the three phases that make up the open-to-closed model of identity.

The development of technology involves transforming the messy, complex, contingent concept of “Identity” into a solidified feature to be used by a system, a Technical Attribute. Different positional perspectives about “Identity” are negotiated and implemented to form a boundary concept, an Attribute. An Attribute is malleable, in that many actors are still making decisions about how to define and represent the Attribute through its Categories throughout the development process (see **Figure 19**). An Attribute is slowly made into something concrete throughout the development process, but finally becomes concretized once it is embedded into a computer vision artifact (a dataset and/or a model). These attributes thus become features of the artifact, Technical Attributes. Technical Attributes can no longer be changed, and do not act as a boundary object for making decisions. They work to enact a specific technical worldview of human identity characteristics. I define each of these terms in **Table 14**. In this section, I describe in more detail what each of the steps in this open-to-closed model of identity theory entails.

Outside Development	**During Development**		**Finalized Artifact**	
“Identity”	**Attribute**	**Categories**	**Technical Attribute**	**Technical Categories**
An intangible meta concept humans use as a mental model for defining the characteristics of a person or group.	A characteristic of human “identity” that has been envisioned as a feature to be used in a product (e.g., “gender”).The Attribute is still malleable in that it has yet to be implemented but is still being defined across the development pipeline. Attributes are defined by their Categories (e.g., “male” and “female”).	The values envisioned to make up the Attribute (e.g., for gender, the categories “male” and “female”). Categories are reflective of beliefs about what defines the Attribute. Like the Attribute, the Categories are still ideas and are still malleable during the development process.	A Technical Attribute is an Attribute that has become a solidified feature of the system (e.g., the Attribute “gender” has been implemented in a model and is now solidly defined by its Technical Categories (e.g., “male” looks like this type of data, “female” looks like this type of data.)	Technical Categories are defined by the data that is used to train a finalized system (e.g., the data collected for “male” now define the category of “male” for the Technical Attribute of “gender”)

Table 15. A table describing the different terms in the open-to-closed model of identity. “Identity” exists outside of the development process but is accessed by workers during development as they attempt to capture identity concepts in an Attribute and its Categories. Once development is complete and an artifact is created, it is imbued with a solidified Technical Attribute and its associated Technical Categories.

“Identity” exists as a meta concept; a concept which broadly encompasses the human experience and how individuals make sense of themselves and others, but which also differs greatly based on individual and cultural experiences. Individuals access their own mental models of “Identity” to make sense of the implementation of human characteristics in technologies. An individuals’ mental model of “Identity” is not necessarily cohesive, but an amalgamation of personal experiences, collective affinities, cultural and temporal values, and, in some cases, favored theories. In other words, these mental models of identities are formed based on individual positionalities. The exact mental model of identity of each individual worker is not necessary to define, and workers in my study largely cannot point towards any specific definitions governing their decisions. However, whether an individual adopts a postmodernist approach to “Identity” versus a more biologically essentialist one influences how they approach defining specific features for development (e.g., gender as fluid vs. gender as binary).

To necessitate the use of human characteristics of a feature, the broad concept of “Identity” is simplified and transformed into an “Attribute.” An Attribute is a characteristic of “Identity” that has been envisioned as a feature to be used in a product. For example, a product team might envision a model for classifying the gender of a person. When conceptualizing this product, the product team identifies “gender” as a key Attribute to the model. In identifying an Attribute, the product team is—often implicitly—referring to their own individual and collective mental model of what “gender” is. Throughout the development process, the Attribute acts as a boundary object—and often, a false friend. Workers may believe they have the same mental model, but every individual is employing a subtly different one. What one worker envisions for the Category of “male” is always going to be subtly different from another worker’s definition.

Positional mental models are reflected in the Categories of the Attribute that the workers identify as salient. Categories are the values envisioned to define the Attribute. For gender, the categories are often values like “male” and “female”—but they can also reflect other mental models of gender, such as “man” and “woman” or “feminine” and “masculine,” or include “non-binary.” The Categories used to represent the Attribute often indicate the beliefs about “Identity” held by the workers at the forefront of the development process—usually white collar traditional tech workers. In the example of “gender,” it may be that the workers have bio-essentialist beliefs about gender as binary and sex-dependent. Yet not all individuals may hold this mental model; some may disagree with the mental model they are expected to help implement (see Chapters 7 and 8). When disagreements occur between workers, they attempt to negotiate their perspectives with one another. However, in the end, “Identity” must be synthesized into one Technical Attribute. Both agreements and disagreements mutually shape the Attribute throughout its development, reflecting not only positional perspectives, but also power. Certain perspectives are always privileged over others. In my data, I describe how some trans workers at MultiplAI disagreed with the binary sex-based approach to gender in the original Attribute to be deployed. While they were unable to fundamentally change gender, they were able to shift the language of gender in the Attribute from “gender” to “gendered appearance” and the Categories to “masculine” and “feminine.” In the cases of retraining or updating the model, workers cycle back to defining an Attribute and its Categories; only when the Attribute becomes Technical, in that it is implemented in a working model, is the definition solidified.

As demonstrated through the MultiplAI example, during the development of the product, both the Attribute and its Categories are still malleable ideas. They may be treated or discussed among workers as if they are solidified, but each decision about the Attribute and its Categories during development shapes the finalized definition in the end product. The product team is always implicitly negotiating different mental models of “gender,” even in cases where there is a belief that their mental models align. Research teams may provide product teams conflicting definitions identified through user research. As data workers collect data for “gender” Categories, they are tacitly collecting data that reflects their own personal mental models. When annotators apply gender labels to data instances, they are enacting subjective judgments informed by their perspective about what “gender” looks like.

Much like traditional workers negotiating perspectives, but ultimately enabling certain perspectives and declining others, traditional workers enact a positional power over the open-to-closed model of turning “Identity” into a Technical Attribute. The diagram below (**Figure 25**) represents that, in developing requirements for computer vision, the positional perspectives of traditional workers are wide open. Many perspectives can be taken into account. As decisions are made in how to action those perspectives into Attributes, the wide-open possibilities of identity in computer vision narrow into specific (or not so specific) guidelines. From there, these narrow guidelines are given to data workers, whose positionality in interpreting and actioning on them is much more constrained by needing to follow the guidelines. Guidelines are artifacts that traditional workers employ to exert power over how identity is defined in computer vision. They are also used to control the role of the positionality of data workers in shaping datasets. As the workers apply the narrow guidelines using their own constrained positionality, identity calcifies into a very narrow, specific, stable artifact - the dataset. Quality assurance checks further allow traditional workers to transform the Attributes in the dataset into Technical Attributes that reflect their worldviews, not data workers.

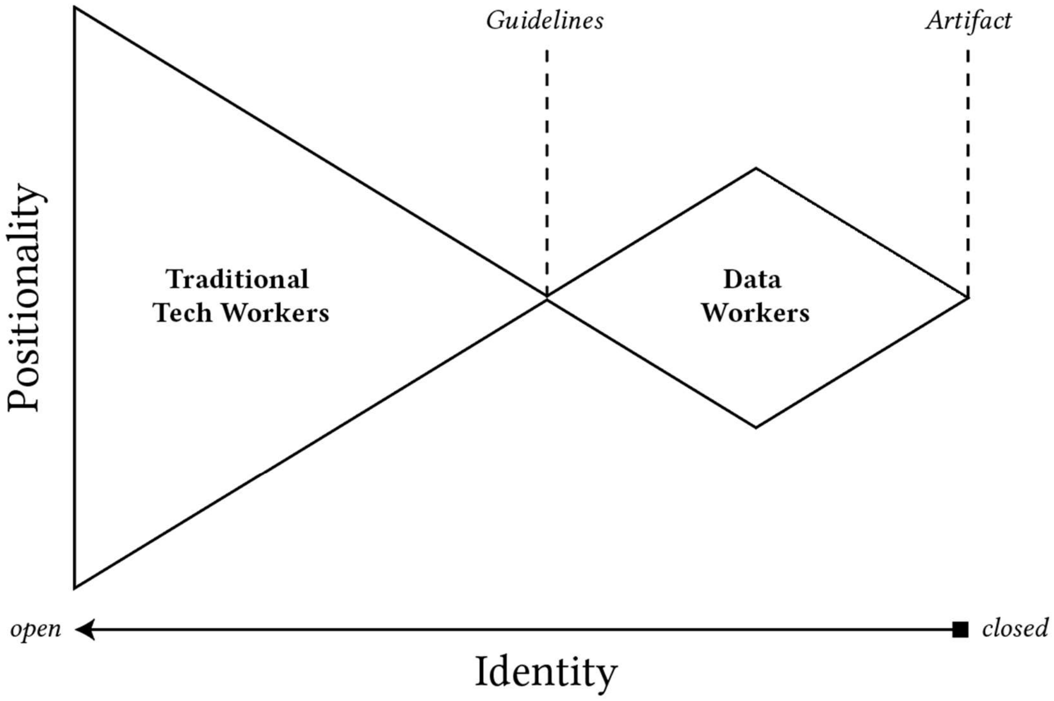


Figure 25. An illustration representing how worker positionality becomes more constrained as identity moves from open ("Identity") to closed (in Technical Attributes embedded in an "Artifact"). Traditional tech workers' positionalities start out more open, becoming increasingly constrained as they define "Identity" for an "Attribute." They use guidelines to ensure data worker positionalities are already constrained by the time they are involved in the process.

Once the dataset governing the intended computer vision product has been finalized and is used to train the model, the Attribute of "gender" is finally solidified into a Technical Attribute. Workers may have envisioned the Attribute itself as a finalized, working definition, but in reality, an Attribute is always in a state of (even if minor) flux. As demonstrated by the five-step transformation process in computer vision, Attributes are updated and refined through each worker's interaction.

On the other hand, a Technical Attribute is an Attribute that has become a solidified feature of the system with one specific definition. "Gender" has been implemented in a model and is now solidly defined by the way that its Technical Categories have been trained. The dataset used to define "gender" and its values are the one working definition of what "gender" is. The category of "male" in the product applies its classification based only on the data it has been told is "male." The new product, the model, now has its own individual mental model of gender—informed by all of the mental models that went into developing it—that it can apply to the world. "Gender" no longer exists as a fluid and intangible concept of "Identity," or as a negotiable and shifting feature like an Attribute. "Gender" now exists as a means of classification, and classifications are determined by its Technical Categories. "Male" represents a summation of learned patterns from the data selected to teach the system "male"; any human that falls outside of this pattern is likely to be unrecognizable as "male."

This open-to-closed model operates beyond the field of computer vision—any attempt to capture "Identity" for the sake of a technical artifact involves transforming it from something nebulous and open to something narrow and closed, something which can act as an affordance of a system. As "Identity" becomes solidified in the narrow reality of a "Technical Attribute," it is inevitable that certain perspectives, experiences, and positions will be marginalized. The "Technical Attribute" can only represent a limited worldview, one which those outside that worldview cannot be represented by. As artifacts defined by the limited worldviews embedded into artifacts via "Technical Attributes" are deployed, they inevitably cause harm to those whose identities aren't represented.

In the next section, I build on my theory that technology development operates within an open-to-closed model. Specifically, I discuss future directions for actively engaging with the open-to-closed model, including ways to imagine adopting a closed-to-open model instead.

Future Directions

Identity is central to the design of technology. In the realm of machine learning, identity is actively embedded into models, like computer vision models. As I have demonstrated through this dissertation, implementing identity in computer vision operates on an open-to-closed model of identity. Further, how closed models of identity in computer vision models have been historically implemented actively marginalizes specific groups, like transgender and non-binary communities.

At the close of this dissertation, I now want to consider the future of identity implementation in computer vision—and technology more broadly. I begin by considering the future for research on understanding identity implementations. This dissertation explored how workers approach identity decisions, and I highlight the gaps which still exist in fully understanding those decisions. Next, I detail the future of development in organizational contexts. In particular, I discuss the realities of product changes due to evolving perspectives on identity. Finally, I conclude by discussing the future of the design of identity in technology. I argue for approaches to identity that prioritize the perspectives of the marginalized and disempowered and contemplate ways to explore a closed-to-open model of identity.

Grounding Identity Decisions

To design technologies meaningful to human life, identity is crucial. Identity is embedded into technical infrastructures as an affordance. This is particularly salient in cases where technologies are expected to make decisions about human identity—like in machine learning technologies such as computer vision. In machine learning, it is common to claim that data is "ground truth," that it is somehow *grounded* in reality. As I have demonstrated in this dissertation, designing identity technologies requires human actors to actively engage with and make decisions about identity concepts, like gender, race, emotion, or individuality. Making such decisions requires individuals to consider how to define identity concepts, even though those considerations are informed by knowledge tacitly learned through their interactions with the world. The tacit knowledge that individuals rely on to make identity decisions for technology designs reflects their positions in the world, echoing their experiences, values, beliefs, and culture.

Those working on implementing computer vision products are certainly aware that identity is important. However, the tacit nature of implementing identity for technical systems makes it impossible for individuals to recognize, let alone explain, exactly how they make decisions. Even what identity means to different individuals is difficult to tease apart, requiring explicit articulation work to define. Certainly, a great deal of articulation work was foundational to this project. Discussions with the participants in this dissertation revealed difficulties in understanding identity practices in industry contexts, often evoking a fundamental question: what even is identity?

There are opportunities for us to reimagine how to *ground* identity in different perspectives, beliefs, and theories. As I proposed in the beginning of this chapter, the concept of identity can be construed as "Identity," a broad and nebulous concept which can be theorized and pursued endlessly. Outside of technical fields, scholars have theorized about the origins of identity and proposed new ways of thinking about identity concepts (e.g., Alcoff, 2006; Butler, 1988; Haraway, 2007). The intangible nature of identity was obvious in discussions with participants, who would quickly ground their perspectives in highly constrained categories, like "demographics" or "identifiers." It is clear that there is an outstanding challenge in grounding identity in a shared language or ways of thinking, so that we might understand how workers view identity concepts concretely and thus use those insights to implement identity differently.

While there are many different theories of identity, some tied to specific types of identities (e.g., gender), industry practitioners largely operate within a pragmatic lens which prioritizes intuition and shared beliefs. However, there is opportunity to develop mappings for what identity could mean in the context of technical development. Connecting different theories of identities to concrete ways of implementing those theories could provide practitioners with new lenses for which to consider their work. For example, what would it mean to take a poststructuralist approach to identity in technology development in comparison with a structuralist approach?

Even when discussing specific concepts of identity, like gender or race, participants couldn't explicitly describe how they conceptualize identity categories or how they ascribe them. While some participants might describe gender as "obvious," they could not describe what aspects of a person make gender obvious. Beyond presenting different approaches to thinking through "Identity," there are opportunities to understand how individuals apply a mental model of identity. For example, what would it mean to explore how individuals actually determine gender for each data instance? Understanding how individuals interpret identity categories at the data level, rather than at a higher conceptual level, can provide more grounded insights into how identity is actually implemented. Understanding implementation at more granular levels can triangulate more behavioral findings as presented in this dissertation. Further, it could lead to methods for providing concrete guidance on how to interpret identity categories, an improvement on current practices which are vague and assume shared interpretations.

Attending to positionality raises questions about what "ground truth." Identity is central to the design of technologies like computer vision, but practitioners developing such technologies currently rely on tacit knowledge gained through their positionalities. Given the tacit nature of implementing identity for technical artifacts, there are open challenges to grounding identity decisions in highly specific models for viewing and interpreting identity. Thus, there are still numerous opportunities for understanding what identity is for technical systems and how it is operationalized concretely.

Organizational Practices

The development of technologies like computer vision is done by individuals embedded in larger organization contexts, whether in academic (as seen in Chapters 3 and 5) or industry (as seen in Chapters 2, 7, and 9). Work conducted in organizational contexts is driven, loosely or tightly, by common standards and practices. Beyond contributing deeper insights into how identity can be defined and applied, there is opportunity to consider the role of organizational practices.

As I have demonstrated in this dissertation, the role of positionality is currently absent from organizational practice. Academic work on computer vision largely prioritizes efficiency, universality, impartiality, and model work. Workers in industrial contexts implicitly express these values or are otherwise constrained by them given they are positioned within companies. Much like I proposed in my discussions of Chapters 7, 8, and 9, shifting current practices towards more positionally-informed ones may result in more nuanced and informed models of identity. Shifting current practices towards new positionally-informed ones opens many new possibilities for research to determine how to do so and whether organizational changes are successful.

One area to focus on would include how to actively engage practitioners in their positional decision-making processes and document them. While there are increasingly tools for documenting the goals and limitations of datasets (Gebru et al., 2021) and models (Mitchell et al., 2018), there are no standards for documenting the social processes embedded in technology production. There is opportunity to, first, determine methods for engaging practitioners with their positionally-informed decision making, and, second, determine best practices for effectively documenting practitioner positions. Documenting positional processes in the development of computer vision would make transparent the reason for the specific identity schemas used in computer vision artifacts. Such transparency would not only make artifacts easier to audit but would help situate their use in specific limited contexts.

Updating legacy systems, in particular, reveal a number of pain points as different workers attempt to negotiate and influence the direction of the system. Ethnographically observing product teams engaged in updating legacy systems could highlight areas of opportunity for direct intervention. Are there ways to explicitly attend to moments of positional negotiation? Practitioners can be trained to actively consider the role of their positionality in the context of their organizational role. Work in this space could help shift the needle towards actively acknowledging and engaging with building artifacts *in context*, rather than for misguided universal goals. Further, it could also contribute methodological guidelines for researchers seeking to unearth tacit knowledge about decision making in many other organizational contexts.

Policy—either at the level of public policy or at the level of internal institutional policy—has the power to constrain or enable certain approaches to identity. This indicates that policy is an effective intervention point for shaping actual practice. Working with public policymakers to establish measures for tactfully and carefully attending to identity and mitigating harms is certainly ripe for new research. Often, policy in the AI space is focused on issues of individual privacy or vague notions of unfair outcomes, but what would it mean for public policy to directly address the gaps between evolving identity in reality and calcified identity in technologies?

Beyond public policy, there are also opportunities to shape corporate policy. Practitioners in my dissertation indicated a desire for better guidance on how best to implement identity concepts. Some participants relied on corporate policy to make arguments for changing legacy systems. Not only are there opportunities for better understanding the role of corporate policy in product development, there is space to engage practitioners directly in their vision for policy. Perhaps co-design activities with practitioners focused on implementing better corporate policies can better enable new developments of identity in technological artifacts.

While there are often many barriers to conducting work with organizations—particularly corporate organizations—organizations fundamentally shape how individuals are able to enact their own positional perspectives and shape identity in technology design. Thus, developing methods for engaging directly with organizations and focusing on shaping organizational practices is a fruitful area to make impacts on how identity can be represented in technology.

Empowering Design Futures

As previously mentioned, identity is crucial to technology design. However, the current state of identity in technology design is undesirable. It is static, calcified, and harmful, particularly to the most marginalized communities. It is often designed top-down, with traditional tech workers in powerful economic and social positions ascribing their values and then releasing products to be used on an uninvolved public. Even better understanding identity decisions and engaging organizations in better practices might not solve the fundamental issue—that identity is rendered onto others without input or consent.

The current state of identity in computer vision is premised on ground truth, that there is some underlying universal truth to be discovered and accurately classified in a model. In computer vision, the "user" is often not the person being classified. The "user" is the one seeking to classify others, applying their own lens of identity to "targets" for the sake of some task—like marketing, security, or assessment. "Targets" are forced to fit into the identity models embedded in the system. If they do not fit in as desired, they face the consequences. Perhaps they get irrelevant ads, perhaps they don't get the job, or perhaps they are even wrongfully arrested. The purpose of identity in computer vision models is to reflect the values of the user, ignoring the realities of those who will go on to be targeted by the system.

It makes sense, then, why computer vision—and other forms of corporate technology—are premised on an open-to-closed model of identity. Identity must be narrowly defined to fit the interests of the intended client. Even in cases where models are developed to be universally adopted, purchased as generic infrastructure by a variety of clients, the practitioners designing them are imagining corporate use cases—cases where a client is enacting a simplistic model of identity. In these cases, practitioners implicitly define identity in ways they believe are shared among their potential clientele.

But what would it look like for the "targets" of the system to become the "users"? How would different groups design computer vision for themselves? Centering those traditionally marginalized by identity in computer vision could result in new, unexplored ways of representing identity in AI models. I imagine how "gender" is represented in computer vision by traditional tech workers would be largely different from how trans individuals would prefer to represent "gender."

Rather than the default open-to-closed model of identity, what would it mean to explore a closed-to-open model of identity? Knowing what the status quo of identity looks like in computer vision, we can begin to imagine ways of purposefully reverse engineering it—transforming it from something that is closed to something that is open. Let's consider the "gender" model example. What would it mean to start from "male" and "female" and then purposefully break them down, disintegrating the logic of the binary currently embedded into models and actively disrupting it? Methods like participatory design, co-design, speculative design, and action research can uncover new visions for computer vision, embedded in the needs and desires of specific communities who are often "targets," and never "users." Centering the positionalities of historically marginalized communities can lead to empowering designs not possible within market-driven corporate organizations. Clearly, the models of identity in such corporate organizations are problematic; they are increasingly scrutinized for their failures, biases, and, as demonstrated in this dissertation, their restrictive models of identities.

As computer vision and other AI models increasingly, rapidly become a part of everyday life, it is necessary to shift lanes and reimagine new, alternative paths to the status quo. If we do not attend to positionality and continue to allow technologies to develop on an open-to-closed model reflective only of those in positions of power, then we will do more than enact specific instances of harm. As AI becomes ubiquitous, portraying its outcomes opaquely as truth, we will reify restrictive models of identity into broader social structures. All of the progress made on identity by social groups could be challenged by technological solutionism, embedded into retail, banking, employment, and social applications. Just as identity is imperative to the design of technologies, the design of technologies is imperative to the future of identity.

REFERENCES

Abdurrahim, S. H., Samad, S. A., & Huddin, A. B. (2018). Review on the effects of age, gender, and race demographics on automatic face recognition. In *The Visual Computer* (Vol. 34, Issue 11, pp. 1617–1630). Springer Berlin Heidelberg. https://doi.org/10.1007/s00371-017-1428-z

Afifi, M., & Abdelhamed, A. (2019). AFIF4: Deep gender classification based on AdaBoost-based fusion of isolated facial features and foggy faces. *Journal of Visual Communication and Image Representation*, *62*, 77–86. https://doi.org/10.1016/j.jvcir.2019.05.001

Afzal, S., Rajmohan, C., Kesarwani, M., Mehta, S., & Patel, H. (2021). *Data Readiness Report*. 42–51. https://doi.org/10.1109/smds53860.2021.00016

Agre, P. E. (1997). Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*. http://polaris.gseis.ucla.edu/pagre/critical.htmlhttp://polaris.gseis.ucla.edu/pagre/

Agüera y Arcas, B. (2017). Do algorithms reveal sexual orientation or just expose our stereotypes? *Medium*. https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Agüera y Arcas, B., Mitchell, M., & Todorov, A. (2017). Physiognomy's New Clothes. *Medium*. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

Alcoff, L. M. (2006). Visible Identities: Race, Gender, and the Self. In *Visible Identities: Race, Gender, and the Self*. Oxford University Press. https://doi.org/10.1093/0195137345.001.0001

Alim, H. S., Rickford, J. R., & Ball, A. F. (2016). *Raciolinguistics : how language shapes our ideas about race*. https://books.google.com/books?id=gVf0DAAAQBAJ&source=gbs_navlinks_s

Altenried, M. (2020). The platform as factory: Crowdwork and the hidden labour behind artificial intelligence. *Capital and Class*, *44*(2), 145–158. https://doi.org/10.1177/0309816819899410

Amazon Rekognition – Video and Image - AWS. (2019). https://aws.amazon.com/rekognition/

Ammari, T., Schoenebeck, S., & Lindtner, S. (2017). The Crafting of DIY Fatherhood. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 1109–1122. https://doi.org/10.1145/2998181.2998270

Amrute, S. (2019). Of Techno-Ethics and Techno-Affects. *Feminist Review*, *123*(1), 56–73. https://doi.org/10.1177/0141778919879744

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Anderson, E. (1995). Knowledge, Human Interests, and Objectivity in Feminist Epistemology. *Philosophical Topics*, *23*(2), 27–58. http://www.jstor.org/stable/43154207

Anderson, J., & Christen, K. (2013). 'Chuck a Copyright on it': Dilemmas of Digital Return and the Possibilities for Traditional Knowledge Licenses and Labels. *Museum Anthropology Review*, *7*(1–2), 105–126. https://scholarworks.iu.edu/journals/index.php/mar/article/view/2169

Anthias, F. (2008). Thinking through the lens of translocational positionality: an intersectionality frame for understanding identity and belonging. *Translocations: Migration and Social Change*, *4*(1), 5–19. https://repository.uel.ac.uk/item/8656x

Anusha, A. V., JayaSree, J. K., Bhaskar, A., & Aneesh, R. P. (2017). Facial expression recognition and gender classification using facial patches. *2016 International Conference on Communication Systems and Networks, ComNet 2016*, 200–204. https://doi.org/10.1109/CSN.2016.7824014

Anwar, M. A., & Graham, M. (2021). Between a rock and a hard place: Freedom, flexibility, precarity and vulnerability in the gig economy in Africa. *Competition and Change*, *25*(2), 237–258. https://doi.org/10.1177/1024529420914473

Anzures, G., Quinn, P. C., Pascalis, O., Slater, A. M., Tanaka, J. W., & Lee, K. (2013). Developmental Origins of the Other-Race Effect. In *Current Directions in Psychological Science* (Vol. 22, Issue 3, pp. 173–178). NIH Public Access. https://doi.org/10.1177/0963721412474459

Arora, P. (2016). Bottom of the Data Pyramid: Big Data and the Global South. *International Journal of Communication*, *10*. https://ijoc.org/index.php/ijoc/article/view/4297

Ashurst, C., Anderljung, M., Prunkl, C., Leike, J., Gal, Y., Shevlane, T., & Dafoe, A. (2020). A Guide to Writing the NeurIPS Impact Statement. *Medium*. https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832

Ashurst, C., Hine, E., Sedille, P., & Carlier, A. (2021). *AI Ethics Statements -- Analysis and lessons learnt from NeurIPS Broader Impact Statements*. https://doi.org/10.48550/arxiv.2111.01705

Ásta. (2018). Categories We Live By. In *Categories We Live By*. https://doi.org/10.1093/oso/9780190256791.001.0001

Attia, M., & Edge, J. (2017). Be(com)ing a reflexive researcher: a developmental approach to research methodology. *Open Review of Educational Research*, *4*(1), 33–45. https://doi.org/10.1080/23265507.2017.1300068

Bacchini, F., & Lorusso, L. (2019). Race, again: how face recognition technology reinforces racial discrimination. *Journal of Information, Communication and Ethics in Society*, *17*(3), 321–335. https://doi.org/10.1108/JICES-05-2018-0050

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Barlas, P., Kyriakou, K., Kleanthous, S., & Otterbacher, J. (2019). Social B(eye)as: Human and Machine Descriptions of People Images. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 583–591. https://ojs.aaai.org/index.php/ICWSM/article/view/3255

Barr, A. (2015). Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms. *Wall Street Journal*. https://www.wsj.com/articles/BL-DGB-42522

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Becher, T. (1987). Disciplinary Discourse. *Studies in Higher Education*, *12*(3), 261–274. https://doi.org/10.1080/03075078712331378052

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 711–720. https://doi.org/10.1109/34.598228

Bellamy, R. K. E., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., Zhang, Y., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., & Mehta, S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4–5). https://doi.org/10.1147/JRD.2019.2942287

Bender-Baird, K. (2015). Peeing under surveillance: bathrooms, gender policing, and hate violence. *Gender, Place & Culture*, *23*(7), 983–988. https://doi.org/10.1080/0966369x.2015.1073699

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Benford, S., Tolmie, P., Ahmed, A. Y., Crabtree, A., & Rodden, T. (2012). Supporting Traditional Music-Making: Designing for Situated Discretion. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 127–136. https://doi.org/10.1145/2145204.2145227

Bennett, C. L., Gleason, C., Scheuerman, M. K., Bigham, J. P., Guo, A., & To, A. (2021). ""It's Complicated"': Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. *CHI Conference on Human Factors in Computing Systems (CHI '21)*.

Benthall, S., & Haynes, B. D. (2019). Racial categories in machine learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 289–298. https://doi.org/10.1145/3287560.3287575

Bentley, R., Hughes, J. A., Randall, D., Rodden, T., Sawyer, P., Shapiro, D., & Sommerville, I. (1992). Ethnographically-Informed Systems Design for Air Traffic Control. *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work*, 123–129. https://doi.org/10.1145/143457.143470

Beraja, M., Yang, D. Y., Yuchtman, N., Kao, A., Lu, S., We, W. P., Hu, S., Liu, J., Ni, S., Quan, Y., Xu, L., Yang, P., Yin, G., Acemoglu, D., Bartelme, D., Bubb, R., Buera, P., Dal Bó, E., Donaldson, D., … Xu, D. (2020). *Data-intensive Innovation and the State: Evidence from AI Firms in China*. https://doi.org/10.3386/W27723

Bettcher, T. M. (2007). Evil Deceivers and Make-Believers: On Transphobic Violence and the Politics of Illusion. *Hypatia*, *22*(3), 43–65. https://doi.org/10.1111/j.1527-2001.2007.tb01090.x

Binder, A. J., Davis, D. B., & Bloom, N. (2016). Career Funneling: How Elite Students Learn to Define and Desire "'Prestigious"' Jobs. *Sociology of Education*, *89*(1), 20–39. https://doi.org/10.1177/0038040715610883

Bingham, G., & Yip, B. (2017). *MORPH-II Dataset: Summary and Cleaning*.

Birhane, A. (2020). Algorithmic Colonization of Africa. *SCRIPT-Ed*, *17*(2), 389–409. https://doi.org/10.2966/SCRIP.170220.389

Birhane, A., & Prabhu, V. U. (2021). Large image datasets: A pyrrhic win for computer vision? *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158

Bivens, R., & Haimson, O. L. (2016). Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. *Social Media + Society*, *2*(4), 1–12. https://doi.org/10.1177/2056305116672486

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and Its Consequences for Online Harassment: Design Insights from HeartMob ACM Reference format. *Proc. ACM Hum.-Comput. Interact. Proc. ACM Hum.-Comput. Interact. Article Proc. ACM Hum.-Comput. Interact*, *1*(2), 1–19. https://doi.org/10.1145/3134659

Bledsoe, W. W. (1964). The Model Method in Facial Recognition. *Technical Report, PRI 15, Panoramic Research, Inc.*

Boellstorff, T. (2008). Coming of age in second life: An anthropologist explores the virtually human. In *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human*. Princeton University Press. https://doi.org/10.1111/j.1757-6547.2009.00060.x

Bogle, A. (2022). Behind "miracle" AI is an army of "ghost workers" — and they're speaking out about Appen. *ABC News*. https://amp-abc-net-au.cdn.ampproject.org/c/s/amp.abc.net.au/article/101531084

Borgman, C. L. (2017). Big Data, Little Data, No Data: Scholarship in the Networked World. In *Big Data, Little Data, No Data*. The MIT Press. https://doi.org/10.7551/mitpress/9963.001.0001

Bourdieu, P. (1987). *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674212770

Bourdieu, P. (1990). *The logic of practice*. Stanford University Press.

Bourdieu, P. (1991). *Language and Symbolic Power: The Economy of Linguistic Exchanges*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674510418

Bowker, G. C. (2006). *Memory Practices in the Sciences*. MIT Press. https://doi.org/10.1108/00220410610688804

Bowker, G. C., & Star, S. L. (2000). *Sorting Things Out: Classification and Its Consequences*. MIT Press. https://doi.org/10.7326/0003-4819-135-10-200111200-00030

Brandom, R. (2019). *Crucial biometric privacy law survives Illinois court fight*. The Verge. https://www.theverge.com/2019/1/26/18197567/six-flags-illinois-biometric-information-privacy-act-facial-recognition

Breeze, R. (2011). Disciplinary values in legal discourse: A corpus study. *Iberica*, *21*, 93–116. https://revistaiberica.org/index.php/iberica/article/view/330

Brown, K., & Jackson, D. D. (2013). The history and conceptual elements of critical race theory. In *Handbook of Critical Race Theory in Education* (pp. 9–22). Routledge. https://doi.org/10.4324/9780203155721

Brubaker, J. R., & Hayes, G. R. (2011). SELECT * FROM USER: infrastructure and socio-technical representation. *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work - CSCW '11*, 369. https://doi.org/10.1145/1958824.1958881

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *. In *Proceedings of Machine Learning Research* (Vol. 81). http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Butler, J. (1988). Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory. *Theatre Journal*, *40*(4), 519. https://doi.org/10.2307/3207893

Cambo, S. A., & Gergle, D. (2022). Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3501998

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 67–74. https://doi.org/10.1109/FG.2018.00020

Casilli, A. A., & Posada, J. (2019). The Platformization of Labor and Society. In M. Graham & W. H. Dutton (Eds.), *Society and the Internet: How Networks of Information and Communication are Changing Our Lives* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780198843498.003.0018

Cealey Harrison, W., & Hood-Williams, J. (2002). *Beyond sex and gender*. SAGE.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press. https://psycnet.apa.org/record/1996-97863-000

Chancellor, S., Baumer, E. P. S., & De Choudhury, M. (2019). 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). https://doi.org/10.1145/3359249

Chang, H., Lu, J., Yu, F., & Finkelstein, A. (2018). PairedCycleGAN: Asymmetric Style Transfer for Applying and Removing Makeup. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 40–48. https://doi.org/10.1109/CVPR.2018.00012

Chardon, A., Cretois, I., & Hourseau, C. (1991). Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science*, *13*(4), 191–208. https://doi.org/10.1111/j.1467-2494.1991.tb00561.x

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. http://www.sxf.uevora.pt/wp-content/uploads/2013/03/Charmaz_2006.pdf

Chase, C. (1998). Hermaphrodites with attitude: Mapping the emergence of intersex political activism. *GLQ*, *4*(2), 189–211. https://doi.org/10.1215/10642684-4-2-189

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the Impact of Gender on Rank in Resume Search Engines. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Cheng, J., & Cosley, D. (2013). How Annotation Styles Influence Content and Preferences. *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, 214–218. https://doi.org/10.1145/2481492.2481519

Chin Leong, K. (2013). *Google Reveals Its 9 Principles of Innovation*. Fast Company. https://www.fastcompany.com/3021956/googles-nine-principles-of-innovation

Clarifai. (2019). https://clarifai.com/

Clemmensen, T., & Roese, K. (2010). *An Overview of a Decade of Journal Publications about Culture and Human-Computer Interaction (HCI) BT - Human Work Interaction Design: Usability in Social, Cultural and Organizational Contexts* (D. Katre, R. Orngreen, P. Yammiyavar, & T. Clemmensen (Eds.); pp. 98–112). Springer Berlin Heidelberg.

Coleman, B. (2009). Race as technology. In *Camera Obscura* (Vol. 24, Issue 1, pp. 177–207). Duke University Press. https://doi.org/10.1215/02705346-2008-018

Collins, P. H. (1990). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment* (Vol. 21, Issue 1). https://doi.org/10.2307/2074808

Collins, P. H. (1998). Some Group Matters: Intersectionality, Situated Standpoints, and Black Feminist Thought. In *Fighting Words Black Women and the Search for Justice*. University of Minnesota Press. https://manifold.umn.edu/read/fighting-words/section/abcf003c-ed0b-401d-afe4-5bb27f98f254

Contractor, D., McDuff, D., Haines, J. K., Lee, J., Hines, C., Hecht, B., Vincent, N., & Li, H. (2022). Behavioral Use Licensing for Responsible AI. *ACM International Conference Proceeding Series*, 778–788. https://doi.org/10.1145/3531146.3533143

Cooper, G., & Bowers, J. (1995). Representing the user: Notes on the disciplinary rhetoric of human-computer interaction. In *The Social and Interactional Dimensions of Human-Computer Interfaces* (pp. 48–66).

Corbett-Davies, S., & Goel, S. (2018). *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*. https://doi.org/10.1063/1.3627170

Crawford, K., & Paglen, T. (2019). *Excavating AI: The Politics of Images in Machine Learning Training Sets*. https://excavating.ai/

Crawford, K., & Schultz, J. (2019). *AI Systems as State Actors*. Columbia Law Review. https://columbialawreview.org/content/ai-systems-as-state-actors/

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color. *Source: Stanford Law Review*, *43*(6), 1241–1299. http://www.jstor.org/stable/1229039

Cunning, D. (2011). Cartesian dualism: theology, metaphysics, science. In *The Cambridge Companion to: Descartes' Meditations*. Cambridge University Press. https://doi.org/10.1017/CCO9781139088220

Currah, P., & Mulqueen, T. (2011). Securitizing Gender: Identity, Biometrics, and Transgender Bodies at the Airport. *Social Research*, *78*(2), 557–582. https://doi.org/10.1353/sor.2011.0030

da Silva, S. M., & Webster, J. P. (2018). Positionality and Standpoint. In *The Wiley Handbook of Ethnography of Education* (pp. 501–512). https://doi.org/https://doi.org/10.1002/9781118933732.ch22

Daigle, R. (2003). Student Workers: The Heart of the Help Desk. *Proceedings of the 31st Annual ACM SIGUCCS Fall Conference*, 193–195. https://doi.org/10.1145/947469.947520

Dame, A. (2016). Making a name for yourself: tagging as transgender ontological practice on Tumblr. *Critical Studies in Media Communication*, *33*(1), 23–37. https://doi.org/10.1080/15295036.2015.1130846

Dance, S. (2019). *Maryland set to add "X" gender designation to driver's licenses under bill by General Assembly*. Baltimore Sun. https://www.baltimoresun.com/news/maryland/politics/bs-md-drivers-licenses-20190313-story.html

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *IJCAI International Joint Conference on Artificial Intelligence*, 4691–4697. https://doi.org/10.24963/ijcai.2017/654

Das, A., Dantcheva, A., & Bremond, F. (2019). Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11129 LNCS*, 573–585. https://doi.org/10.1007/978-3-030-11009-3_35

Davani, A. M., Díaz, M., Research, G., & Prabhakaran, V. (2021). *Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations*. https://arxiv.org/abs/2110.05719v1

Davis, H. (2020). A Dataset is a Worldview. *Medium*. https://towardsdatascience.com/a-dataset-is-a-worldview-5328216dd44d

Davis, H. F. (2017). *Beyond Trans: Does Gender Matter?* https://books.google.com/books?id=uHA4DQAAQBAJ&source=gbs_navlinks_s

De Vries, K. M. (2012). Intersectional Identities and Conceptions of the Self: The Experience of Transgender People. *Symbolic Interaction*, *35*(1), 49–67. https://doi.org/10.1002/SYMB.2

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data and Society*, *8*(2), 205395172110359. https://doi.org/10.1177/20539517211035955

Denton, E., Hanna, A., Amironesei, R., Smart, A., Nicole, H., & Scheuerman, M. K. (2020). *Bringing the People Back In: Contesting Benchmark Machine Learning Datasets*. http://arxiv.org/abs/2007.07399

Dervin, F. (2015). Discourses of Othering. In *The International Encyclopedia of Language and Social Interaction* (pp. 1–9). Wiley. https://doi.org/10.1002/9781118611463.wbielsi027

Descartes, R. (1993). *Meditations on first philosophy* (T. Donald A. Cress (Ed.); 3rd ed.). Hackett Publishing Company.

Dias, E., Haberman, M., & Durston, E. A. (2019). Trump's Order to Combat Anti-Semitism Divides Its Audience: American Jews. *The New York Times*. https://www.nytimes.com/2019/12/12/us/politics/trump-anti-semitism-jews.html

Díaz, J. L. P., Dorn, A., Koch, G., & Abgaz, Y. (2020). A Comparative Approach between Different Computer Vision Tools, Including Commercial and Open-source, for Improving Cultural Image Access and Analysis. *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, 815–819. https://doi.org/10.1109/ACIT49673.2020.9208943

Díaz, M., Kivlichan, I., Rosen, R., Baker, D., Amironesei, R., Prabhakaran, V., & Denton, E. (2022).

CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2342–2351. https://doi.org/10.1145/3531146.3534647

Doffman, Z. (2019). *Is Microsoft AI Helping To Deliver China's "Shameful" Xinjiang Surveillance State?* Forbes. https://www.forbes.com/sites/zakdoffman/2019/03/15/microsoft-denies-new-links-to-chinas-surveillance-state-but-its-complicated/#4cb624f73061

Dombrowski, L., Harmon, E., & Fox, S. (2016). Social justice-oriented interaction design: Outlining key design strategies and commitments. *DIS 2016 - Proceedings of the 2016 ACM Conference on Designing Interactive Systems: Fuse*, 656–671. https://doi.org/10.1145/2901790.2901861

Dong, Z., Shi, C., Sen, S., Terveen, L., & Riedl, J. (2021). War Versus Inspirational in Forrest Gump: Cultural Effects in Tagging Communities. *Proceedings of the International AAAI Conference on Web and Social Media*, *6*(1), 82–89. https://doi.org/10.1609/icwsm.v6i1.14258

Driskill, Q.-L. (2016). *Asegi Stories: Cherokee Queer and Two-Spirit Memory*.

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(15). https://doi.org/10.1073/pnas.1322355111

Dunn, M. (2016). Digital Work: New Opportunities or Lost Wages? *Academy of Management Proceedings*, *2016*(1), 11689. https://doi.org/10.5465/ambpp.2016.11689abstract

Dym, B., Brubaker, J. R., Fiesler, C., & Semaan, B. (2019). "Coming Out Okay": Community Narratives for LGBTQ Identity Recovery Work. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–28. https://doi.org/10.1145/3359256

Easley, W., Hamidi, F., Lutters, W. G., & Hurst, A. (2018). Shifting Expectations: Understanding Youth Employees' Handoffs in a 3D Print Shop. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW). https://doi.org/10.1145/3274316

Edgerton, J. D., & Roberts, L. W. (2014). Cultural capital or habitus? Bourdieu and beyond in the explanation of enduring educational inequality. *Theory and Research in Education*, *12*(2), 193–220. https://doi.org/10.1177/1477878514530231

Edwards, W. K., Newman, M. W., & Poole, E. S. (2010). The infrastructure problem in HCI. *Conference on Human Factors in Computing Systems - Proceedings*, *1*, 423–432. https://doi.org/10.1145/1753326.1753390

Eisenthal, Y., Dror, G., & Ruppin, E. (2006). Facial attractiveness: Beauty and the machine. *Neural Computation*, *18*(1), 119–142. https://doi.org/10.1162/089976606774841602

England, K. V. L. (1994). Getting personal: Reflexivity, positionality, and feminist research. *Professional Geographer*, *46*(1), 80–89. https://doi.org/10.1111/j.0033-0124.1994.00080.x

Enwukwe, N. E. (2021). The Employment Status of Nigerian Workers in the Gig Economy: Using Uber as a Case Study. *Journal of Law, Policy and Globalization*. https://doi.org/10.7176/jlpg/107-08

Erikson, E. H. (Erik H., & Erikson, J. M. (Joan M. (1982). *The life cycle completed*. 134. https://www.worldcat.org/title/life-cycle-completed/oclc/916049006

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. https://www.google.com/books/edition/Automating_Inequality/pn4pDwAAQBAJ?hl=en&gbpv=0

Face API - Facial Recognition Software | Microsoft Azure. (2019). https://azure.microsoft.com/en-us/services/cognitive-services/face/

Face Recognition Data. (n.d.). Retrieved January 13, 2020, from https://cswww.essex.ac.uk/mv/allfaces/

Face Recognition Technology (FERET). (2017). *National Institute of Standards and Technology (NIST)*. https://www.nist.gov/programs-projects/face-recognition-technology-feret

Fang, R., Gallagher, A. C., Chen, T., & Loui, A. (2013). Kinship classification by modeling facial feature heredity. *2013 IEEE International Conference on Image Processing, ICIP 2013 - Proceedings*, 2983–2987. https://doi.org/10.1109/ICIP.2013.6738614

Farkas, L. G. (1994). *Anthropometry of the head and face*. Raven Press.

Farkas, L. G., Katic, M. J., Forrest, C. R., Alt, K. W., Bagič, I., Baltadjiev, G., Cunha, E., Čvičelová, M., Davies, S., Erasmus, I., Gillett-Netting, R., Hajniš, K., Kemkes-Grottenthaler, A., Khomyakova, I., Kumi, A., Kgamphe, J. S., Kayo-Daigo, N., Le, T., Malinowski, A., … Yahia, E. (2005). International anthropometric study of facial morphology in various ethnic groups/races. *Journal of Craniofacial Surgery*, *16*(4), 615–646. https://doi.org/10.1097/01.scs.0000171847.58031.9e

Fausto-Sterling, A. (2000). *Sexing the Body: Gender Politics and the Construction of Sexuality.* Basic

Books.

Feinberg, M. (2017). A Design Perspective on Data. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2952–2963. https://doi.org/10.1145/3025453.3025837

Feinberg, M., Carter, D., & Bullard, J. (2014). A Story Without End : Writing the Residual into Descriptive Infrastructure. *DIS '14 Proceedings of the Designing Interactive Systems Conference*, 385–394. https://doi.org/10.1145/2598510.2598553

Fellbaum, C. (2012). WordNet. In *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405198431.wbeal1285

Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media and Society*, *4*(1), 205630511876336. https://doi.org/10.1177/2056305118763366

Forman, J., & Damschroder, L. (2007). Qualitative Content Analysis. In *Advances in Bioethics* (Vol. 11, pp. 39–62). Emerald Group Publishing Limited. https://doi.org/10.1016/S1479-3709(07)11003-7

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon Mechanical Turk: Gold Mine or Coal Mine? *Comput. Linguist.*, *37*(2), 413–420. https://doi.org/10.1162/COLI_a_00057

Foucault, M. (1970). *The Order of Things: An Archaeology of the Human Sciences*. Pantheon Books.

Foucault, M. (1976). *The History of Sexuality: An Introduction*. Knopf Doubleday Publishing Group.

Friedman, B. (1996). Value-sensitive design. *Interactions*, *3*(6), 16–23. https://doi.org/10.1145/242485.242493

Fu, S., He, H., & Hou, Z. G. (2014). Learning race from face: A survey. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Vol. 36, Issue 12, pp. 2483–2509). IEEE Computer Society. https://doi.org/10.1109/TPAMI.2014.2321570

Fussell, S. (2019). *San Francisco Wants to Ban Government Face Recognition*. The Atlantic. https://www.theatlantic.com/technology/archive/2019/02/san-francisco-proposes-ban-government-face-recognition/581923/

Future of Privacy Forum. (2017). Unfairness By Algorithm: Distilling the Harms of Automated Decision-Making. *Future of Privacy Forum*, *December*, 1. https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/

Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., & Zhao, D. (2008). The CAS-PEAL large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *38*(1), 149–161. https://doi.org/10.1109/TSMCA.2007.909557

Garcia, P., & Cifor, M. (2019). Expanding Our Reflexive Toolbox: Collaborative Possibilities for Examining Socio-Technical Systems Using Duoethnography. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW). https://doi.org/10.1145/3359292

Garcia, P., Sutherland, T., Cifor, M., Chan, A. S., Klein, L., D'Ignazio, C., & Salehi, N. (2020). No: Critical Refusal as Feminist Data Practice. *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, 199–202. https://doi.org/10.1145/3406865.3419014

Garden, C. (2018). Disrupting Work Law: Arbitration in the Gig Economy. *University of Chicago Legal Forum*, *2017*. https://heinonline.org/HOL/Page?handle=hein.journals/uchclf2017&id=211&div=&collection=

Gargesha, M., & Panchanathan, S. (2002). A hybrid technique for facial feature point detection. *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, *2002-Janua*, 134–138. https://doi.org/10.1109/IAI.2002.999905

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. In *Communications of the ACM* (Vol. 64, Issue 12, pp. 86–92). https://doi.org/10.1145/3458723

Gehi, P. S., & Arkles, G. (2007). Unraveling Injustice: Race and Class Impact of Medicaid Exclusions of Transition-Related Health Care for Transgender People. *Sexuality Research & Social Policy Journal of NSRC Sexuality Research and Social Policy Journal of NSRC*, *4*(4), 7–35. https://doi.org/10.1525/srsp.2007.4.4.7

Geiger, S. R., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 325–336. https://doi.org/10.1145/3351095.3372862

Geva, M., Goldberg, Y., & Berant, J. (2019). Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. *Conference on Empirical Methods in Natural Language Processing*, 1161–1166. https://doi.org/10.18653/v1/d19-1107

Ghaffary, S., & Molla, R. (2019). Facial recognition: A map of where surveillance technology is in the US. *Vox*. https://www.vox.com/recode/2019/7/18/20698307/facial-recognition-technology-us-government-fight-for-the-future

Ghode, A. (2019). Data Work by Frontline Health Workers in Pregnancy Care. *Proceedings of the 10th Indian Conference on Human-Computer Interaction*. https://doi.org/10.1145/3364183.3364201

Giere, R. N. (2006). *Scientific Perspectivism*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226292144.001.0001

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., Goodall, J. L., Karlstrom, L., Lee, H., Mills, H. J., Oh, J.-H., Pierce, S. A., Pope, A., Tzeng, M. W., Villamizar, S. R., & Yu, X. (2016). Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science*, *3*(10), 388–415. https://doi.org/https://doi.org/10.1002/2015EA000136

Gilbert, S., Vitak, J., & Shilton, K. (2021). Measuring Americans' Comfort With Research Uses of Their Social Media Data. *Social Media + Society*.

Gillis, A. (2001). Digital sweatshops. *This*, *34*(4), 6. https://colorado.idm.oclc.org/login?url=https://www.proquest.com/magazines/digital-sweatshops/docview/203549187/se-2?accountid=14503

Goffman, E. (1956). The Presentation of Self in Everyday Life. *The Production of Reality: Essays and Readings on Social Interaction*, 262. https://books.google.com/books/about/The_Presentation_of_Self_in_Everyday_Lif.html?id=Sdt-cDkV8pQC

Goldman, L. R. (2000). *Social impact analysis: An applied anthropology manual*. Routledge. https://www.routledge.com/Social-Impact-Analysis-An-Applied-Anthropology-Manual/Goldman/p/book/9781859733929

Goldstein, A. G. (1979). Race-related variation of facial features: Anthropometric data I. *Bulletin of the Psychonomic Society*, *13*(3), 187–190. https://doi.org/10.3758/BF03335055

Gong, S., Liu, X., & Jain, A. K. (2019). *DebFace: De-biasing Face Recognition*. http://arxiv.org/abs/1911.08080

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org/

Gordon, L. R. (2007). Thinking through Identities: Black Peoples, Race Labels, and Ethnic Consciousness. In *The Other African Americans: Contemporary African and Caribbean Immigrants in the United States* (Vol. 45, Issue 07, pp. 69–92). https://books.google.com/books?hl=en&lr=&id=RVweAAAAQBAJ&oi=fnd&pg=PA69&dq=lewis+gordon+race&ots=z5jl27ODMI&sig=D3Nq5Ng_orDomTJkFWT9o7yw29Q#v=onepage&q=lewis gordon race&f=false

Goyal, N., Kivlichan, I. D., Rosen, R., & Vasserman, L. (2022). *Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation; Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation*. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Graham, D. B., & Allinson, N. M. (1998). Characterising Virtual Eigensignatures for General Purpose Face Recognition. In *Face Recognition* (pp. 446–456). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-72201-1_25

Graham, M., Hjorth, I., & Lehdonvirta, V. (2017). Digital labour and development: impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transfer*, *23*(2), 135–162. https://doi.org/10.1177/1024258916687250

Gray, M. L., & Siddharth, S. (2019). Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. In *Journal of Chemical Information and Modeling* (Vol. 53, Issue 9). https://www.hmhbooks.com/shop/books/ghost-work/9781328566287

Grother, P., Ngan, M., Hanaoka, K., & Ross, W. L. (2018). Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification. *NIST Interagency/Internal Report (NISTIR)*. https://doi.org/10.6028/NIST.IR.8238

Grudin, J. (1994). Computer-supported cooperative work: history and focus. *Computer*, *27*(5), 19–26.

https://doi.org/10.1109/2.291294

Guide to Social Science Data Preparation and Archiving: 6th Ed. (2012). ICPSR - Interuniversity Consortium for Political and Social Research. https://www.icpsr.umich.edu/web/pages/deposit/guide/

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). MS-celeb-1M: A dataset and benchmark for large-scale face recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9907 LNCS*, 87–102. https://doi.org/10.1007/978-3-319-46487-9_6

Gutta, S., Wechsler, H., & Phillips, P. J. (1998). Gender and ethnic classification of face images. *Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition, FG 1998*, 194–199. https://doi.org/10.1109/AFGR.1998.670948

Haimson, O. L., Brubaker, J. R., Dombrowski, L., & Hayes, G. R. (2015). Disclosure, Stress, and Support During Gender Transition on Facebook. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*, 1176–1190. https://doi.org/10.1145/2675133.2675152

Haimson, O. L., & Hoffmann, A. L. (2016). Constructing and enforcing "authentic" identity online: Facebook, real names, and non-normative identities. *First Monday*, *21*(6). https://doi.org/10.5210/fm.v21i6.6791

Halberstam, J. (1998). *Female Masculinity*. https://books.google.com/books/about/Female_Masculinity.html?id=5BqOswEACAAJ&source=kp_book_description

Hall, S. (2012). Introduction: Who Needs 'Identity'? In *Questions of Cultural Identity* (pp. 1–17). SAGE Publications Ltd. https://doi.org/10.4135/9781446221907.n1

Halverson, C. A., Erickson, T., & Ackerman, M. S. (2004). Behind the Help Desk: Evolution of a Knowledge Management System in a Large Organization. *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, 304–313. https://doi.org/10.1145/1031607.1031657

Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018). Gender Recognition or Gender Reductionism? The Social Implications of Automatic Gender Recognition Systems. *2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*.

Hanafi, H. (1998). Middle East in whose world? *Papers from the Fourth Nordic Conference on Middle Eastern Studies*, 1–9.

Hanley, M., Barocas, S., Levy, K., Azenkot, S., & Nissenbaum, H. (2021). Computer Vision and Conflicting Values: Describing People with Automated Alt Text. *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 543–554. https://doi.org/10.1145/3461702.3462620

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2019). Towards a Critical Race Methodology in Algorithmic Fairness. *FAT**. https://doi.org/10.1145/3351095.3372826

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575. https://doi.org/10.2307/3178066

Haraway, D. (2007). A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late 20th Century. In *The International Handbook of Virtual Learning Environments* (pp. 117–158). Springer Netherlands. https://doi.org/10.1007/978-1-4020-3803-7_4

Harding, S. (2004). *The Feminist Standpoint Theory Reader: Intellectual and Political Controversies*. Routledge. http://0-search.ebscohost.com.mercury.concordia.ca/login.aspx?direct=true&db=a9h&AN=12506415&site=ehost-live

Harrell, D. F. (2009). Computational and cognitive infrastructures of stigma. *Proceeding of the Seventh ACM Conference on Creativity and Cognition - C&C '09*, 49. https://doi.org/10.1145/1640233.1640244

Harrison, S., & Pan, B. (2020). *Mitigating Bias in Facial Recognition with FairGAN*.

Hawkins, A. (2017). *KFC China is using facial recognition tech to serve customers - but are they buying it?* The Guardian.

Hayes, G. R. (2011). The Relationship of Action Research to Human-Computer Interaction. *ACM Trans. Comput.-Hum. Interact.*, *18*(3). https://doi.org/10.1145/1993060.1993065

Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., & Ferrer, C. C. (2021). *Towards measuring fairness in AI: the Casual Conversations dataset*.

Heeks, R., Eskelund, K., Gomez-Morantes, J. E., Malik, F., & Nicholson, B. (2020). Digital Labour Platforms in the Global South: Filling or Creating Institutional Voids? *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3645389

Heeks, R., Graham, M., Mungai, P., Van Belle, J.-P., & Woodcock, J. (2020). Systematic Evaluation of Platform Work Against Decent Work Standards: Development of a New Framework and Application in the Global South. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3645151

Heikkilä, P., Honka, A., & Kaasinen, E. (2018). Quantified Factory Worker: Designing a Worker Feedback Dashboard. *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*, 515–523. https://doi.org/10.1145/3240167.3240187

Heinzerling, B. (2019). NLP's clever hans moment has arrived. *The Gradient*. https://thegradient.pub/nlps-clever-hans-moment-has-arrived/

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women Also Snowboard: Overcoming Bias in Captioning Models. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11207 LNCS*, 793–811. https://doi.org/10.1007/978-3-030-01219-9_47

Hendricks, M. L., & Testa, R. J. (2012). A conceptual framework for clinical work with transgender and gender nonconforming clients: An adaptation of the Minority Stress Model. *Professional Psychology: Research and Practice*, *43*(5), 460–467. https://doi.org/10.1037/a0029597

Herman, J. L. (2013). Gendered restrooms and minority stress: The public regulation of gender and its impact on transgender people's lives. *Journal of Public Management and Social Policy*, 65–80. https://williamsinstitute.law.ucla.edu/wp-content/uploads/Herman-Gendered-Restrooms-and-Minority-Stress-June-2013.pdf

Hicks, M. (2019). Hacking the Cis-tem: Transgender Citizens and the Early Digital State. *IEEE Annals of the History of Computing*, *41*(1), 1–1. https://doi.org/10.1109/mahc.2019.2897667

Hill Collins, P., & Bilge, S. (2016). *Intersectionality*.

Hirschman, C. (2004). The Origins and Demise of the Concept of Race. *Population and Development Review*, *30*(September), 385–415.

Hodge, J., Foley, S., Brankaert, R., Kenning, G., Lazar, A., Boger, J., & Morrissey, K. (2020). Relational, Flexible, Everyday: Learning from Ethics in Dementia Research. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376627

Hogg, M. A. (2016). Social Identity Theory. In S. McKeown, R. Haji, & N. Ferguson (Eds.), *Understanding Peace and Conflict Through Social Identity Theory: Contemporary Global Perspectives* (pp. 3–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-29869-6_1

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards*. http://arxiv.org/abs/1805.03677

Hollister, S. (2019). Google contractors reportedly targeted homeless people for Pixel 4 facial recognition. *The Verge*. https://www.theverge.com/2019/10/2/20896181/google-contractor-reportedly-targeted-homeless-people-for-pixel-4-facial-recognition

Holloway, I., & Biley, F. C. (2011). Being a Qualitative Researcher. *Http://Dx.Doi.Org/10.1177/1049732310395607*, *21*(7), 968–975. https://doi.org/10.1177/1049732310395607

Holstein, K., Daumé III, H., Dudík, M., Wallach, H., & Wortman Vaughan, J. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *CHI Conference on Human Factors in Computing Systems Proceedings*. ACM. https://doi.org/10.1145/3290605.3300830

Honeychurch, K. G. (1996). Researching dissident subjectivities: Queering the grounds of theory and practice. *Harvard Educational Review*, *66*(2), 339–355. https://doi.org/10.17763/haer.66.2.322km3320m402551

Hong, T., Wang, Z., Luo, X., & Zhang, W. (2020). State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, *212*, 109831. https://doi.org/https://doi.org/10.1016/j.enbuild.2020.109831

Howard, A., & Borenstein, J. (2018). The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering Ethics*, *24*(5), 1521–1536.

https://doi.org/10.1007/s11948-017-9975-2

Huang, X. D., Gates III, W. H., Horvitz, E. J., Goodman, J. T., Brunell, B. A., Dumais, S. T., Flake, G. W., Griffin, T. J., & Hurst-Hiller, O. (2006). *Targeted advertising in brick-and-mortar establishments* (Patent No. US8725567B2). https://patents.google.com/patent/US8725567B2/en

Hube, C., Fetahu, B., & Gadiraju, U. (2019). Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300637

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2021). Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188

IBM Data Privacy Passports. (n.d.). IBM. https://www.ibm.com/products/data-privacy-passports

Introduction to Race and Ethnic (Hispanic Origin) Data for the Census 2000 Special EEO File. (n.d.). U.S. Equal Employment Opportunity Commission. Retrieved March 16, 2023, from https://www.eeoc.gov/data/introduction-race-and-ethnic-hispanic-origin-data-census-2000-special-eeo-file

Irani, L. (2016). The Hidden Faces of Automation. *XRDS*, *23*(2), 34–37. https://doi.org/10.1145/3014390

Irani, L. C., & Silberman, M. S. (2016). Stories We Tell About Labor: Turkopticon and the Trouble with "Design." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4573–4586. https://doi.org/10.1145/2858036.2858592

Irani, L. C., & Silberman, M. S. (2013). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. *Conference on Human Factors in Computing Systems - Proceedings*, 611–620. https://doi.org/10.1145/2470654.2470742

Irani, L., Salehi, N., Pal, J., Monroy-Hernández, A., Churchill, E., & Narayan, S. (2019). Patron or poison? Industry funding of HCI research. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 111–115. https://doi.org/10.1145/3311957.3358610

Irving Seidman. (2006). *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences*. Teachers College Press. https://doi.org/10.1080/00220671.2014.938514

Jacobs, A. Z., & Wallach, H. (2019). *Measurement and Fairness*. http://arxiv.org/abs/1912.05511

James, S. E., Herman, J. L., Rankin, S., Keisling, M., Mottet, L., & Anafi, M. (2016). *The Report of the 2015 U.S. Transgender Survey*. http://www.transequality.org/sites/default/files/docs/usts/USTS Full Report - FINAL 1.6.17.pdf

Jeffrey Dastin. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 5–9. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306–316. https://doi.org/10.1145/3351095.3372829

Johnson, T., & Scott, J. C. (1999). *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press. https://doi.org/10.2307/525426

Jonathon Phillips, P., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(10), 1090–1104. https://doi.org/10.1109/34.879790

Julia Kapusta, S. (2016). Misgendering and Its Moral Contestability. *Hypatia*, *31*(3), 502–519. https://doi.org/10.1111/hypa.12259

Kaeser-Chen, C., Dubois, E., Schüür, F., & Moss, E. (2020). Positionality-Aware Machine Learning: Translation Tutorial. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 704. https://doi.org/10.1145/3351095.3375666

Kak, A. (2020). "the Global South is everywhere, but also always somewhere": National policy narratives & AI Justice. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 307–312. https://doi.org/10.1145/3375627.3375859

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, *583*(7815), 169. https://doi.org/10.1038/D41586-020-02003-2

Kamiran, F., & Calders, T. (2009). Classifying without discriminating. *2009 2nd International Conference on Computer, Control and Communication, IC4 2009*. https://doi.org/10.1109/IC4.2009.4909197

Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proceedings - 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000*, 46–53. https://doi.org/10.1109/AFGR.2000.840611

Kannabiran, G., & Petersen, M. G. (2010). Politics at the interface: A Foucauldian power analysis. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries (NordiCHI '10)*, 695–698. https://doi.org/10.1145/1868914.1869007

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018, October 27). Progressive growing of GANs for improved quality, stability, and variation. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Katz, L. F., & Krueger, A. B. (2019). The Rise and Nature of Alternative Work Arrangements in the United States, 1995–2015. *ILR Review*, *72*(2), 382–416. https://doi.org/10.1177/0019793918820008

Katzman, J., Wang, A., Scheuerman, M. K., Blodgett, S. L., Laird, K., Wallach, H., & Barocas, S. (2023). Taxonomizing and Measuring Representational Harms: A Look at Image Tagging. *AAAI*.

Kempe-Cook, L., Sher, S. T.-H., & Su, N. M. (2019). Behind the Voices: The Practice and Challenges of Esports Casters. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300795

Kemper, J., & Kolkman, D. (2018, June 18). Transparent to whom? No algorithmic accountability without a critical audience. *Information Communication and Society*, 1–16. https://doi.org/10.1080/1369118X.2018.1477967

Keyes, O. (2018). The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–22. https://doi.org/10.1145/3274357

Khan, L. (2011). Transgender Health at the Crossroads: Legal Norms, Insurance Markets, and the Threat of Healthcare Reform. *Yale Journal of Health Policy, Law & Ethics*, *11*(c), 375–418. https://heinonline.org/HOL/Page?handle=hein.journals/yjhple11&id=381&collection=journals&index=

Khan, M., & Hanna, A. (2022). The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4217148

Khosla, A., Bainbridge, W. A., Torralba, A., & Oliva, A. (2013). Modifying the memorability of face photographs. *Proceedings of the IEEE International Conference on Computer Vision*, 3200–3207. https://doi.org/10.1109/ICCV.2013.397

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., & Torralba, A. (2012). Undoing the damage of dataset bias. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7572 LNCS*(PART 1), 158–171. https://doi.org/10.1007/978-3-642-33718-5_12

Kirabo, L., Carter, E. J., Barry, D., & Steinfeld, A. (2021). Priorities, technology, and power: Co-designing an inclusive transit agenda in kampala, uganda. *Conference on Human Factors in Computing Systems - Proceedings*, 11. https://doi.org/10.1145/3411764.3445168

Klare, B. F., Burge, M. J., Klontz, J. C., Vorder Bruegge, R. W., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212

Knight, W. (2018). *Facial recognition has to be regulated to protect the public, says AI report*. MIT Technology Review. https://www.technologyreview.com/s/612552/facial-recognition-has-to-be-regulated-to-protect-the-public-says-ai-report/

Koch, B., Denton, E., Hanna, A., & Foster, J. G. (2021). *Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research*. https://paperswithcode.com

Kogan, S. L., & Muller, M. J. (2006). Ethnographic study of collaborative knowledge work. *IBM Systems Journal*, *45*(4), 759–771. https://doi.org/10.1147/sj.454.0759

Kong, Y. (2022). Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 485–494. https://doi.org/10.1145/3531146.3533114

Koopman, C. (2019). *How We Became Our Data: A Genealogy of the Informational Person*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226626611.001.0001

Kornilova, A. (2022). *Data annotation guidelines and best practices*. Snorkel AI. https://snorkel.ai/data-annotation/

Köstinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. *Proceedings of the IEEE International Conference on Computer Vision*, 2144–2151. https://doi.org/10.1109/ICCVW.2011.6130513

Krippendorff, K. (2018). Content Analysis: An Introduction to Its Methodology. In *Content Analysis: An Introduction to Its Methodology* (4th ed.). SAGE Publications, Inc. https://doi.org/10.4135/9781071878781

Krüger, K. (2010). The Destruction of Faces in Rwanda 1994: Mutilation as a Mirror of Racial Ideologies. *L'Europe En Formation*, *357*(3), 91. https://doi.org/10.3917/eufor.357.0091

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. The University of Chicago Press. https://doi.org/10.5840/philstudies196413082

Kuligowski, K. (2019). *Facial Recognition Advertising Targets Customers*. Business News Daily.

Kumar, N., Berg, A. C., Belhumeur, P. N., & Nayar, S. K. (2009). Attribute and simile classifiers for face verification. *Proceedings of the IEEE International Conference on Computer Vision*, 365–372. https://doi.org/10.1109/ICCV.2009.5459250

Kumar, R. S. S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., & Xia, S. (2020). Adversarial Machine Learning-Industry Perspectives. *2020 IEEE Security and Privacy Workshops (SPW)*, 69–75. https://doi.org/10.1109/SPW50608.2020.00028

Kunzel, R. (2014). The Flourishing of Transgender Studies. *TSQ: Transgender Studies Quarterly*, *1*(1–2), 285–297. https://doi.org/10.1215/23289252-2399461

Kutlu, M., McDonnell, T., Elsayed, T., & Lease, M. (2020). Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research*, *69*, 143–189. https://doi.org/10.1613/jair.1.12012

Lambrecht, A., & Tucker, C. E. (2016). Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. In *SSRN*. https://doi.org/10.2139/ssrn.2852260

Lamont, M. (2001). Culture and Identity. In J. H. Turner (Ed.), *Handbook of Sociological Theory* (pp. 171–185). Springer US. https://doi.org/10.1007/0-387-36274-6_9

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and Emotion*, *24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076

LaSala, M. C., Jenkins, D. A., Wheeler, D. P., & Fredriksen-Goldsen, K. I. (2008). LGBT faculty, research, and researchers: Risks and rewards. *Journal of Gay and Lesbian Social Services*, *20*(3), 253–267. https://doi.org/10.1080/10538720802235351

Leavitt, A. (2015). "This is a Throwaway Account": Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 317–327. https://doi.org/10.1145/2675133.2675175

Lee, J. U., Klie, J. C., & Gurevych, I. (2022). Annotation Curricula to Implicitly Train Non-Expert Annotators. *Computational Linguistics*, *48*(2), 343–373. https://doi.org/10.1162/coli_a_00436

Lesala Khethisa, B., Tsibolane, P., & Van Belle, J. P. (2020). Surviving the Gig Economy in the Global South: How Cape Town Domestic Workers Cope. *IFIP Advances in Information and Communication Technology*, *601*, 67–85. https://doi.org/10.1007/978-3-030-64697-4_7/TABLES/3

Lewis, A. E. (2003). Everyday Race-Making: Navigating Racial Boundaries in Schools. In *American Behavioral Scientist* (Vol. 47, Issue 3, pp. 283–305). https://doi.org/10.1177/0002764203256188

Li, S., & Deng, W. (2019). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, *28*(1), 356–370. https://doi.org/10.1109/TIP.2018.2868382

Li, S. Z., & Jain, A. K. (2011). Handbook of Face Recognition. In *Handbook of Face Recognition*. https://doi.org/10.1007/b138828

Liang, C. A., Munson, S. A., & Kientz, J. A. (2021). Embracing Four Tensions in Human-Computer Interaction Research with Marginalized People. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *28*(2), 14. https://doi.org/10.1145/3443686

Lien, J. J., Cohn, J. F., Kanade, T., & Li, C. C. (1998). Automated facial expression recognition based on FACS action units. *Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition, FG 1998*, 390–395. https://doi.org/10.1109/AFGR.1998.670980

Lin, A. (2017). *Facial recognition is tracking customers as they shop in stores, tech company says*.

CNBC. https://www.cnbc.com/2017/11/23/facial-recognition-is-tracking-customers-as-they-shop-in-stores-tech-company-says.html

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8693 LNCS*(PART 5), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Ling, H., Soatto, S., Ramanathan, N., & Jacobs, D. W. (2010). Face verification across age progression using discriminative methods. *IEEE Transactions on Information Forensics and Security*, *5*(1), 82–91. https://doi.org/10.1109/TIFS.2009.2038751

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*, *47*(2), 519–528. https://doi.org/10.3758/s13428-014-0483-x

Liu, J. (2023). What Big Tech companies are paying, based on new public salary data. *CNBC*. https://www.cnbc.com/2023/01/06/what-big-tech-companies-are-paying-based-on-new-public-salary-data.html

Locke, J. (1689). Of Identity and Diversity. In *The Works of John Locke, vol. 1 (An Essay concerning Human Understanding Part 1)*.

Lu, X., & Jain, A. K. (2004). Ethnicity Identification from Face Images. *Proceedings of SPIE*, *5404*, 114–123. https://doi.org/10.1117/12.542847

MacKay, W. E. (1999). Is Paper Safer? The Role of Paper Flight Strips in Air Traffic Control. *ACM Trans. Comput.-Hum. Interact.*, *6*(4), 311–340. https://doi.org/10.1145/331490.331491

Mackenzie, N., & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. In *Issues In Educational Research* (Vol. 16). http://msessd.ioe.edu.np/wp-content/uploads/2017/04/Handout4L4pages11-Research-Dilemmas-etc.pdf

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Conference on Human Factors in Computing Systems - Proceedings*, 1–14. https://doi.org/10.1145/3313831.3376445

Mager, M., Gutierrez-Vasques, X., Sierra, G., & Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the {A}mericas. *Proceedings of the 27th International Conference on Computational Linguistics*, 55–69. https://aclanthology.org/C18-1006

Mahalingam, G., & Kambhamettu, C. (2011). Can discriminative cues aid face recognition across age? *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, 206–212. https://doi.org/10.1109/FG.2011.5771399

Mahalingam, G., & Ricanek, K. (2013). Is the eye region more reliable than the face? A preliminary study of face-based recognition on a transgender dataset. *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS 2013)*, 1–7. https://doi.org/10.1109/BTAS.2013.6712710

Makena Kelly. (2019). *Pressure mounts on Google, Microsoft, and Amazon's facial recognition tech*. The Verge. https://www.theverge.com/2019/1/15/18183789/google-amazon-microsoft-pressure-facial-recognition-jedi-pentagon-defense-government

Mansoor Roomi, S. M., Virasundarii, S. L., Selvamegala, S., Jeevanandham, S., & Hariharasudhan, D. (2011). Race classification based on facial features. *Proceedings - 2011 3rd National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, NCVPRIPG 2011*, 54–57. https://doi.org/10.1109/NCVPRIPG.2011.19

Marcia, J. E. (1966). Development and Validation of Ego-Identity Status. In *Journal ol Personality and Social Psychology* (Vol. 3, Issue 5). https://pdfs.semanticscholar.org/f145/f3fbada1eb7a01052255f586094301669287.pdf

Marciano, A. (2019). Reframing biometric surveillance: from a means of inspection to a form of control. *Ethics and Information Technology*, *21*(2), 127–136. https://doi.org/10.1007/s10676-018-9493-1

Marcus, G. E. (1995). Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography. *Annual Review of Anthropology*, *24*, 95–117. http://www.jstor.org/stable/2155931

Markham, A. (2012). Fabrication as Ethical Practice: Qualitative Inquiry in Ambiguous Internet contexts. *Information Communication and Society*, *15*(3), 334–353. https://doi.org/10.1080/1369118X.2011.641993

Martinez, A. M. (1998). The AR face database. *CVC Technical Report 24*. https://doi.org/10.1023/B:VISI.0000029666.37597

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. In *Proceedings of the ACM on Human-Computer Interaction* (Vol. 3, Issue CSCW, p. 23). https://doi.org/10.1145/3359174

McDowell, L. (1992). Doing Gender: Feminism, Feminists and Research Methods in Human Geography. *Transactions of the Institute of British Geographers*, *17*(4), 399. https://doi.org/10.2307/622707

McFarland, M. (2016). Terrorist or pedophile? This start-up says it can out secrets by analyzing faces. *Washington Post*. https://www.washingtonpost.com/news/innovations/wp/2016/05/24/terrorist-or-pedophile-this-start-up-says-it-can-out-secrets-by-analyzing-faces/

McGarry, G., Chamberlain, A., Crabtree, A., & Greehalgh, C. (2021). The Meaning in “the Mix”: Using Ethnography to Inform the Design of Intelligent Tools in the Context of Music Production. *Proceedings of the 16th International Audio Mostly Conference*, 40–47. https://doi.org/10.1145/3478384.3478406

McGarry, G., Tolmie, P., Benford, S., Greenhalgh, C., & Chamberlain, A. (2017). “They’re All Going out to Something Weird”: Workflow, Legacy and Metadata in the Music Production Process. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 995–1008. https://doi.org/10.1145/2998181.2998325

McKinnon, J. D., & Horwitz, J. (2019). HUD Action Against Facebook Signals Trouble for Other Platforms. *The Wall Street Journal*. https://www.wsj.com/articles/u-s-charges-facebook-with-violating-fair-housing-laws-11553775078

McLemore, K. A. (2015). Experiences with Misgendering: Identity Misclassification of Transgender Spectrum Individuals. *Self and Identity*, *14*(1), 51–74. https://doi.org/10.1080/15298868.2014.950691

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). *A Survey on Bias and Fairness in Machine Learning*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Meiling, B. (2021). *Employers bow to tech workers in hottest job market since the dot-com era*. Los Angeles Times. https://techxplore.com/news/2021-07-employers-tech-workers-hottest-job.html

Meissner, J. L., Pretterhofer, N., Bergmann, N., & Haselsteiner, E. (2022). The Hidden Technological Labour of Service Workers in Health and Beauty Shops. *2022 Symposium on Human-Computer Interaction for Work*. https://doi.org/10.1145/3533406.3533413

Merler, M., Ratha, N., Feris, R. S., & Smith, J. R. (2019). *Diversity in Faces*.

Merriam, S. B., Johnson-Bailey, J., Lee, M. Y., Kee, Y., Ntseane, G., & Muhamad, M. (2001). Power and positionality: Negotiating insider/outsider status within and across cultures. *International Journal of Lifelong Education*, *20*(5), 405–416. https://doi.org/10.1080/02601370120490

Metcalf, J., & Crawford, K. (2016). Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society*, *3*(1), 2053951716650211. https://doi.org/10.1177/2053951716650211

Metcalf, J., Moss, E., & Boyd, D. (2019). Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly*, *86*(2), 449–476. https://doi.org/10.1353/SOR.2019.0022

Metz, C. (2019). Is Ethical A.I. Even Possible? *The New York Times*. https://www.nytimes.com/2019/03/01/business/ethics-artificial-intelligence.html

Metz, R. (2019). *Amazon shareholders want it to stop selling facial-recognition tech to the government*. CNN. https://www.cnn.com/2019/01/17/tech/amazon-shareholders-facial-recognition/index.html

Miceli, M., & Posada, J. (2021). *Wisdom for the Crowd: Discoursive Power in Annotation Instructions for Computer Vision*. https://arxiv.org/abs/2105.10990v1

Miceli, M., & Posada, J. (2022). *The Data-Production Dispositif*. https://doi.org/10.48550/arxiv.2205.11963

Miceli, M., Posada, J., & Yang, T. (2022). Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP). https://doi.org/10.1145/3492853

Miceli, M., Schuessler, M., & Yang, T. (2020). Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 25. https://doi.org/10.1145/3415186

Miceli, M., Yang, T., Naudts, L., Schuessler, M., Serbanescu, D., & Hanna, A. (2021). Documenting computer vision datasets: An invitation to reflexive data practices. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 161–172. https://doi.org/10.1145/3442188.3445880

Miller, S. (2022). U.S. Tech Salaries Averaged Above Six Figures in 2021. *Society for Human Resource Management*. https://www.shrm.org/resourcesandtools/hr-topics/compensation/pages/us-tech-salaries-averaged-above-six-figures-in-2021.aspx

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2018). Model Cards for Model Reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Monticelli, R. De. (2002). Personal Identity and Depth of the Person: Husserl and the Phenomenological Circles of Munich and Gottingen. In *Phenomenology World-Wide* (pp. 61–74). Springer Netherlands. https://doi.org/10.1007/978-94-007-0473-2_4

Moreno, A. B., & Sánchez, A. (2004). GavabDB: A 3D Face Database. *Proc. 2nd COST275 Workshop on Biometrics on the Internet, 2004*, 75–80.

Morning, A. (2008). Ethnic classification in global perspective: A cross-national survey of the 2000 census round. *Population Research and Policy Review*, *27*(2), 239–272. https://doi.org/10.1007/s11113-007-9062-5

Moses, J. & Knutsen, T. (2019). *Ways of knowing: Competing methodologies in social and political research*. https://www.bloomsbury.com/us/ways-of-knowing-9781352005530/

Mozur, P. (2019). One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. *New York Times*. https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html

Mtsweni, E. S., & Mavetera, N. (2018). Individual Barriers of Tacit Knowledge Sharing within Information System Development Projects. *Proceedings of the First International Conference on Data Science, E-Learning and Information Systems*. https://doi.org/10.1145/3279996.3280036

Muiruri, D., Lwakatare, L. E., Nurminen, J. K., & Mikkonen, T. (2022). Practices and Infrastructures for Machine Learning Systems: An Interview Study in Finnish Organizations. *Computer*, *55*(6), 18–29. https://doi.org/10.1109/MC.2022.3161161

Mukhopadhyay, C. C. (2018). Getting Rid of the Word "Caucasian." In *Privilege: A Reader* (pp. 231–236). Routledge. https://doi.org/10.4324/9780429494802-26

Muller, M., Aragon, C., Guha, S., Kogan, M., Neff, G., Seidelin, C., Shilton, K., & Tanweer, A. (2020). Interrogating Data Science. *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, 467–473. https://doi.org/10.1145/3406865.3418584

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C., & Erickson, T. (2019). How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300356

Muller, M., Wolf, C. T., Andres, J., Desmond, M., Joshi, N. N., Ashktorab, Z., Sharma, A., Brimijoin, K., Pan, Q., Duesterwald, E., & Dugan, C. (2021). Designing Ground Truth and the Social Life of Labels. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445402

Murgia, M. (2019). Who's using your face? The ugly truth about facial recognition. *Financial Times*. https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e

Murphy, M. (2017). The Economization of Life. In *The Economization of Life*. Duke University Press. https://doi.org/10.2307/j.ctv11smr89

Musa, N. C. and M. (2019). *The new assembly lines: Why AI needs low-skilled workers too*. World Economic Forum. https://www.weforum.org/agenda/2019/08/ai-low-skilled-workers/

Muth, L. (2018). *Why the Gender on My License Is Female Even Though I'm Nonbinary*. Allure. https://www.allure.com/story/nonbinary-gender-identity-drivers-license

Muthukumar, V., Pedapati, T., Ratha, N., Sattigeri, P., Wu, C.-W., Kingsbury, B., Kumar, A., Thomas, S., Mojsilovic, A., & Varshney, K. R. (2018). *Understanding Unequal Gender Classification Accuracy from Face Images*. https://github.com/ox-vgg/vgg_face2

Namaste, V. K. (2000). Invisible Lives: The Erasure of Transsexual and Transgendered People. *Contemporary Sociology*, *31*(3), 264. https://doi.org/10.2307/3089651

Narayanan, S. (2019). *An Update About Face Recognition on Facebook*. Facebook. https://about.fb.com/news/2019/09/update-face-recognition/

Natalia Drozdiak. (2019). *Microsoft Seeks to Restrict Abuse of its Facial Recognition AI*. Bloomberg. https://www.bloomberg.com/news/articles/2019-01-23/microsoft-seeks-to-restrict-abuse-of-its-facial-recognition-ai

National Center for Transgender Equality. (2015). *2015 US Transgender Survey Report on the Experiences of Black Respondents*. http://www.transequality.org/sites/default/files/docs/usts/USTS-Black-Respondents-Report.pdf

Newton, C. (2019). The secret lives of Facebook moderators in America. *The Verge*. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Ng, C. B., Tay, Y. H., & Goi, B. M. (2015). A review of facial gender recognition. *Pattern Analysis and Applications*, *18*(4), 739–755. https://doi.org/10.1007/s10044-015-0499-6

Ngan, M., & Grother, P. (2015). Face Recognition Vendor Test (FRVT) - Performance of Automated Gender Classification Algorithms. In *National Institute of Standards and Technology (NIST)*. https://doi.org/10.6028/NIST.IR.8052

Nik, M. A., Nik, M. A., Dehshibi, M. M., & Bastanfard, D. A. (2007). *Iranian Face Database and Evaluation with a New Detection Algorithm*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.418.771

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. https://nyupress.org/books/9781479837243/

O'Brien, S. A. (2018). *What is Amazon's responsibility over its facial recognition tech?* CNN. https://money.cnn.com/2018/07/26/technology/amazon-facial-recognition/index.html

O'Sullivan, J. (2019). *Washington Senate approves consumer-privacy bill to place restrictions on facial recognition*. The Seattle Times. https://www.seattletimes.com/seattle-news/politics/senate-passes-bill-to-create-a-european-style-consumer-data-privacy-law-in-washington/

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Ogbonnaya-ogburu, I. F., Smith, A. D. R., To, A., & Toyama, K. (2020). Critical Race Theory for HCI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. https://doi.org/10.1145/3313831.3376392

Olson, G. M., & Olson, J. S. (2000). Distance matters. *Human-Computer Interaction*, *15*(2–3), 139–178. https://doi.org/10.1207/S15327051HCI1523_4

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

OpenQA. (2016). *SeleniumHQ - Browser Automation*. Selenium. https://selenium.dev/

Oshodi, A. (2022). *How to Write Better Annotation Guidelines for Human Labelers: 4 Top Tips*. Superb AI. https://superb-ai.com/blog/how-to-write-better-annotation-guidelines-for-human-labelers-4-top-tips/

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in Deploying Machine Learning: A Survey of Case Studies. *ACM Comput. Surv.*, *55*(6). https://doi.org/10.1145/3533378

Pandey, Y. N., Rastogi, A., Kainkaryam, S., Bhattacharya, S., & Saputelli, L. (2020). Machine Learning in the Oil and Gas Industry. *Machine Learning in the Oil and Gas Industry*. https://doi.org/10.1007/978-1-4842-6094-4

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the reuse of scientific data. *Data Science Journal*, *16*(0), 1–9. https://doi.org/10.5334/DSJ-2017-008/METRICS/

Passi, S., & Jackson, S. J. (2018). Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW). https://doi.org/10.1145/3274405

Patton, D., Blandfort, P., Frey, W., Gaskell, M., & Karaman, S. (2019). Annotating Social Media Data From Vulnerable Populations: Evaluating Disagreement Between Domain Experts and Graduate Student Annotators. *Hawaii International Conference on System Sciences 2019 (HICSS-52)*. https://aisel.aisnet.org/hicss-52/dsm/critical_and_ethical_studies/4

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11), 100336. https://doi.org/https://doi.org/10.1016/j.patter.2021.100336

PBS. (2015). *A Map of Gender-Diverse Cultures*. PBS. http://www.pbs.org/independentlens/content/two-spirits_map-html/

Peng, K., Mathur, A., & Narayanan, A. (2021). *Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers*. https://arxiv.org/abs/2108.02922v1

Perrigo, B. (2022a). *Inside Facebook's African Sweatshop*. Time. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/

Perrigo, B. (2022b). Meta Accused Of Human Trafficking and Union-Busting in Kenya. *Time*. https://time.com/6175026/facebook-sama-kenya-lawsuit/

Perrigo, B. (2023). *OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive*. Time. https://time.com/6247678/openai-chatgpt-kenya-workers/

Phillips, A. (2020). Making a Face: Quantizing Reality in Character Animation and Customization. In *Gamer Trouble* (pp. 66–99).

Phillips, P. J., Narvekar, A., O'Toole, A. J., Jiang, F., & Ayyad, J. (2011). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception*, *8*(2), 1–11. https://doi.org/10.1145/1870076.1870082

Pincus, F. L. (2019). From Individual to Structural Discrimination. In *Race and Ethnic Conflict* (pp. 120–124). Routledge. https://doi.org/10.4324/9780429497896-13

Pittman, M., & Sheehan, K. (2016). Amazon's Mechanical Turk a Digital Sweatshop? Transparency and Accountability in Crowdsourced Online Research. *Journal of Media Ethics*, *31*(4), 260–262. https://doi.org/10.1080/23736992.2016.1228811

Polanyi, M. (2009). The Tacit Dimension. In *The Tacit Dimension*. The University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/T/bo6035368.html

Pomranz, M. (2017). *Beer Billboard Uses Facial Recognition to Advertise Only to Women*. Food & Wine. https://www.foodandwine.com/fwx/drink/beer-billboard-uses-facial-recognition-advertise-only-women

Ponterotto, D. (2016). Resisting the Male Gaze: Feminist Responses to the "Normatization" of the Female Body in Western Culture. *Journal of International Women's Studies*, *17*(1), 133–151. http://vc.bridgew.edu/jiws

Porcheron, A., Mauger, E., Soppelsa, F., Liu, Y., Ge, L., Pascalis, O., Russell, R., & Morizot, F. (2017). Facial contrast is a cross-cultural cue for perceiving age. *Frontiers in Physiology*, *8*(JUL), 1208. https://doi.org/10.3389/fpsyg.2017.01208

Porter, J. P., & Olson, K. L. (2001). Anthropometric facial analysis of the African American woman. *Archives of Facial Plastic Surgery : Official Publication for the American Academy of Facial Plastic and Reconstructive Surgery, Inc. and the International Federation of Facial Plastic Surgery Societies*, *3*(3), 191–197. https://doi.org/10.1001/archfaci.3.3.191

Powell, J. L. (2015). "Disciplining" Truth and Science: Michel Foucault and the Power of Social Science. *World Scientific News*, *13*, 15–29. www.worldscientificnews.com

Prewitt, K. (2005). Racial classification in America: Where do we go from here? *Daedalus*, *134*(1), 5–17. https://doi.org/10.1162/0011526053124370

Qinghao, W., & Dengkai, Y. (2017). A New Method for Evaluating Air Traffic Control Safety. *Proceedings of the 2017 VI International Conference on Network, Communication and Computing*, 217–221. https://doi.org/10.1145/3171592.3171640

Qu, S. Q., & Dumay, J. (2011). The qualitative research interview. In *Qualitative Research in Accounting and Management* (Vol. 8, Issue 3, pp. 238–264). Emerald Group Publishing Ltd. https://doi.org/10.1108/11766091111162070

Race Reporting Guide. (2015). Race Forward. https://www.raceforward.org/reporting-guide

Raengo, A. (2013). On the sleeve of the visual: Race as face value. In *On the Sleeve of the Visual: Race as Face Value*.

Raji, I. D., & Buolamwini, J. (2019). *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving Face: Investigating the ethical concerns of facial recognition auditing. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–151. https://doi.org/10.1145/3375627.3375820

Raji, I. D., Scheuerman, M. K., & Amironesei, R. (2021). You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 515–525. https://doi.org/10.1145/3442188.3445914

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW1). https://doi.org/10.1145/3449081

Ramanathan, N., & Chellappa, R. (2006). Modeling age progression in young faces. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *1*, 387–394. https://doi.org/10.1109/CVPR.2006.187

Ramey, A., & Salichs, M. A. (2014). Morphological Gender Recognition by a Social Robot and Privacy Concerns. *Proceedings of the 2014 ACM/IEEE Iternational Conference on Human-Robot Interaction (HRI '14)*, 272–273. https://doi.org/10.1145/2559636.2563714

Ramnani, M. (2022). Is AI becoming a technology built on worker exploitation? *Analytics India Mag*. https://analyticsindiamag.com/is-ai-fast-becoming-a-technology-built-on-worker-exploitation-from-global-south/

Rankin, Y. A., & Thomas, J. O. (2019). Straighten up and fly right: Rethinking intersectionality in HCI research. *Interactions*, *26*(6), 64–68. https://doi.org/10.1145/3363033

Raviv, S. (2020). The Secret History of Facial Recognition. *Wired*.

Reeves, B. N., & Shipman, F. (1996). Tacit Knowledge: Icebergs in Collaborative Design. *SIGOIS Bull.*, *17*(3), 24–33. https://doi.org/10.1145/242206.242212

Reicher, S. (2004). The Context of Social Identity: Domination, Resistance, and Change. *Political Psychology*, *25*(6), 921–945. https://doi.org/https://doi.org/10.1111/j.1467-9221.2004.00403.x

Repo, J. (2015). *The Biopolitics of Gender*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190256913.001.0001

Ricanek, K., & Tesafaye, T. (2006). MORPH: A longitudinal image database of normal adult age-progression. *FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, *2006*, 341–345. https://doi.org/10.1109/FGR.2006.78

Richardson, R., Schultz, J., & Crawford, K. (2019). Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *N.Y.U. L. Rev. Online*, *15*, 15–55. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423

Risman, B. J. (2018). Gender as a Social Structure. In *Handbook of the Sociology of Gender* (pp. 19–43). Springer, Cham. https://doi.org/10.1007/978-3-319-76333-0_2/FIGURES/1

Roberto, K. R. (2011). Inflexible bodies: Metadata for transgender identities. *Journal of Information Ethics*, *20*(2), 56–64. https://doi.org/10.3172/JIE.20.2.56

Robitzki, D. (2019). Japanese Taxis Are Using Facial Recognition to Target Ads to Riders. *Futurism*. https://futurism.com/japanese-taxis-facial-recognition-target-ads-riders

Rode, J. A. (2011). Reflexivity in digital anthropology. *Conference on Human Factors in Computing Systems - Proceedings*, 123–132. https://doi.org/10.1145/1978942.1978961

Rodriguez, C. E. (2001). Changing Race: Latinos, the Census, and the History of Ethnicity in the United States. *The Journal of American History*, *88*(2), 744. https://doi.org/10.2307/2675257

Rodríguez, P., Cucurull, G., Gonfaus, J. M., Roca, F. X., & Gonzàlez, J. (2017). Age and gender recognition in the wild with deep attention. *Pattern Recognition*, *72*, 563–571. https://doi.org/10.1016/J.PATCOG.2017.06.028

Roh, Y., Heo, G., & Whang, S. E. (2019). A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/tkde.2019.2946162

Roldan, W., Gao, X., Hishikawa, A. M., Ku, T., Li, Z., Zhang, E., Froehlich, J. E., & Yip, J. (2020). Opportunities and Challenges in Involving Users in Project-Based HCI Education. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3313831.3376530

Rolin, K. (2006). The Bias Paradox in Feminist Standpoint Epistemology. *Episteme*, *3*(1–2), 125–136. https://doi.org/DOI: 10.3366/epi.2006.3.1-2.125

Rolin, K. (2009). Standpoint Theory as a Methodology for the Study of Power Relations. *Hypatia*, *24*(4), 218–226. https://doi.org/10.1111/j.1527-2001.2009.01070.x

Rose, G. (1997). Situating knowledges: Positionality, reflexivities and other tactics. *Progress in Human Geography*, *21*(3), 305–320. https://doi.org/10.1191/030913297673302122

Rose, J. (2019). *I'm a trans woman – here's why algorithms scare me | Dazed*. Dazed. https://www.dazeddigital.com/science-tech/article/43211/1/trans-algorithm-machine-learning-bias-discrimination-chelsea-manning-edit

Roth, L. (2009). Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity. *Canadian Journal of Communication*, *34*(1). https://doi.org/10.22230/cjc.2009v34n1a2196

Rothe, R., Timofte, R., & Van Gool, L. (2018). Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. *International Journal of Computer Vision*, *126*(2–4), 144–157. https://doi.org/10.1007/s11263-016-0940-3

Rubin, G. (2013). Literary theory: an anthology. In J. Rivkin & M. Ryan (Eds.), *Choice Reviews Online* (Vol. 35, Issue 10, pp. 35-5478-35–5478). https://doi.org/10.5860/choice.35-5478

Ruyter, A. de, Brown, M., & Burgess, J. (2018). GIG WORK AND THE FOURTH INDUSTRIAL REVOLUTION: CONCEPTUAL AND REGULATORY CHALLENGES. *Journal of International Affairs*, *72*(1). https://www.jstor.org/stable/26588341?casa_token=UPCRbL9O-BYAAAAA%3AtIhfIuCCgcYW2FS8Ti0xS009J42nKrx_dOVITqIzrU3LFmWXgysUN8TutCuVFC3I-aiFepvPBgJt0fg9nSi54-zuftEiDaU2FaMEKal_eauJLsKd5g&seq=1

Salter, A. J., & Martin, B. R. (2001). The economic benefits of publicly funded basic research: A critical review. *Research Policy*, *30*(3), 509–532. https://doi.org/10.1016/S0048-7333(00)00091-3

Sambasivan, N., Kapania, S., & Highfill, H. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445518

Sambasivan, N., & Veeraraghavan, R. (2022). The Deskilling of Domain Expertise in AI Development. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3517578

Santarcangelo, V., Farinella, G. M., & Battiato, S. (2015). Gender recognition: Methods, datasets and results. *2015 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2015*, 1–6. https://doi.org/10.1109/ICMEW.2015.7169756

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., & Smith, N. A. (2022). Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5884–5906. https://doi.org/10.18653/v1/2022.naacl-main.431

Savage, R. (2019). Nonbinary? Intersex? 11 U.S. states issuing third gender IDs. *Reuters*. https://www.reuters.com/article/us-us-lgbt-lawmaking/nonbinary-intersex-11-us-states-issuing-third-gender-ids-idUSKCN1PP2N7

Scheuerman, M. K., Branham, S. M., & Hamidi, F. (2018). 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*, 29.

Scheuerman, M. K., Pape, M., & Hanna, A. (2021). Auto-essentialization: Gender in automated facial analysis as extended colonial project: *Https://Doi.Org/10.1177/20539517211053712*, *8*(2), 205395172110537. https://doi.org/10.1177/20539517211053712

Scheuerman, M. K., Paul, J. M., & Brubaker, J. R. (2019). *How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services*. *144*, 33. https://doi.org/10.1145/3359246

Scheuerman, M. K., Spiel, K., Haimson, O., Hamidi, F., & Branham, S. M. (2019). *HCI Guidelines for Gender Equity and Inclusivity*. https://www.morgan-klaus.com/gender-guidelines.html

Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proc. ACM Hum.-Comput. Interact.*, *4*(CSCW1).

Schmelzer, E. (2018). *Colorado to allow use of X as sex identifier on driver's licenses starting this month*. The Denver Post.

Schrupp, L. (2019). *Why We Created a Gender-Inclusive Stock Photo Library*. Broadly. https://broadly.vice.com/en_us/article/qvyq8p/transgender-non-binary-stock-photos-gender-spectrum-collection

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc.

https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

Seidelin, C., Dittrich, Y., & Grönvall, E. (2018). Data Work in a Knowledge-Broker Organisation: How Cross-Organisational Data Maintenance Shapes Human Data Interactions. *Proceedings of the 32nd International BCS Human Computer Interaction Conference*. https://doi.org/10.14236/ewic/HCI2018.14

Sen, S., Giesel, M. E., Gold, R., Hillmann, B., Lesicko, M., Naden, S., Russell, J., Wang, Z. K., & Hecht, B. (2015). Turkers, Scholars, "arafat" and "peace": Cultural communities and algorithmic gold standards. *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, 826–838. https://doi.org/10.1145/2675133.2675285

Serano, J. (2017). Transgender People and "Biological Sex" Myths. *Medium*.

Serrant-Green, L. (2002). Black on black: methodological issues for black researchers working in minority ethnic communities. In *Nurse researcher* (Vol. 9, Issue 4, pp. 30–44). https://doi.org/10.7748/nr2002.07.9.4.30.c6196

Setty, S., Husain, M., Beham, P., Gudavalli, J., Kandasamy, M., Vaddi, R., Hemadri, V., Karure, J. C., Raju, R., Rajan, B., Kumar, V., & Jawahar, C. V. (2013). Indian Movie Face Database: A benchmark for face recognition under wide variations. *2013 4th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, NCVPRIPG 2013*. https://doi.org/10.1109/NCVPRIPG.2013.6776225

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World*. http://arxiv.org/abs/1711.08536

Shapiro, A. (2017). Between autonomy and control: Strategies of arbitrage in the "on-demand" economy: *Https://Doi.Org/10.1177/1461444817738236*, *20*(8), 2954–2971. https://doi.org/10.1177/1461444817738236

Sharma, R., Moon, H., & Jung, N. (2007). *Automatic detection and aggregation of demographics and behavior of people* (Patent No. US8351647B2). https://patents.google.com/patent/US8351647B2/en

Sheth, F. A. (2009). The Technology of Race and the Logics of Exclusion: The Unruly, Naturalization, and Violence. In *Toward a Political Philosophy of Race* (pp. 21–35). SUNY Press. https://doi.org/10.5840/soctheorpract201036439

Sim, T., Baker, S., & Bsat, M. (2002). The CMU Pose, Illumination, and Expression (PIE) database. *Proceedings - 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR 2002*, 53–58. https://doi.org/10.1109/AFGR.2002.1004130

Simonite, T. (2021). What Really Happened When Google Ousted Timnit Gebru. *WIRED Magazine*, 20–39. https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/

Simpson, E., & Semaan, B. (2021). For You, or For"You"? Everyday LGBTQ+ Encounters with TikTok. *Proc. ACM Hum.-Comput. Interact.*, *4*(CSCW3). https://doi.org/10.1145/3432951

Singer, P. A. (2020). *Colonialism, Two-Spirit Identity, and the Logics of White Supremacy*. 1–7. https://www.academia.edu/2259929/Colonialism_Two_Spirit_Identity_and_the_Logics_of_White_Supremacy

Slota, S. C., Fleischmann, K. R., Greenberg, S., Verma, N., Cummings, B., Li, L., & Shenefiel, C. (2020). Good systems, bad data?: Interpretations of AI hype and failures. *Proceedings of the Association for Information Science and Technology*, *57*(1). https://doi.org/10.1002/pra2.275

Smith, M. M. (2006). *How Race Is Made: Slavery, Segregation, and the Senses*. https://doi.org/10.2307/25094613

Smith, R. (2001). Measuring the social impact of research. *BMJ*, *323*(7312), 528–528. https://doi.org/10.1136/bmj.323.7312.528

Snow, J. (2018). *Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots*. ACLU.

Sopelsa, B. (2018). *Gender "X": New York City to add third gender option to birth certificates*. NBC News. https://www.nbcnews.com/feature/nbc-out/gender-x-new-york-city-add-third-gender-option-birth-n909021

Soper, T. (2020). Retired UW computer science professor embroiled in Twitter spat over AI ethics and 'cancel culture.' *GeekWire*. https://www.geekwire.com/2020/retired-uw-computer-science-professor-embroiled-twitter-spat-ai-ethics-cancel-culture/

Stodden, V., Krafczyk, M. S., & Bhaskar, A. (2018). Enabling the Verification of Computational Results: An Empirical Evaluation of Computational Reproducibility. *Proceedings of the First International Workshop on Practical Reproducible Evaluation of Computer Systems*. https://doi.org/10.1145/3214239.3214242

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

Stoler, A. L. (2014). *Race and the Education of Desire*. Duke University Press. https://doi.org/10.1215/9780822377719

Stryker, S., Whittle, S., von Krafft-Ebing, R., Cauldwell, D. O., Haraway, D., Butler, J., Bornstein, K., & Halberstam, J. (2006). The Transgender Studies Reader. In *Journal of Chemical Information and Modeling* (Vol. 53). https://doi.org/10.1017/CBO9781107415324.004

Stuart Geiger, R., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 325–336. https://doi.org/10.1145/3351095.3372862

Su, N. M., Lazar, A., & Irani, L. (2021). Critical Affects: Tech Work Emotions Amidst the Techlash. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW1). https://doi.org/10.1145/3449253

Suchman, L. (1993). Do Categories Have Politics? The language/action perspective reconsidered. In *Proceedings of the Third European Conference on Computer-Supported Cooperative Work 13–17 September 1993, Milan, Italy ECSCW '93* (pp. 1–14). Springer Netherlands. https://doi.org/10.1007/978-94-011-2094-4_1

Suresh, H., & Guttag, J. V. (2019). *A Framework for Understanding Unintended Consequences of Machine Learning*. www.aaai.org

Szlávik, Z., & Szirányi, T. (2004). Face Analysis Using CNN-UM. *Proceedings IEEE International Workshop on Cellular Neural Networks and Their Applications (CNNA 2004)*, 190–195.

Taherdoost, H. (2018). Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3205035

Takezawa, Y. (2012). Problems with the Terms : "Caucasoid", "Mongoloid" and "Negroid. *ZINBUN*, *43*, 61–68. https://doi.org/10.14989/155688

Tan, X., Li, Y., Liu, J., & Jiang, L. (2010). Face liveness detection from a single image with sparse low rank bilinear discriminative model. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6316 LNCS*(PART 6), 504–517. https://doi.org/10.1007/978-3-642-15567-3_37

Tavanti, M., Marti, P., & Bourgois, M. (2006). Tacit Knowledge and Frugal Artifacts: A Challenge for Technology Mediated Collaboration. *Proceedings of the 13th Eurpoean Conference on Cognitive Ergonomics: Trust and Control in Complex Socio-Technical Systems*, 105–108. https://doi.org/10.1145/1274892.1274909

Taylor, J. L., Soro, A., Roe, P., Lee Hong, A., & Brereton, M. (2017). Situational When: Designing for Time Across Cultures. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6461–6474. https://doi.org/10.1145/3025453.3025936

Teo, T. (2014). Epistemological Violence. In *Encyclopedia of Critical Psychology* (pp. 593–596). Springer New York. https://doi.org/10.1007/978-1-4614-5583-7_441

Tommasi, T., Patricia, N., Caputo, B., & Tuytelaars, T. (2017). A deeper look at dataset bias. In *Advances in Computer Vision and Pattern Recognition* (Issue 9783319583464, pp. 37–55). Springer London. https://doi.org/10.1007/978-3-319-58347-1_2

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B. J., & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006

Towle, E. B. (2005). ROMANCING THE TRANSGENDER NATIVE: Rethinking the Use of the "Third Gender" Concept. *GLQ: A Journal of Lesbian and Gay Studies*, *8*(4), 469–497.

https://doi.org/10.1215/10642684-8-4-469

Tubaro, P., Casilli, A. A., & Coville, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data and Society*, *7*(1), 25. https://doi.org/10.1177/2053951720919776

Valentine, D. (2016). Imagining Transgender: An Ethnography of a Category. In *Journal of Anthropological Research* (Vol. 65, Issue 3, pp. 516–517). Duke University Press. https://doi.org/10.1086/jar.65.3.25608249

Valentino-DeVries, J. (2020). How the Police Use Facial Recognition, and Where It Falls Short. *The New York Times*. https://www.nytimes.com/2020/01/12/technology/facial-recognition-police.html

Valla, J. M., Ceci, S. J., & Williams, W. M. (2011). The accuracy of inferences about criminality based on facial appearance. *Journal of Social, Evolutionary, and Cultural Psychology*, *5*(1), 66–91. https://doi.org/10.1037/h0099274

Vallas, S., & Schor, J. B. (2020). What do platforms do? Understanding the gig economy. In *Annual Review of Sociology* (Vol. 46, pp. 273–294). Annual Reviews. https://doi.org/10.1146/annurev-soc-121919-054857

van Doorn, N. (2017). Platform labor: on the gendered and racialized exploitation of low-income service work in the 'on-demand' economy. *Information Communication and Society*, *20*(6), 898–914. https://doi.org/10.1080/1369118X.2017.1294194

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Veale, M., Kleek, M. Van, & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3173574

Veron, J., Preston, S. H., Heuveline, P., & Guillot, M. (2006). Demography: Measuring and Modeling Population Processes. *Population (French Edition)*, *57*(3), 591. https://doi.org/10.2307/1535065

Vertesi, J., & Dourish, P. (2011). The value of data: Considering the context of production in data economies. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 533–542. https://doi.org/10.1145/1958824.1958906

Vincent, B., & Manzano, A. (2017). History and Cultural Diversity. In *Genderqueer and Non-Binary Genders* (pp. 11–30). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-51053-2_2

Vincent, J. (2017, August 22). *Transgender YouTubers had their videos grabbed to train facial recognition software*. The Verge.

Vision API - Image Content Analysis | Cloud Vision API | Google Cloud. (2019). https://cloud.google.com/vision/

Wacquant, L. (2006). Habitus. In *International encyclopedia of economic sociology*. Routledge.

Walker, A. M., & Devito, M. A. (2020). "'More gay' fits in better": Intracommunity Power Dynamics and Harms in Online LGBTQ+ Spaces. *Conference on Human Factors in Computing Systems - Proceedings*, 1–15. https://doi.org/10.1145/3313831.3376497

Wallace, L. R. (2019). *The View from Somewhere: Undoing the Myth of Journalistic Objectivity*.

Wang, D., Prabhat, S., & Sambasivan, N. (2022, April 29). Whose AI Dream? In search of the aspiration in data annotation. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3491102.3502121

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., & Ordonez, V. (2019). Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wang, Y., & Kosinski, M. (2017). Deep Neural Networks Can Detect Sexual Orientation From Faces. *Journal of Personality and Social Psychology*, 1–47. https://doi.org/10.17605/OSF.IO/HV28A

Wang, Z., Huang, Z., & Luo, Y. (2020). Human Consensus-Oriented Image Captioning. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.

Watson Visual Recognition. (2019). https://www.ibm.com/watson/services/visual-recognition/

Wevers, R. (2018). Unmasking Biometrics ' Biases: Facing Gender, Race , Class and Ability in Biometric Data Collection. *Tijdschrift Voor Mediageschiedenis*, *21*(2), 89–105. https://doi.org/10.18146/tmg21368

Whitehill, J., & Movellan, J. R. (2008). Personalized facial attractiveness prediction. *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008*.

https://doi.org/10.1109/AFGR.2008.4813332

Wilkins, D. J., Meijer, M., Lindley, S., Hulikal Muralidhar, S., & Lascău, L. (2022). Gigified Knowledge Work: Understanding Knowledge Gaps When Knowledge Work and On-Demand Work Intersect. *Proceedings of the ACM on Human-Computer Interaction*, *6*, 28. https://doi.org/10.1145/3512940

Williams, B. A., Miceli, M., Gebru, T., & Williams, A. (2022). *The exploited labor behind artificial intelligence*. NOEMA. https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/

Williamson, V. (2016). On the Ethics of Crowdsourced Research. *PS: Political Science & Politics*, *49*(1), 77–81. https://doi.org/DOI: 10.1017/S104909651500116X

Wilson, O. (2013). *Violence and Mental Health in the Transgender Community*. *December*. https://search.proquest.com/docview/1647175438?pq-origsite=gscholar

Wilson, S. C., Morrison, S. D., Anzai, L., Massie, J. P., Poudrier, G., Motosko, C. C., & Hazen, A. (2018). Masculinizing Top Surgery: A Systematic Review of Techniques and Outcomes. In *Annals of Plastic Surgery* (Vol. 80, Issue 6, pp. 679–683). https://doi.org/10.1097/SAP.0000000000001354

Winner, L. (1980). Do Artifacts have Politics? *Daedalus*, *109*(1), 19–39. https://doi.org/10.2307/20024652

Winner, L. (1986). *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University of Chicago Press.

Wodda, A., & Panfil, V. (2015). “Don’t talk to me about deception”: The necessary erosion of the trans* panic defense. *Albany Law Review*, *78*(3), 927–971. https://doi.org/10.1017/CBO9781107415324.004

Wolf, C. T., Asad, M., & Dombrowski, L. S. (2022). Designing within Capitalism. *Designing Interactive Systems Conference*. https://doi.org/10.1145/3532106

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18)*, *2018-April*, 1–14. https://doi.org/10.1145/3173574.3174230

Woolgar, S., & Suchman, L. A. (1989). Plans and Situated Actions: The Problem of Human Machine Communication. *Contemporary Sociology*, *18*(3), 414. https://doi.org/10.2307/2073874

Wouters, N., Kelly, R., Velloso, E., Wolf, K., Ferdous, H. S., Newn, J., Joukhadar, Z., & Vetere, F. (2019). Biometric mirror: Exploring values and attitudes towards facial analysis and automated decision-making. *DIS 2019 - Proceedings of the 2019 ACM Designing Interactive Systems Conference*, 447–461. https://doi.org/10.1145/3322276.3322304

Wu, X., & Zhang, X. (2016). *Automated Inference on Criminality using Face Images*.

Wylie, A. (2003). Why standpoint matters. In R. Figueroa & S. G. Harding (Eds.), *Science and Other Cultures: Issues in Philosophies of Science and Technology*. Routledge.

Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). *Learning Face Representation from Scratch*. http://arxiv.org/abs/1411.7923

Young, I. M. (1990). Five Faces of Oppression. In *Justice and the politics of difference* (pp. 39–65). Princeton University Press.

Yuan, L. (2018). How Cheap Labor Drives China’s A.I. Ambitions. *New York Times*, 1–5. https://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html

Zdanowska, S., & Taylor, A. S. (2022). A study of UX practitioners roles in designing real-world, enterprise ML systems. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3491102.3517607

Zhang, Q., Elsweiler, D., & Trattner, C. (2020). Visual Cultural Biases in Food Classification. *Foods*, *9*(6). https://doi.org/10.3390/foods9060823

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2979–2989. https://doi.org/10.18653/v1/d17-1323

Zheng, Y. (2020). Digital Economies at Global Margins. *Economic Geography*, *96*(2), 190–192. https://doi.org/10.1080/00130095.2020.1721278

Zliobaite, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, *24*(2), 183–201. https://doi.org/10.1007/s10506-016-9182-5

APPENDIX

Appendix A: Traditional Tech Workers Semi-Structured Interview Protocol

Introduction

Hello, my name is Morgan Scheuerman. I'm a PhD student at University of Colorado, Boulder.

Thanks for taking the time to talk with me today. The purpose of this study is to dig into the role human identity characteristics play in technology, specifically machine learning applications. I know that is really vague and broad right now - we want to start that way, then dig into more specifics about this as we go.

Your participation in this conversation is strictly confidential - what we talk about stays between us. We will not identify you by name, team, or company and will anonymize all information for your protection. You also absolutely can decline to answer any of the questions I ask. You are also free to stop the interview at any time, you just have to let me know.

I just want you to know that there are no wrong answers to my questions. I want to hear your stories and your experiences. This is your interview.

- Do you have any questions about this interview before we start?

Before we move forward, do I have permission to record this interview?

Our scope of identity is pretty broad here. We are really interested in anything that could be important to people and their identities or senses of self, so demographics, practices, beliefs, cultures, and traditions.

Occupational Questions

- Can you briefly describe your job and what is your role within the company/business?

- Can you describe your role in developing computer vision algorithms?

Technical and Business Purposes

- What use cases do you develop computer vision algorithms for?

- Use cases
  - Where will computer vision algorithms be used?

    - Are there any ideal use-cases for computer viz? ("I'm sure your team has talked about a lot of different use cases for these algos. Is there one that you especially like? Your favorite? And can you tell me why?")
    - Are there any use cases you think it'd be useful for in the future?

- Users
    - Who are your end users?
        - How will these algorithms impact these end users?
    - How do you imagine these end users will benefit or make use of these algorithms?

People in Algorithms

Specifically interested in the way algorithm interacts with people…

- Can you describe how you describe classifying people in the context of your job or your users means to you?

- What does classifying people look like in a machine learning algorithms?

Inputs

Next, I want to talk specifically about inputs…

- What kind of features about people do you embed into the algorithms? For each, why? [Start open-ended.]
    - [Be certain to cover the following]
        - Race/Ethnicity
        - Age
        - Gender
        - Sexual Orientation
        - Other

- In the development of [your API/algo]...
    - Are there any features/characteristics you intentionally do not include?
    - Why?

- What do you consider when weighing the options?
    - What is your metric for success?
    - Accuracy?

- How do these identifiers interconnect?
    - Does this impact accuracy?

- Things that would be beneficial that…
    - Are technically hard to do?
    - Socially or politically hard to do?

- Do you ever work with third parties for data collection or annotation?

    - If so, how?

Outputs

So we have talked about inputs, let's talk about the outputs.

- Based on the above, what does a user see or understand about the characteristics in the algorithm?
    - Do they know the identifying characteristics?
    - Can they see them?
        - Input them?
        - Change them?
    - How does this impact the user experience?
- Is there any information the user cannot see or change?
- Is there any information the user is unaware is being collected / categorized about them?
    - What is the benefit of not knowing?

Ethics

- Have you ever had to make any decisions about incorporating/using these characteristics into a system you did not agree with?
    - How did you handle this?
    - What did you see the impact being?
    - If there were no barriers, how would you change this scenario in the future?

- Have you ever thought one of these categories should be implemented, but for some reason it could not be?
    - What was the reason?
    - How did you handle this?
    - What did you see the impact being?
    - If there were no barriers, how would you change this scenario in the future?

Changes

- Have you ever had to change anything in your product based on feedback from clients?
    - Inputs
    - Characteristics

- How would [your product] change if you added these specific characteristics?
    - What would the consequences of that be?
    - How would you build it?
    - What information could it provide?
    - How would this be useful?
    - Not useful?

Examples to use if need be:

- A resume parsing API that did not categorize gender
- A facial recognition software that did not categorize gender, age
- Tailored ad tools that do not categorize gender

- How would [your product] change if you removed one specific characteristic?
    - For example, gender...
    - What would the consequences of that be?
    - How would you build it?
    - What information could it provide?
    - How would this be useful?
    - Not useful?

- How would [your product] change if all identity categories were removed?
    - What would the consequences of that be?
    - How would you build it?
    - What information could it provide?
    - How would this be useful?
    - Not useful?
    - What about user input?

- On your team, what role do you play in shaping what these APIs/algos are what they should do?

Opinions and Values

Technical Futures

Based on your expertise and knowledge of this domain, I wanted to ask you what your thoughts were more broadly - outside the context of your specific product or company…

- Are there any use cases you think it'd be useful for in the future?
    - Gender recognition
    - Race recognition
    - Age recognition
    - Sexual orientation
    - Other characteristics - disability?
    - How should these categories NOT be used?

- Challenges and opportunities
    - Gender
        - What do you think are some the most interesting challenges or opportunities for G/R in the next 10 years?
        - How do you see the future of how G/R is used?
    - Race
        - What do you think are some the most interesting challenges or opportunities for R/R in the next 10 years?
        - How do you see the future of how R/R is used?
    - Age
        - What do you think are some the most interesting challenges or opportunities for A/R in the next 10 years?
        - How do you see the future of how A/R is used?

    - Sexual orientation
        - What do you think are some the most interesting challenges or opportunities for R/R in the next 10 years?
        - How do you see the future of how R/R is used?
    - Other characteristics - disability?

- Have you heard of any positive scenarios using identity categories?
- Have you heard of any negative use cases or disaster scenarios using identity categories?

Values/Personal Opinions

- Feelings about identity categories...
    - So Vision API can detect all kinds of things. How important do you think these ID chars are in relationship to the others?
    - Q2 -- Based on your expertise, what do you PERSONALLY feel are the most important benefits of ID chars in APIs like these?
    - Are there any possible negative outcomes that you personally care about?

- [Follow up, if needed. When possible, use the scenarios provided in previous responses.]
    - What impacts do you think identity classifiers might have on society?
        - Positive
        - Negative

- What do you think a mistake should look like?
    - How should it be handled by the system?
    - By the users?
    - How do you currently handle mistakes?
    - Can users report mistakes?

- Should end users be able to interact with algorithmic categories (the way their personal identity is characterized by a system)?
    - Why or why not?
    - How?

- What is a team's responsibility for how API is used by other people/systems?

Wrap Up

- Is there anything else you'd like to discuss?
- Would you like to be informed about the results of our research?
- Would you feel comfortable interviewing again as this research progresses?

Demographic Questions

- What is your age?
- What is your racial identity?
    - Clarification questions: Is this different from your ethnicity? How would you describe your race?
- What is your gender identity?

    - Clarification questions: How would you describe your gender?
- Where do you live?
- What is your current occupation?

Appendix B: Annotator Semi-Structured Interview Protocol

Hello, my name is Morgan Scheuerman. I'm a PhD student at University of Colorado Boulder. Thanks for taking the time to talk with me today. The purpose of this study is to dig into the role you play in shaping the data annotation used to power computer vision. I am doing this research to understand the importance of annotators' work in shaping ethical computer vision systems. As such, I will ask you about your opinions, beliefs, practices, and identity in relation to your work.

What is it like to participate in an interview study?

- Your participation in this conversation is strictly confidential - what we talk about stays between us.
- I will not identify you by name, team, or company and will anonymize all information for your protection.
- You only have to answer the questions you want to! You can decline to answer any of the questions I ask.
- You are also free to stop the interview at any time, you just have to let me know.
- [If you want to change your mind, or strike/delete/remove something you've said, that is okay too.]
- Most important: I just want you to know that there are no wrong answers to my questions. I want to hear your stories and your experiences. This is your interview.

What will be shared with my employer?

- Will responses be confidential from my employer?
    - Yes, no individual interviews will be shared with EnVision Data. Any information provided to EnVision Data will be aggregated (e.g., "Annotators found task X to be the most difficult…")
- What else will be shared with my employer?
    - The only other thing that will be shared with EnVision Data are high level findings from the study.

Do you have any questions about this interview before we start?

Types of Questions I Will Ask During the Interview

- I want to hear about you and your work and how you approach your work, such as what brought you to work for EnVision Data, what types of projects you work on, and what kinds of annotations you find easy/difficult
- Specific projects you are working on with EnVision Data, including training you undergo and the process of labeling images
- Examples of images and the step by step processes you go through

- Things that could help you do your job or make your job easier
- Personal experiences and perspectives that help you do your work

Interview Guide

- Tell me a bit about yourself…
    - What brought you to HITL?
    - What is your background like?
    - What work did you do prior to HITL?
    - What has your experience been like since you began at HITL?
    - What is the community like with fellow workers?

- At a high level, describe the work you do
    - Do you have a favorite type of project to work on?
    - Least favorite?
    - What are easy projects to work on?
    - What are difficult projects to work on?

Training

I'd like to talk to you more about the x project…

- Tell me a bit about the training you underwent for x
    - Can you walk me through how the training usually goes
    - Is training always the same for all projects
    - Who conducts the training

- What interfacing did you have with the client if any?

- What happens when there are confusions in training?
    - How do you get clarifications?

Annotating

- How do you go about annotating x?
    - Walk me through your decision making process

- How long would you say it takes you to annotate…
    - X
    - X
    - X

- What sort of edge cases came up in the project?
    - Can you show me an example?
    - How do you handle edge cases like that?
    - How confident do you feel dealing with such edge cases?

- What sorts of annotations do you find…

  - Subjective?
  - Objective?

- How confident are you in annotating…
  - X
  - X
  - X
  - What do you do in instances you don't feel confident?

- Have you ever had to annotate x in another project also?
  - Were the instructions the same or different?
  - If different…

- What kind of personal experiences do you rely on when making annotation decisions?
  - For example…

- X is based in X … Were there ever any culturally confusing things about the annotation process?
  - E.g., what a certain concept looks like is different in your culture than the client's
  - Did cultural difference get incorporated into training?

- What about language differences or challenges?

- How do you feel your perspective shapes your annotations about people?
  - What are the positive aspects?
  - Are there any difficulties?

Aftermath

- How is the quality of your work assessed?
  - How are you given feedback?

- What happens if you make a mistake on an annotation?
  - How often do you feel that happens?
  - How is it handled by HITL / your manager?

- Have you ever come into contact with AI like you work on in real life?

- Have you ever disagreed with the way a label like x was defined?

- Have you ever disagreed with a task or project?
  - Would there be any tasks or projects you would find offensive or uncomfortable to do?
  - Would you be able or willing to turn down such projects?

- Is there anything you think could help improve your work?
  - Anything in the work environment that could be changed
  - What makes x challenging?

HITL Questions

- HITL positions itself as a provider of ethical AI…
    - How do they incorporate that into training?

- HITL really values the diversity of its workforce in terms of gender and country…
    - What are your thoughts on that?
    - How do you feel you fit into that mission?

- Do you know anyone who worked here who quit?
    - Moved on?

Appendix C: EnVision Data Ethical AI Training Assessment

- What did you already know going into the course?
    - What did you learn?

- What did you find useful?
    - What did you find boring?

- What did you find easy?
    - What did you find hard?

- How did you feel about the point that annotators from diverse countries are useful for AI?
    - Why do you think they are?
    - Why would they not be?

- Why do you feel you need to know the aim of the client/project?
    - How is it helpful?

- How do you feel about the process of reporting confusions?

- How do you feel about the process of handling unsure edge cases?

- If you had to collect a dataset of people, what would you do?
    - Did you previously have this in mind?
    - Race
    - Gender
    - Age
    - Disability
    - Did you consider…
        - Scale
        - Pose
        - Occlusion
        - State

        - Origin
        - Illumination

- Do you feel you have enough guidance to collect a sufficiently diverse dataset?
    - Have you ever had feedback a dataset wasn’t diverse enough?
    - What challenges are there to collecting a diverse dataset?

- What would happen in a model if your labels were not consistent?
    - With team members?

- How would you define ethical AI after the course?
    - Did it change your perspective?

- What about bias?
    - How do you feel your own values and judgments affect your annotations?
        - Were you conscious of those?

- Did you personally agree with everything in the course?

- Was anything missing from the course?