Reflections on Algorithmic Reputation: Judgment and Equity in a Digitally Mediated Society

Michael A. Katell

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Adam D. Moore, Chair

Ryan Calo

Batya Friedman

Anna Lauren Hoffmann

Program Authorized to Offer Degree

The Information School

University of Washington

Abstract

Reflections on Algorithmic Reputation: Judgment and Equity in a Digitally Mediated Society

Michael A. Katell

Chair of the Supervisory Committee:

Adam D. Moore

The Information School

Contemporary lives are digitally-mediated lives. A significant amount of communication, knowledge-seeking, employment, politics, and transactions take place through devices and on platforms where they are captured, aggregated, and analyzed computationally. This work, often labeled "algorithmic profiling," is directed toward predicting what people will do and for calculating their potential worth and risk in many domains of life, from policing to employment to dating. Descriptive accounts of algorithmic profiling typically fail to signal how computational judgment and categorization are *socio-technical,* subject to human culture and politics. I employ *reputation* as a means of evaluating algorithmic profiling, redirecting attention from technical issues and constraints to its social, political, and economic features. I employ and modify the political philosophy of John Rawls to gain perspective on the moral content of algorithmic reputation. I analyze algorithmic reputation for its role in structuring governing institutions and constructing relations of power. I argue that the requirements of a just, equitable, and stable society include algorithmic reputation processes and practices that are transparent, accountable, and which demonstrate fundamental respect for persons.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I gratefully acknowledge the tremendous support and mentorship I received from my PhD advisor Adam Moore and "shadow advisor" Batya Friedman, whose years of guidance and engagement indelibly shaped my development as a scholar. I am also deeply humbled by the patient mentorship and dedication of Ryan Calo and Anna Lauren Hoffmann, whose perspectives and critiques broadened my horizons and challenged me to explore unfamiliar terrain. I was also guided, encouraged, and supported in uncountable ways by the faculty and fellow Ph.D. students (current and former) of the University of Washington Information School, including Norah Abokhodair, Negin Dahya, Megan Finn, Ricardo Gomez, David Hendry, Mary Hotchkiss, Wendy Elizabeth King, David Levy, Lassana Magassa, Amanda Menking, Elizabeth Mills, Sonali Mishra, Bryce Newell, Rose Paquet, Matthew Saxton, Annie Searle, Araba Sey, Jaime Snyder, Richard Sturman, Hazel Taylor, Nicholas Weber, and Jevin West. I am grateful for the significant moral support, advice, and collaboration by my research colleagues in the Critical Platform Studies Group: Peaks Krafft and Meg Young. I owe a special debt of gratitude to Francien Dechesne, Bert-Jaap Koops, and Jens-Erik Mai for welcoming me into their scholarly communities. I benefitted tremendously from my association with the University of Washington eScience Institute, the Tech Policy Lab, and the Value Sensitive Design Lab. I am also grateful to the dedicated staff of the Information School for their many years of assistance in overcoming challenges large and small, and also to my generous and meticulous copyeditor, Cindy Fester.

I am also grateful for the ongoing love and support of my family and friends, including my sister, Geraldine Copitch, and members of the Ballantine, Bowen, Cahn, Lawrence, and Leyrer families, as well as John Adair, Mia Boyle, Matthew and Leslie Dresdner, Jed Dunkerley, Kathleen Hall, Sonia Honeydew, Daniel Thornton, Jonathan Radosevich, Yvette Iribe Ramirez, Susan Wilk, and all my fellow 'Strugglers.'

DEDICATION

This dissertation is dedicated to my late wife and best friend Sarah Leyrer (1978-2020).

May her commitment to justice and equity for the oppressed, and the love, enthusiasm, and generosity she shared with so many, be an inspiration to all who knew her.

CHAPTER 1 — ALGORITHMIC REPUTATION AND DIGITALLY MEDIATED LIFE

1.1 INTRODUCTION

Billions of people live digitally mediated lives in which they are routinely rated, scored, and categorized by computational means. Beginning with the introduction of the browser cookie and proceeding through the development of an array of tracking and inference technologies, profiling practices have evolved to capture a significant share of human behaviors and choices. This behavioral visibility is instantiated in records that become the raw materials of digital profiling and prediction. Mobile apps, credit card purchases, online searches, electronic books, and participation in the on-demand economy are just a few sites of computational rating and scoring. The purchase of a plane ticket is an opportunity to identify either a thrifty or extravagant traveler for advertising partners. Riding in a Lyft is an opportunity to rate the drive and be rated in return. Being rude to a service representative on the phone may lead to longer hold times on future calls. Having a social media connection to a target of law enforcement may increase the likelihood of becoming one. In addition to the entities with whom we directly transact or communicate, an industry of data brokers evaluates the data collected from our connected lives and uses it to categorize us for various interventions. Assessing and evaluating people for worth or risk is nothing new, but information technologies have expanded the reach and frequency of this work, turning any digitally mediated activity into an opportunity be characterized and anticipated. Which activities are noticed and what kinds of assessments are made reflect the interests and worldviews of the observers and the audiences for whom the assessments are produced.

It is likely that humans have been characterizing other humans for others for as long as there has been anything called a "society." *Reputation* describes the many processes of forming and

sharing characterizations of people, typically with the intention of assessing them for risk or opportunity, and for influencing how others interact with them. Each of us has as many reputational profiles as social contexts in which we operate, and each captures a different dimension or perception of who we are. These profiles vary over time with changes in ourselves, within the conditions of our lives, and within the shifting scope and necessities of the contexts of interaction from which they are constructed. In digitally mediated lives, the scale and detail of information available to characterize and predict has expanded dramatically. So too has the domain of observers. Government and commercial actors, working toward their own and for mutual interests, acquire, interpret, and exchange personal details about people, often without the person who is being evaluated being aware of it.

No matter how acquired or processed, reputations are interpretive. They emerge from information passed through the sensing capabilities of people and their technologies and result in assessments that are sensitive to culture and moral commitments. The sources of the information used in reputational profiling are as varied as the affordances of observation and interpretation. Individual humans watch, listen, and tell stories. Organized entities like news organizations do something similar but at a greater scale and with more concrete objectives. At even greater scale and detail, digital information firms use their access to data to profile and assess anyone touched by networked technologies. It is this domain of activity—the work of digital information firms and their effects on the lives of users—that is the focus of this dissertation.

1.1.1 Algorithmic Reputation

When computational systems are employed in the work of acquiring, interpreting, and producing reputational information, I call this "algorithmic reputation." Algorithmic reputation is a variant reputation that involves similar mechanics of reputational judging and sharing but is done through

the affordances of information systems. It comprises systems of observation, interpretation, and production of information about people that produce assessments and characterizations. Algorithmic reputation is primarily a commercial practice with roots in targeted advertising—using the ability to observe or acquire information about people to predict what they might buy. Targeted advertising is a decades-old practice that began by using non-digital information—zip codes, warranty registrations, surveys, membership lists—to derive assumptions about interests and beliefs.[1] With the development of digital affordances to log, store, and share a significant portion of people's lives, a more granular version of targeted advertising has become possible. It is now the financial heart of the information economy. Modern data brokers remain interested in anticipating what people might buy (and nudging them toward buying choices) but have also pushed farther into a larger range of predictions about what people will do and who they are. Systems for evaluating potential employees, tenants, insureds, dates, and criminal defendants are examples of digital services that employ the power of information systems to manage people and intervene in their lives. On some accounts, this work is merely a business model; firms developing new ways of processing and packaging information for their customers. On others, this is a project of social control, with the intended outcome being both the anticipation and shaping of behavior to instantiate specific commercial and juridical objectives. However, it is facile to separate the two. Business practices reflect assumptions about the world including the nature and goodness of markets and the obligations people have, or do not have, to one another.

The affordances of algorithmic reputation take many forms. Among them are variations of data mining in which machine learning algorithms trawl through databases containing the various histories of mobile device locations, digital communications, information searches, media choices,

[1] Daniel Solove, *The Digital Person: Technology and Privacy in the Information Age* (New York University Press 2004).

online transactions, etc., to surface patterns that indicate interests, beliefs, and potential behaviors.[2] More recent technological affordances operate in real-time at the point of observation, in which some form of artificial intelligence processes sensor data to predict a subject's personality or state of mind, interpreting visual, audible, haptic and other physical cues to produce psychometric insights.[3] While the technologies involved are sophisticated and potentially inscrutable to lay observers, the work is generally a mimicry or amplification of similar work done by humans. For example, a human marketer could manually review a person's credit card transactions and make reasonably accurate inferences about the person's income, family composition, diet, leisure preferences, and so on. Similarly, human employers use the information from their senses while interviewing job candidates to infer aspects of a person's personality. What is crucially different is speed and scale. A single provider of a reputation product can access and interpret many thousands of data points about a person from numerous sources and produce algorithmically fueled judgments about them almost instantly. Another difference is the appearance of objectivity. Computational systems are presumed to be neutral and value-free, but upon deeper inspection, they are often revealed to reproduce the moral and political contexts of their makers and social moment. The combination of speed and scale with a belief in objectivity (and thereby infallibility) makes algorithmic reputation especially important for analysis by technologists, sociologists, legal scholars, and philosophers. The importance becomes more acute when such tools determine the economic or existential fates of human lives.

By labeling algorithmic pattern matching and interpretive sensing as a form of reputation, it

---

[2] Marion Fourcade and Kieran Healy, 'Classification Situations: Life-Chances in the Neoliberal Era' (2013) 38 Accounting, Organizations and Society 559; Solove (n 1).

[3] Luke Stark, 'Algorithmic Psychometrics and the Scalable Subject' (2018) 48 Social Studies of Science 204; Jon-Mark Sabel, 'How Artificial Intelligence Helps Humans Find the Best Talent [Infographic]' (*HireVue*, 5 October 2017) <https://www.hirevue.com/blog/how-artificial-intelligence-helps-humans-find-the-best-talent> accessed 4 December 2018.

is my purpose to make the link between human and machine activities explicit. I aim to tie together a familiar, longstanding, and under-theorized human practice with similar activities conducted with digital tools. There is a spectrum of reputational work that ranges from fully human at one end to fully computational at another. For example, deep learning algorithms can sort through millions of data points about people and identify patterns of similarity without direct human intervention. However, there are numerous hybrid activities in which humans choose categories and seek out information that places people within them. Sitting between these approaches is the use of humans to label information, such as images, to train algorithms to do similar work at scale.

1.1.2 Artifacts Have Politics

Algorithmic processes may appear to be neutral, simply extracting "what is there" and using the power of statistical modeling to draw correlations, but it is arguably true that even seemingly neutral computational processes are more nuanced than they may appear. The political scientist Langdon Winner, whose ground-breaking analyses of private and public sector technologies are widely studied by design and science and technology studies scholars, famously argues that "artifacts have politics." Employing examples including a New York City highway designed to impede the bus-bound inner-city poor (particularly people of color) from easily reaching Jones Beach, and the costly and counterproductive mechanization of a Chicago factory to disrupt labor organizing,[4] Winner argues that a thorough analysis of design begins with an examination of the context in which it occurs. Writing specifically about digital technologies, the technology historian Melvin Kranzberg[5] argues that the technological artifacts that reach mainstream adoption are shaped not only by good ideas and engineering talent but also by the contexts in which design,

[4] Langdon Winner, 'Do Artifacts Have Politics?' (1980) 109 Daedalus 121.
[5] 'Technology and History: "Kranzberg's Laws"' (1986) 27 Technology and Culture 544.

production, and implementation take place. Human politics and confluences of history enter into decisions about what to design, how to design it, and who ultimately adopts the result. Unpredictable markets, corporate struggles, world events, and public policy may result in the "best" technologies being overwhelmed by competitors. As data ethicist and critical technology scholar Anna Hoffmann argues, algorithmic systems "are generated by and through existing tools, methods, and practices and, further, are framed by the political and economic contexts out of which they emerge. Rather than transcending the material and the political, big data is firmly mired in the people and tools that make it possible"[6]. These perspectives jointly suggest that technologies are not inevitable, they are expressions of human desires and perspectives. Technology participates in these expressions, shaping people's possibilities and perspectives. Furthermore, the distribution of benefits and harms of a technology can be similarly traced back to the context and conditions in which it emerges. A move to "modernize" land ownership records in India that disenfranchised lower-caste landholders is more fully understood alongside an appreciation of Indian caste politics and historical discrimination.[7]

This framing is reflected in the conceptualization of technical artifacts as *sociotechnical*. The work of Science and Technology Studies (STS) scholars, such as Bruno Latour and Tarleton Gillespie are instructive here. Latour argues that technological artifacts, rather than being inert and stable, are actors in the drama of human lives.[8] On this view, technologies negotiate with people and institutions, participating in their choices and options. Gillespie extends this view by illustrating ways that society co-constructs the sociotechnical landscape, where technologies are

[6] Anna Lauren Hoffmann, 'Making Data Valuable: Political, Economic, and Conceptual Bases of Big Data' (2018) 31 Philosophy & Technology 209, 210.

[7] Jeffrey Alan Johnson, 'From Open Data to Information Justice' (2014) 16 Ethics and Information Technology 263.

[8] Bruno Latour, 'Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts' in Bijker, Wiebe E and John Law (eds), *Shaping Technology/Building Society: Studies in Sociotechnical Change* (MIT Press 1992).

reflections and instantiate human wills. Rather than providing a bridge from human frailty to perfection, technology merely extends what is already present. As he states: "technology is society rendered durable,"[9] meaning that technology is an expression of society rather than merely an external solution to its complex problems. Given a sociotechnical understanding of algorithmic systems, perhaps then it should not be surprising when critical and feminist scholars identify historical patterns of discrimination finding new expressions in algorithmic processes. A popular internet search platform appears at first glance to be merely an *information* service (qua neutral), but as African-American and Ethnic Studies scholar Safiya Noble argues, Google Search is an *advertising* service; a business product intended to serve specific business needs and audiences.[10] Noble argues that Google search results are racialized, frequently depicting members of non-dominant social groups in a degrading manner not simply due to the types of information available about them but also because of the design of the system to suit particular business logics.

Following these authors, I argue that algorithmic reputations are *designed* reputations, serving particular needs and particular audiences, casting doubt on presumptions of their neutrality. In addition to the ways an intended audience or customer shapes a system, algorithmic systems frequently rely on data whose production is shaped by human politics and history. Every collection of data and the models trained from it are constructed for particular reasons. The source data that flows into reputational algorithms is frequently political. For example, there is evidence to suggest that in several countries non-White people are more often convicted of "nuisance crimes," such as drug possession, than Whites despite similar incidence of drug dealing and use across demographic

[9] Tarleton Gillespie, *Wired Shut: Copyright and the Shape of Digital Culture* (MIT Press 2007) 77.

[10] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press 2018).

differences.[11] This provides for claims that criminal justice is racialized; suspects are targeted, at least in part, based on their race. Disproportionate treatment in life scenarios for members of some groups influences the shape and content of record-keeping about people that become fodder for algorithmic analysis. When data shaped by human politics is used for ad inputs for algorithmic systems, the outputs of those systems reflect a pattern of prior, politicized decisions, including decisions that disfavor members of marginalized race, gender, and other groups.

1.1.3 What is at Stake

What is interesting to me, and which motivates this project, is the moral dimension of this work, by which I mean the normative features of reputational judging and opportunity-shaping. Given a normative perspective on algorithmic reputation, we might approach its use with greater clarity and care. Reputation constructs the identity of the person in the eyes of others and therefore plays a role in important decisions. This includes decisions of extraordinary importance, such as who should have access to gainful employment and whether someone should go to jail. In practice, the components of reputation are rarely, if ever, believably objective. Choices about what to observe, how to interpret what has been observed, and how to tell others about the observation are based in specific views of the world. The move to make reputation computational is a move to strip away the visible evidence of its subjective content. While we might agree that there are many forms of reputational judgment—algorithmic or otherwise—that meet some context-specific criteria of acceptability, we can only do so when the bases of the assessment are visible and auditable in some fashion. A reputational assessment is only as trustworthy as the entity who issues it. The glamorous

[11] Bryan Warde, 'Black Male Disproportionality in the Criminal Justice Systems of the USA, Canada, and England: A Comparative Analysis of Incarceration' (2013) 17 Journal of African American Studies 461.

promise of using algorithmic systems to perfect such processes is only that—a promise. Its fulfillment should not be assumed.

While there may be rough consensus about the legitimacy of employing particular views and priorities in important decisions, they remain interesting targets for moral evaluation. Why this standard and not that one? Who gains most when this fact is considered, or considered in this particular way, while another is ignored? Reputation is a squishy process. It has uncertain effects on its domain of influence. By associating this complicated process of observing, synthesizing, and producing judgments with digital practices, I attempt to pull such practices into view as very human and therefore subject to our human politics, whether or not they appear as human or algorithmic.

1.2 Reputation and Political Philosophy

I focus on the determination of human reputations through consumer and citizen profiling technologies as a matter that unites political philosophy and information ethics. We live in the era of risk. Calculating and controlling for risk occupies a central place in the politics and business of the contemporary societies of the industrialized nations. Governments concentrate on national security and law enforcement while business risk discourse primarily revolves around maximizing efficiency, boosting profitability, and protecting intellectual goods. At the center of the risk paradigm are human subjects, whose inscrutability and variability produce uncertainty and untapped potential in the era of risk. Simultaneously, the extraction of maximum value from these subjects as an economic practice has become a touchstone for a significant share of contemporary capitalist ventures. The business model of generating revenue from tracking and predicting human behavior has been characterized as “extractive” and exploitative, and assigned the evocative label

"surveillance capitalism."[12] Whether the business model is precisely targeting individuals for marketing or juridical interventions or producing insights for uptake by others to do similar work, the contemporary information industry appears focused on commodifying human experience. Some business models may also seek to *orchestrate* that experience to suit specific objectives. There are efforts offered as good faith exercises in paternalism or based in a firm belief in societal benefits of economic activity, even when such activity includes nudging or manipulating people to make particular choices. Other efforts are more cynical; these are clearly manipulative and focused on exploiting people for their financial or other value and channeling it into a few, well-placed hands. Still others occupy some middle ground, where data practices may arguably offer a mix of harms and benefits. What would be helpful in analyzing these practices is a framework that sets aside common 'economic benefit' and 'innovation' narratives to focus attention on matters of justice and human well-being. Where information systems or practices threaten basic liberties and collective goods, I argue that we need to examine them closely and propose systems to hold them, and those who adopt them, accountable.

In this dissertation, I argue that forms of computationally enhanced profiling are a novel form of reputation. Having evolved well-beyond its roots in targeted advertising, automated profiling is used to evaluate worth and risk in decisions concerning employment, housing, dating, criminal justice, and many other domains of life. Reputation is similarly a process for making decisions about people by assessing them for risk or opportunity. Conceiving of automated profiling as a digital expression of reputation positions us to hold these processes up to the light and see beyond their presentation as an engineering practice based in infallible math to its *sociotechnical* nature as a set of artifacts and practices freighted with human desires and shaped by power. Assessing people

[12] Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (First edition, PublicAffairs 2018).

for risk and opportunity has implications for their basic justice interests. Additionally, deciding what qualities and categories matter most create normative forces that transmit value commitments and shape the structure and ordering of human society. Because of this, I argue that the benefits and harms of algorithmic reputation and the distribution of those risks deserves close attention and moral evaluation.

I argue that we should interrogate systems of algorithmic reputation to surface the logics of the people and institutions that build them and the audiences for whom they are produced. This evaluative work is a first step to holding the systems and persons that shape society and its member's lives fully accountable. My work contributes to an information and technology ethics based in a political justice that transcends engineering and market narratives. I open up reputation to examination, in both digital and non-digital variants, and conduct a case study in the employment sector to situate algorithmic reputation and related technologies as expressions of specific worldviews and frequently unverifiable claims.

1.3 ETHICAL STANCE

This dissertation is motivated by a moral argument. At its basis, it relies on ideal theory of John Rawls, which in turn is motivated by the philosophy of Immanuel Kant. While a thorough account of Kantian ethics is not undertaken here, I include a fundamental concept from his philosophy, which is that we are obligated to treat persons with basic humanity and a base level of dignity. This is perhaps best expressed by the second formulation of Kant's categorical imperative, which indicates that we have a duty to recognize all of humanity not merely as means to our ends but as

ends in themselves.[13] For both Kant and Rawls, persons (once defined) are accorded fundamental respect by virtue of their humanity.

I begin this inquiry with a set of assumptions. The first is that the abuse or exploitation of human beings by other humans and non-human entities like firms, states, or autonomous machines, is morally indefensible. By abuse I mean emotional and physical harm as well other acts that denigrate or diminish the basic dignity of a living being unless such acts are a commensurate and limited response to acts by the target (*e.g*. incarcerating a person who commits a violent act). By exploitation I mean the use of one agent's power, including economic, political, juridical, and social power, to extract something of value from a weaker party under conditions they would not voluntarily accept, given an opportunity to choose (*e.g*. denying a deserved good or benefit). Not all conditions of abuse or exploitation are the focus of this inquiry. I limit my concern to abusive and exploitive conditions that affect matters of basic justice. I employ the meaning of basic justice from Rawls who defines it as those features of society that govern the liberties, opportunities, and the ability to participate in society as an equal. Examples include the right to vote, the right to have one's religion tolerated, to be assured fair equality of opportunity, and the ability to hold property.[14] While there may be mild forms of abuse or exploitation that do not rise to a level of significant concern—such as convincing a reluctant family member to prepare dinner. Such trivial and commonplace events that do not diminish a person's fundamental dignity are unlikely to be matters of basic justice on this meaning. Locating the threshold of what counts as a matter of basic justice is one of the goals of this dissertation.

In making a claim about the moral status of exploitation, I rely on the work of John Rawls,

---

[13] Immanuel Kant, 'Groundwork for the Metaphysics of Morals,' *The Philosophy of Kant: Immanuel Kant's Moral and Political Writings* (Random House 1949). Kantian obligations motivate Rawls's philosophical stance, which I adopt here.

[14] John Rawls, *Political Liberalism* (Expanded ed, Columbia University Press 2005) 214.

who adapts Kant's commitment to the dignity and rationality of persons to place consent and choice at the center of his philosophy. As a liberal egalitarian, Rawls holds that all people are deserving of being treated with equal consideration rather than within tiers of greater or lesser consideration. This is not only a matter of what people deserve out of compassion but also for practicality sake; a society of equals is more likely to be a stable and prosperous society than an authoritarian order marked by unbreakable hierarchies that give rise to fierce resistance. Under egalitarian theory, people neither earn nor lose access to equal consideration; whether one's claims are accepted or rejected by others is subject to the reason and rationality that produces negotiated rules and customs that all can accept. This is not to suggest that criminals be treated like innocents, people of bad intent treated as trusted friends, that one's personal virtue is meaningless, etc. Rather, this view presumes that the fundamental interests of all persons are claims that must be considered, to be honored or rejected for justifiable reasons, rather than arbitrarily or capriciously. For a society's rules and structures to be just, they must be seen as *legitimate;* providing for both rational and reasonable bases for action. I will defend a view of legitimacy as being based in the conditions another person ought to accept for themselves and others if given a choice. So, a person who is taller than most other people should not be denied access to basic education simply due to her height. A society would have to identify a reason for such treatment that all could accept, including the person potentially denied a benefit. Otherwise, the legitimacy of this treatment is in question. Clearly it is a challenge for societies to construct their rules to suit such a standard and yet we see evidence of it in our own society. While it is certainly not universal, it is likely that even most petty thieves do not entirely reject rules against theft and drivers who routinely exceed the posted speed limit similarly endorse rules against speeding.

Exploitation is similarly irrational as a long-term strategy for social order because it

undermines a potential source of stability, the sense that the actions of one to another are legitimate and aim toward collective happiness. If we accept the premise that exploitation is morally unacceptable, then it follows that persons or systems that promote exploitation or that do little or nothing to prevent it are implicated in the same moral violation. Persons, practices, laws, customs, and the normative commitments of philosophers should therefore not excuse or create the conditions for exploitation, at a minimum. Wherever possible, they should actively work to defeat or weaken the conditions of exploitation. To this latter point, legal scholars have argued that a fundamental role for law, in particular, is to restrain power. The law and technology theorist Bert-Jaap Koops argues, for example that "in a lawless society power will reign supreme, while in a society of law, power is reined in by law."[15] On this view, the law is a compensating factor that recognizes the unequal status of employers vs. employees, police vs. the accused, and so on. Similarly, a role for moral philosophy is to instruct us in how to act toward others, first in general, and second, in light of the power relations that exist between differently situated persons. If I can easily rob you because of my physical strength or possession of a weapon, moral philosophy provides reasons why I am required to refrain from doing so. In keeping with this instructive idea of moral philosophy, I propose to offer instruction regarding algorithmic reputation to limit its participation and enablement of exploitation.

1.4 RESEARCH QUESTIONS

I attempt to answer the following research questions.

1.4.1 What is algorithmic reputation?

- How should reputation be understood in information ethics?
- Who are the participants in a reputational process and what are their roles?

[15] Bert-Jaap Koops, 'Law, Technology, and Shifting Power Relations' (2010) 25 Berkeley Technology Law Journal 973, 974.

- What features of reputation matter to information ethicists?

1.4.2 What is institutional reputation?

- What are the roots of institutional reputation in commerce and culture?
- How can we account for the authority of institutional knowledge over social organization?
- What is the connection between institutional reputation and algorithmic reputation?

1.4.3 What is the moral status of algorithmic reputation?

- How can we evaluate algorithmic reputation within a Rawlsian account of justice?
- Who is obligated to be open and accountable in matters of reputation and why?

1.5 STRUCTURE OF THE DISSERTATION

This dissertation is organized as follows. In Chapter 2, I provide an understanding of reputation as a process. Reputation is a concept that has received relatively little attention in the information ethics, human computer interaction (HCI), science and technology studies (STS), or adjacent literatures. Notable exceptions include a book-length treatment by Daniel Solove and Clay Shirky's anecdotal accounts.[16] While these accounts are valuable, they mainly focus on the features and effects of digital technologies in extending the reach of mob judgment intending to cause shame and embarrassment over transgressive behaviors. Algorithmic reputation can take the form of mob judgment and shaming, but I seek to unpack more of its features, such as the construction of identity and rational processes of accountability.

Like many features of the social and moral landscape, the mechanics and importance of digital identity and accountability become more salient with the emergence of technologies and sociotechnical practices that engage in reputational judging. In Chapter 2, I expand the exploration of reputation by considering various definitions that have been offered and settle on a preferred option. I also identify the key roles in processes of reputation, that of the *subjects* of reputational

[16] Daniel Solove, *The Future of Reputation: Gossip, Rumor, and Privacy on the Internet* (Yale University Press 2007); Clay Shirky, *Here Comes Everybody: The Power of Organizing without Organizations* (Penguin Press 2008).

judging whose behaviors and actions, and *assessors* whose observations and judgments produce the content of reputation. Also in this chapter, I unpack the process of reputation into three stages, collection, synthesis, and production, to describe aspects of reputation formation and the many choices involved. In addition, I consider the role of reputation in both transmitting and enforcing normative commitments, foreshadowing questions about what behaviors and whose assessments are most likely the subject of *algorithmic* reputation.

Having gestured here and in Chapter 2 toward a Rawlsian analysis of algorithmic reputation, in Chapter 3, I review several key concepts from John Rawls's two, interrelated theories of society, "justice as fairness" and "political liberalism." Justice as fairness is a detailed conception for the design of a society as a fair system of cooperation for mutual advantage. Rawls employs a thought experiment, the original position, in which we are to imagine a set of ideal conditions for the creation of a new society. In consideration of the institutions, structures, and rules imagined emerging from this conception, Rawls prompts us to reflect upon the societies we know and possible paths toward their improvement in light of the ideal. In his theory of political liberalism, Rawls considers the requirements for a democratic society to remain both fair and stable over time. Central to both of his major theories is Rawls's conception of the person. Persons are said to have two moral powers that motivate them to pursue their own interests and goals while mitigating and shaping them in light of the need to cooperate with others for mutual advantage. The extent to which persons regard each other as equals—deserving of reciprocal respect and consideration—demonstrates the degree of commitment to a just, democratic society.

This dynamic is further explored in political liberalism in which the coercive force of the social contract is understood to threaten the legitimacy, and thereby the stability, of democratic society. Reputation plays a role in both domains. First, reputation plays a functional role in systems

of voluntary cooperation by providing a means for persons who are not well-known to each other to be evaluated as partners in joint projects. Next, reputation provides role in social organization by providing a basis of mutual respect required for full participation in society. Finally, the normative force of reputation is enshrined in the basic structure of society which sets out the rights and obligations of citizens—what we expect from them—thereby constructing and shaping their civic identity. Here, a distinction begins to emerge between the arenas of social and associational life in which people may act capriciously based on personal beliefs and contexts in which moral commitments and standards of accountability are expected. While we may not hold reputational judging to especially high standards of accountability in casual interactions, we are required to employ a higher standard of openness and accountability within the deliberative domain Rawls calls the *political conception* of society, where questions of basic liberty and social organization are decided.

In Chapter 4, I revisit reputation to further investigate its role in the organization of society. I narrow the scope of interest to focus on reputational processes that affect the economic, political, or existential fate of the person. This closely mirrors the scope of deliberation in the "political conception" of society outlined by Rawls. Reputational judgment that affect a person's employment prospects or their likelihood of being imprisoned raise important moral questions. These decisions are typically the purview of established institutions, such as companies and courts, and other entities whose choices have a governing effect on people's lives. Reputations produced by such entities I label "institutional reputation" to reflect the actors who produce or consume reputational assessments and make decisions with that information. In this chapter, I conjure the roots of institutional reputation within common narratives that describe the development of modern society and the parallel emergence of long-distance commodities trading. Here, reputation

is understood as a process of building relationships beyond clan and kin. Institutional reputation appears to follow developments in, first, transportation, and later, information technologies that enabled economic activity at a distance. This moved market interactions and other communications from familiar social contexts to those involving largely unknown others. Standards and practices of evaluating others appear to have emerged to suit this development, leading to contemporary features of the information industry concerned with making assurances and mitigating the risks of entering into financial and other relationships with any number of persons near and far.

Institutional reputation is often assumed to be objective and standardized, and it has many aspects that meet or approach such characterizations. However, reputation is a process that is extremely sensitive to non-objective worldviews, including beliefs and commitments that have little consistency across cultures or other institutions. In this dissertation, I argue that institutional reputations are constrained and shaped to meet the interests of the institution and not necessarily contributive to the best interests of individuals and communities. The justice they provide is imperfect and culturally shaped. We see evidence of this when comparing laws between countries. The laws of one country may provide for same-sex marriage while the laws of another condemn homosexuals to death. While there may be wider social value to the decisional work of some institutions, we cannot assume that all institutional reputations are socially beneficial, and many may indeed be harmful to collective human flourishing.

In Chapter 5, I conduct a critical case study of an algorithmic reputation system employing aspects of the Value Sensitive Design methodological framework. I investigate the job candidate screening system produced and sold by HireVue, Inc. I evaluate HireVue's patents, marketing materials, public statements, and third-party commentary to surface the stakeholders affected by the HireVue system, its core technologies and features, and the company's worldview and

epistemic commitments. I develop an inventory of values revealed by HireVue technologies and associated narratives.

Following Luke Stark,[17] I find that HireVue engages in "algorithmic psychometrics," or the attempt to reveal a person's inner psychological state from physical and verbal cues. While there are serious questions about the accuracy and reliability of such analyses, particularly as it succeeds or fails for various social groups, attempts to reveal a person's inner-life raises significant privacy and autonomy concerns. Ultimately, I conclude that HireVue employs their technology to participate in the power relations that exist between societal members who already hold significant power (*e.g.* employers) and those who have far less (*e.g.* job applicants, particularly in times of economic precarity). HireVue very clearly targets its powerful technologies to the benefit of the former while attending little to the latter, thereby contributing to power relations of potential domination.

HireVue's key defense is that their technology "works," meaning that it provides job candidates to employers that are successful, as defined by the employer. HireVue also claims that their technologies are objective and overcome racial and other discrimination in hiring. Both claims are worthy of analysis. The first claim is vulnerable to criticism that when employers claim to find "successful" hires, they may be reproducing the discriminatory patterns of prior hiring practices. The second claim is offered without sufficient evidence by which we can fairly judge it. HireVue issues the findings of their own analyses without subjecting them to external audits. We are left to trust in the priorities and claims of employers and HireVue.

In Chapter 6, I conduct an analysis of algorithmic reputation—digital tools for decision-making—within a moral framework. I offer arguments about the moral status of algorithmic

[17] Stark (n 3).

reputation situated in the liberal political theory of John Rawls. First, I argue that reputation is an appropriate frame for understanding certain technical practices and artifacts. Next I offer three arguments from Rawlsian theory that indicate how we can fairly assess persons as rational and reasonable members of a cooperative society. First, I argue that reputation is a "primary good" closely tied to the good of "self-respect." Self-respect develops within individuals but depends also on recognition by others. A Kantian view of moral desert, adapted by Rawls, holds that there is a baseline of deep moral respect owed to all persons that exists prior to notions of what they deserve based on their acts, utterances, or other demonstrations of personal virtue. Rawls argues that in the absence of fundamental respect from others, the person cannot develop the "sense of justice" necessary to develop a sense of worth and membership in society. Only when a person is granted the opportunity to develop and then act on this sense of justice can we fairly judge the them reputationally. Next, I argue that reputation flows from the "basic structure" of society constructed by its participants to govern public life. Reputation is embedded in the basic structure by setting expectations for persons and then treating them in accordance with those expectations. The basic structure provides a set of public rules and obligations that indicate that "what he will be entitled to, and what a person is entitled to depends on what he does."[18] Third, I argue that algorithmic reputation is at risk of moral failure when it is both coercive and inscrutable. Following from and adaption of Rawls ideal of "public reason," I argue that while not all reputational standards and decisions are likely to be fully transparent in any context, there are life choices and opportunities that are so fundamental to human flourishing that they require an open and objective deliberative process to meet a baseline standard of justice. Ultimately, I close with a proposal for addressing problems of both justice and fairness for algorithmic reputation.

[18] John Rawls, *A Theory of Justice, Revised Edition* (Rev ed, Belknap Press of Harvard University Press 1999) 74.

Ultimately, this dissertation does two main things. First, it posits the notion of "algorithmic reputation," first by unpacking what reputation means in human society and then associating that meaning within computational processes of human assessment and evaluation. Then this dissertation places algorithmic reputation before us for a technical and moral analysis. As a technical artifact, algorithmic reputation contains and telegraphs a set of epistemic commitments and functional constraints that merit analysis for both fairness and accuracy. A moral analysis along Rawlsian lines indicates that algorithmic reputations may fail to confer basic moral respect for persons and lead to assessments and decisions that violate fundamental principles of justice and fairness.

CHAPTER 2 — REPUTATION AS AN INFORMATION PROCESS

2.1 INTRODUCTION

The purpose of this chapter is to describe the concept of reputation and to make it a legible concept for the arguments of this dissertation project. The definitions and descriptions offered here will be used to scaffold investigations into the evolution of reputation from an interpersonal social process grounded in collective action to formalized business systems of judgment and sorting that are increasingly mediated by digital information systems. First, in this chapter, I offer multiple definitions of reputation and ultimately settle on a preferred account. Each definition emphasizes something different about opportunities to subject reputation to moral evaluation. In so doing, I also discuss some related concepts, including cooperation, social performance, and trust. I follow by settling on a procedural account of reputation. Here, reputation is understood to be an evaluative process in which *subjects* are evaluated by *assessors.* To the extent they are able, subjects seek to influence how they are perceived. Assessors employ their conceptions of the world in choosing which signals about subjects to notice, how to think about them, and what to do with their assessments. Next, I explore reputation for its "normative force," in which I demonstrate how reputation is both a signal and enforcement mechanism of the prevailing moral and social norms that guide an assessor. Finally, I introduce the concept of *algorithmic reputation* as a variant of reputation in which subjects are assessed as a byproduct of digitally-mediated lives and the power of artificial intelligence technologies.

2.2 ORGANIZATION OF THIS CHAPTER

The chapter is organized as follows. In section 2.3, I offer a definition of reputation by considering three types of definition. Settling on a definition, I then offer an account of reputation as a process

through which reputational assessments are generated. While describing the process of reputation, I focus on the mechanics of collection, synthesis, and production of information that becomes a reputation. Next, I venture into categories of reputation to get at the degree of agency individual subjects have over their reputations and the degree of objectivity we can assert on them. Finally, in 2.9, I describe the "normative force" of reputation and its power to shape human behavior. Here, I argue that reputation shapes the actions and expressions of individuals as they work to meet the expectations of individuals, groups, and institution from which they desire acceptance or other benefits. This implicates certain questions about the application of governmental and non-governmental power to set the standards of reputation and thereby shape the behaviors of subjects. I foreshadow questions about the legitimacy of such power and the subsequent task of evaluating reputational standards and practitioners for their legitimacy and reasonableness. I close with a preliminary account of algorithmic reputation which will be revisited and expanded upon throughout this dissertation project.

2.3 DEFINING REPUTATION

In this section, I offer some definitions of reputation found in diverse literatures and then settle on a definition that provides a basis for the political and normative arguments in subsequent chapters. Prior to stating and discussing these definitions, I explain the roles that persons or entities play in the construction of reputation and then offer some foundational features of reputation before moving into various definitions and their relative strengths and weaknesses.

2.3.1 Roles

Reputation is generally a binary process involving two agents that I will call an *assessor* and a *subject.* The person or entity who performs the evaluative work is an assessor. A subject is the

agent about whom a reputation is produced. Assessors observe and/or acquire information about subjects and then do something with this information.

2.3.2 Individual vs. Entity

While reputation can flow from individual human beings to other human beings, between non-human entities, such as companies or government bodies, or between individuals and non-human entities, my scope is narrowly focused on reputation subjects who are human beings. I also focus on reputation subjects in conditions and relations in which the assessor is in a position of significantly greater economic, epistemic, or political power than the subject. So, while the reputation of a business owner who is assessed by an international trade group may be of interest here, the reputation of a wealthy CEO of a company who is assessed by a mail room worker is most likely not. I offer three reasons for choosing to narrow the scope in this way. One is efficiency; not all possible reputational relations can be accounted for without a lengthy treatment. Second, I choose to focus on the experience of human reputation subjects because there appears to be more fertile ground for investigating the normative status of judging them. While there may be normative questions about how we evaluate non-human entities in reputation settings,[19] the questions about individual human persons seem more urgent and acute. Third is the aspect of reputational stakes as a byproduct of power relations. While the fortunes of people or entities who hold positions of power can be affected by their reputations as assessed by relative weaklings, access to other forms of power make it easier for the powerful to neutralize or ride out unfavorable assessments, whereas those without access to other forms of power have more to lose in processes of reputation.

[19] An interesting, and possibly urgent non-human entity that merits normative inquiry is that of artificial entities (a.k.a. "robots"). See for example: David J Gunkel, *The Machine Question: Critical Perspectives on AI, Robots, and Ethics* (The MIT Press 2012).

2.3.3 Control

Subjects, even those with relatively little power, may have varying amounts of control over how they are judged. In some cases, a well-positioned or shrewd subject may have extraordinary control over how others see her. However, it is often the case that subjects have very little control or none at all. This makes intuitive sense when we consider that reputational assessments are the work of assessors. Assessors acquire information about subjects, consider the information, and then choose what to do with it. The degree to which a subject controls an assessment is the degree to which she controls the flow of information about herself. Controlling information to this degree is challenging for anyone and requires anticipating the attention and interests of her assessors. A person who desires to be seen as a "hard worker" in a typical U.S. firm might arrive early and stay late at her job, deliver high quality work to her supervisors, and strategically send emails to her coworkers during early or late work hours to signal her fulfillment of the highest expectations of the firm.[20]

While asserting control over reputation might be possible in certain contexts where the expectations are clear and the subject is capable of meeting them, control is particularly challenging under other circumstances, such as when reputations are attached to immutable characteristics over which the subject has no control. For example, assessors who make rash judgments based on one's skin hue, country of origin, or mental health status, and assign meanings to such characteristics, leave subjects with little chance of influencing those meanings. There are other aspects of the person that assessors may choose to privilege over others that may be unknown or incomprehensible to the subject, or for which the subject is incapable of responding. I discuss the influence of subjects on their reputations in more detail in section 0.

[20] The degree of control subjects have over their reputation is discussed in more detail in subsequent sections of this chapter.

2.4 DEFINITIONS

There appear to be three general approaches to defining reputation. As information object, as a performance, and as a process. Each manner of conceptualizing reputation exposes the concept to different types of analysis. The first two accounts of reputation are primarily descriptive, by which I mean they provide some facts about what reputation looks like without investigating whether what is observed ought to take place in the way it does. In contrast, a normative account may include descriptive elements while also delving into questions of the legitimacy and morality of how things are done. I present all three accounts and then discuss why we should favor the third procedural account, primarily because of its normative content.

2.4.1 Reputation as Object

Authors who define reputation in terms that render it an information object presents it as a fairly fixed set of information and its relationship to a subject. For example, the law and privacy scholar Daniel Solove defines reputation as "a shared, or collective perception about a person"[21]. This definition has some intuitive appeal because each of us experiences our own reputation and that of others as something stable and in terms of its effects. When I consider my own reputation, I may wonder if it matches how I wish to be perceived by others and what I can do about it if it does not. Another definition offered by social psychologists Cameron Anderson and Aiwa Shirako is "a set of beliefs, perceptions, and evaluations a community forms about one of its members."[22] This definition resembles Solove's but also brings in the social embeddedness of reputation, distributing attention between the subject and her location in social life. Author and reputation management entrepreneur Michael Fertik's definition is mainly focused on the experience of the subject but

[21] Solove (n 16) 11.

[22] Cameron Anderson and Aiwa Shirako, 'Are Individuals' Reputations Related to Their History of Behavior?' (2008) 94 Journal of Personality and Social Psychology 320. 320.

also brings in a transactional aspect: "Your reputation defines how people see you and what they will do for you."[23] Fertik and co-author David Thompson are concerned with reputation as either an asset or burden to a subject as she attempts to pursue her economic and social ends in a "reputation economy."[24]

Each of these definitions says something about the social nature of reputation, but the focus is primarily with the subject. While subjects are players in the dramas of their reputations in these accounts, the assessors are mainly out of view. When they are discussed, normative questions about what values they employ and the legitimacy of employing them are largely unexamined. For example, Fertik and Thompson describe a world in which the minute details of a person's life, including every financial transaction and their tone of voice on customer service phone calls, is tabulated into a subject's reputation. Yet Fertik and Thompson do not question whether or why we might question this type of judging; they focus instead on how the subject can maximize their perceived value in the eyes of their assessors. Such advice lacks an inquiry into questions about privacy, autonomy, justice, democracy, or any other values we might consider using to evaluate the legitimacy of the system. For reputation objectifiers, reputation is negotiated as, at best, a give-and-take between assessors and subjects. In the portrayal offered by Fertik & Thomson, assessors set most of the terms in relation to reputation but they themselves are out of view. In democratic societies, there is typically a requirement that systems of decision-making in matters of grave importance be open to accountability. Similarly, where being assessed with a good reputation means gaining access to a crucial good or the avoidance of a harsh sanction there can be a commensurate demand for accountability. Definitions of reputation that limit inquiry to the

[23] Michael Fertik and David C Thompson, *The Reputation Economy: How to Optimize Your Digital Footprint in a World Where Your Reputation Is Your Most Valuable Asset* (First edition, Crown Business 2015).
[24] ibid.

reputation itself and possibly to the subject are incomplete if they lack a similar inquiry into the details of the social frameworks that make the production of reputation—and its effects—possible.

2.4.2 Reputation as Performance

Another approach to defining reputation engages the subject more actively as a *performer* attempting to shape the impressions others adopt about her. Reputation in this conception is a performance in the drama of social interaction and human politics. The sociologist Erving Goffman argued that social life is performative; it is an ongoing process of settling the terms of mutual trust and expectation. "Information about the individual helps to define the situation, enabling others to know in advance what he will expect of them and what they may expect of him."[25] For Goffman, this exchange[26] can be understood as a collection of conscious and unconscious performances in which agents serve both performers and audience, producing and exchanging perceptions about where each person fits into a given social context.

Goffman's characterization captures the reflexive, self-conscious aspects of reputation formation for a person. In this conception, a subject becomes aware of her reputation and the types of performances that may lead to the reputational character she desires. She then can choose to shape her performances towards those ends. Depending on the type of role being considered for a reputation subject, she has a range of options in how to perform, from using signals like clothing and facial expressions, to calculated presentations, such as carefully chosen narratives about one's history or attitudes. Goffman describes a range of performances, such as spoken phrasing, physical behaviors, and the setting of background using the "props" of environment, furnishings, and other elements to support the performative narrative desired by the subject.

---

[25] Erving Goffman, *The Presentation of Self in Everyday Life* (Repr, Penguin 1990) 13.

[26] Reputation appears to lurk under the surface of Goffman's conception of social performances, but he does not employ the term.

Goffmanian performances are not only physical and temporal but may include both curated and involuntary/accidental trails of information artifacts, such as resumes, letters of reference, reports, background checks, etc. Particularly contemporary examples may include social media posts, music playlists, and consumer scores. Goffman acknowledges that performances are not always voluntary and curated by the subject. Assessors may seek documentary evidence of past behaviors or the impressions of associates who may offer a flattering account of the subject, or assessors may simply observe the subject without their awareness.

The performative account of reputation is more appealing than the reputation-as-object account because it provides a basis for understanding the subject's role in the construction of her reputation. On this view, the subject co-constructs her reputation with her audience of assessors. However, this conception overlooks the many functional and normative obstacles faced by subjects wishing to influence how they are evaluated and fails to account for the power of the audience whose conceptions of good, bad, right, and wrong play a larger role than Goffman allows. This unnecessarily limits our inquiry to the performer rather than paying equal attention to the work, perspective, and positionality of the assessor. There is little discussion in Goffman's account, for example, about the role of an assessor who harbors ill-intent toward the subject. While Goffman acknowledges that an audience may be unmoved even by an excellent performance, he does not pursue any normative questions about the values employed by the audience and whether those values ought to be viewed as legitimate or acceptable. As a result, the audience appears morally neutral rather than dynamic and directed by a broad array of interests and beliefs. I suggest that we need a more thorough, "end-to-end" account of reputation where we have more entry points into the normative spaces from which reputations emerge.

2.4.3 Reputation as Process

While the preceding definitions shed light on the experience of reputation and indicate how subjects may participate in their reputations, there is another account that offers more opportunities to consider how reputations are constructed against the backdrop of a society, including its values, prevailing rules, and institutions. I now turn to exploring reputation in procedural terms—how reputations are formed through processes of observation, synthesis, and production. In the procedural account, the focus is not on the reputation that is created so much as the process of creating it. Where other accounts focus mainly on the content of reputations and the role of the subject, the account I offer focuses primarily on the work of assessors. While I have indicated that subjects may influence how assessors respond to them, there are many situations in which assessors do most of the work of producing reputations. These include contexts in which the assessor is economically, epistemically, or politically powerful relative to the subject. Examples include relations between government agencies (powerful) and individual citizens (relatively weak) as well as those between some private institutions such as credit agencies (powerful) and individual consumers (relatively weak). The stakes in such relations of reputation can be extremely high for subjects whose access to income, liberty, powers, and opportunities may hang in the balance. A full account of the assessors in such domains should reveal their methods and the legitimacy and acceptability[27] of employing them in a free and fair society.

As a process, reputation is a system in which agents evaluate others in a collaborative and ongoing construction of meaning against the background conditions of a societal context and a set of values. Understood this way, reputational standards are constructed within and among groups

[27] I do not assume that assessments by powerful assessors are necessarily suspect or unacceptable. My meaning here is to suggest that we may have good reasons (*e.g.* justice, fairness, other values) to evaluate assessors rather than merely accept their authority without question.

following patterns of human relationships, culture, environment, history, and so on. Many features of reputational standards are influenced by individual and institutional viewpoints. This is not to say such standards are *arbitrary* but that the standards may not be inevitable.[28] Regardless of the process by which reputational standards are produced, a reputation is never solely about the person being assessed. While some reputations employ evidence and reliable measurements (*e.g.* "a fast runner"), there are degrees of interpretation as to the meaning of whatever facts/measurements lead to a reputational assessment. Reputations then are politicized judgments constructed to reflect myriad choices about what to notice, how to think about, and who to tell.

2.4.4 Who Decides

The aspects of reputation that reflect human politics extends to the distribution of particular reputation roles. In purely social situations, people occupy the role of assessor or subject interchangeably, potentially occupying both at the same time. For example, two people on a first date simultaneously judge and are judged by each other. Reputation can also be used to determine the distribution of what philosophers of distributive justice, such as John Rawls, label, "primary social goods,"[29] such as a desirable job, or the allocation of sanctions, such as jail sentence. Where the distribution of primary social goods and the allocation of sanctions are involved, the roles of reputation are foreordained. They may be constructed through the design of government or through the accretion of significant power to a private entity. Here, the roles in relations of reputation are stable, flowing in one-direction from assessor to subject. Judges deliberate and hand out jail sentences to defendants, not the other way around.[30] Societal structures that place people in relation

[28] I am not arguing here that reputations are necessarily capricious or based in nothing. What I argue is that reputation is not solely about facts. It is about which facts matter and what those facts mean in a particular context.

[29] John Rawls, *A Theory of Justice, Original Edition* (Original ed, Belknap Press 2005).

[30] A defendant may have some choice contributions to make to a judge's reputation, but they are unlikely to matter much to the judge's fortunes.

to each other as assessors and subjects reflect the distribution of power in a society. I mean to include all of the common forms of constructed power through which people assert their dominance over others. This includes economic power (wealth), social power (cultural and aesthetic preferences), juridical power (law, threat of force), and epistemic power (systems of knowledge).

2.5 THE MECHANICS OF REPUTATION

As described in the following sections, reputations occur as a process. The process occurs in stages as assessors acquire and process information into a reputational assessment. Reputation is a process consisting of phases. Each phase is characterized by a mechanism of action. The first mechanism is *collection*, in which some particular phenomena are observed, inferred, or acquired by the assessor; Next is *synthesis*, in which the assessor shapes and analyzes the collected information; Third is *production*, in which the assessor packages and conveys reputational information to others. Subjects passively emit or actively send signals through their interactions with the social world. Assessors make myriad choices about what to notice about subjects, what to think about it, and who to tell. Understanding and evaluating how those choices are made is the goal of this dissertation. In later sections, I closely examine the normative features of these decisions. In this chapter, I briefly discuss some of the salient aspects of reputation decisions for purposes of unpacking the mechanics of reputation and scaffolding normative questions for later discussions.

2.5.1 Collection

Subjects exist in the social world (including the *sociotechnical* world) where they passively and actively signal aspects of character. An assessor chooses, consciously or otherwise, which signals are important enough to notice and retain, failing to notice or actively disregarding others. Some

of the choices reflect the doctrinal commitments of the assessor, while other choices reflect the functional constraints of collection. Still others reflect active promotion by the subject. If the reputational information emerges from the performative signals of the subject, there will be limits to what the assessor can take in. First, there are challenges to sensing the subject. If reputational information emerges from direct observation, is the subject visible and audible? Is the subject legible to the assessor? Is the subject sending misleading signals?[31] Does the assessor have the benefit of repeated, sufficiently detailed experience with the subject? Is the assessor relying on information from others? Are they reliable? (What reputation precedes them?) The answers to these questions are among the contributors to the richness and reliability of the reputational input.

By example, consider a supervisor who notices one of her employees at the company holiday party. Being physically present and paying attention to the subject places the manager in an ideal position to render a judgment about the employee's interactions and other behaviors (*e.g.* "witty," "a bit of a lush," "shy," etc.). Yet the supervisor's ability to make an assessment is constrained by the limitations of her senses, attention, and memory. The employee may not be observable at all times. His conversations may take place out of listening range. A motivated and strategic observer will make optimization choices about how closely to observe a subject, but only if it is important enough to do so. Generally, the observer makes conscious and unconscious choices about where to concentrate their observational focus and will likely allow many details to go unnoticed. The employee may be masterful at giving off a particular impression that is beyond the skills of the supervisor to discern. The supervisor, who may herself have had a lot to drink at the party, also must have the ability and commitment to do the cognitive work of remembering what she has observed beyond the moment of observation.

[31] See section 2.6 for a fuller discussion of the role of the subject in reputational processes.

If the assessor cannot directly observe a subject or is otherwise constrained from getting enough information on her own, she may rely on a mix of third-party reports, including hearsay (gossip) or documentary evidence. Another employee may report to the supervisor what *she* saw at the party. This third-party information introduces additional functional constraints and optimization choices. The second employee may not themselves be particularly trustworthy, leading to a lower confidence level in the reputation information. Perhaps time has gone by and memories have faded. There may even be competing information from different sources, with multiple employees sharing wildly different accounts.

In addition to the functional constraints of gaining desired information for assessing a subject, the assessor brings her worldview to the act. A worldview shapes what an assessor thinks is important enough to notice and informs every act of assessment. The worldview may reflect one's own beliefs or be a projection of a belief system that governs an established context, such as a work setting. Before the supervisor at the holiday party takes note of an employee for being obnoxious, hilarious, polite, etc., she has to care (or feel a duty to care) about some aspect of character or behavior enough to pay attention to it. One could separately or simultaneously choose to focus on a person's tone of voice, facial expressions, or hairstyle. An assessor could focus on particular conversation topics while ignoring or forgetting others. Assuming the supervisor is capable of observing many aspects of the employee, she will only attend to those that matter to her.

2.5.2 Synthesis

Once something has been observed by an assessor, her next (or simultaneous) step is to convert what has been observed into an assessment. The assessor interprets what she has observed or acquired and considers what matters to her. This takes place through the lens of the assessor's

worldview and interests. There are many variables here. Among them are the purpose for making an observation and assessment in the first place. Assessing a subject as a prospective romantic partner provokes a different analytic process than assessing a subject as a candidate for political office. Both assessments may operate on the same information but with different emphasis because the interests being served are different. In the first case, the interests include intimacy or sex. In the second, the interests touch on one's vision for government. By example, an assessor may be attentive to the subject's apparent level of education or display of fashion for signals to guide her thinking. Using such signals, different assessors may see the same subject and same observed facts as alternately relatable or alienating, attractive or intimidating.

Beyond the purpose of the assessment, the assessor also brings numerous other priorities and beliefs—based on a combination of cultural context, social development, and personal psychology—to bear on the act of synthesizing information. Here again, a seemingly objective characteristic—apparent level of education—may have a varying impact on a reputational assessment. Someone who views themselves as "down to earth" may be inclined to view someone with an advanced degree as "elitist." Meanwhile, the same individual may be widely admired by her colleagues and students. Synthesis is the form of reputation most heavily informed by social or moral norms. Once an assessor decides to notice something and make a judgment about it, why it matters and the content of the conclusions is a function of her values, ranging from trivial, personal values to moral values that reflect commitments to conceptions of goodness, badness, right, and wrong.

2.5.3 Production

The third phase of the process of reputation is production, or the sharing of a reputational analysis with others. Reputations perform their social function when they emerge from the observation and

synthesis phases into the stories and labels we use to characterize a person. For any reputational assessment, an assessor produces a reputation to achieve some perceived end.

The production of reputation is also subject to constraints. An assessor is functionally constrained by the extent of her social reach. One person's impression can have very little authority or weight beyond one-to-one interactions. As argued by Anderson and Shirako, a reputation becomes more robust and impactful to the subject when it is shared among social networks in which there is rough agreement[32]. Rough agreement is not guaranteed, nor is a willing audience. Having one's views accepted by a social group depends on a range factors, including access to the network and an assessor's influence in that network. Some voices are likely to be heard more than others, a reflection of the position and power of the speaker. Here we encounter the possibility that the reputation of the assessor becomes part of the reputation process. An assessor who is presumed to be reliable, trustworthy, insightful, etc. is more likely to have her assessments taken up by others. An assessor who is not trusted or whose credibility is weak is less likely to be as impactful when assessing the reputations of others. Some assessors are granted credibility as a function of their position in society. For example, the elders of the Mormon Church are considered extremely credible among Mormons as a function of their authoritative role in the church. An individual elder gets a boost in his credibility simply through his position, regardless of his personal trustworthiness. So, in addition to the credibility that an assessor earns through her own performance, an assessor can benefit from her access to or association with the centers of power available in a given social context.

An assessor in a position to be heard by others chooses what to report and to whom. Here, we return to questions about what matters to the assessor and her relationship to her audience. When

[32] Anderson and Shirako (n 22).

reputations are produced, they are typically intended for some audience and that intention influences the product. A job reference is produced for a prospective employer. The producer of a reputation for an employer has to care, at least superficially, about what the employer values. If the employer is a fast-food franchise, the reference is unlikely to include the details of the subject's bawdy sense of humor and love of anime and more likely to include discussion of his efficiency, punctuality, and reliability. This aspect of reputation is important to consider. Recall that Anderson & Shirako define reputation in terms of how a person is understood within a network. From this we can imply that the network also shapes the reputation. Only those aspects of a person, presented in particular ways, are likely to be taken up in any given network. In another network, other aspects and other presentations will matter more.

2.6 VALUES AND CONCEPTIONS

As discussed above, there are both functional and normative features to the mechanics of reputation. The functional constraints have so far been better described than the normative content. To address this, I turn to the political philosophy of John Rawls, which I discuss in more depth in the following chapter. Rawls argued that people are guided in their lives by their individual conceptions of the good, or "comprehensive doctrines." These are a person's motivating framework that consists of a family of "ends" that indicates what an individual finds valuable and worthwhile in life.[33] Comprehensive doctrines reflect commitments to ideas based in philosophy, religion, and so on. Comprehensive doctrines can be entirely unique to an individual but are often shared with others. For example, people can share with others a belief in a religious doctrine that includes a code of conduct toward self and others, dietary rules, and manners of dress. The officers of a corporation may share a belief in free market capitalism and their company's appropriate place

[33] John Rawls, *Justice as Fairness: A Restatement* (Erin Kelly ed, Harvard University Press 2001) 19.

within it. Comprehensive doctrines provide a basis from which to evaluate actions and the persons who commit them. I have so far used terms like "worldview" and "priorities" to describe what guides normative choices in reputation. These terms coincide with the Rawlsian concept of comprehensive doctrines, which sufficiently describe the source of value-laden judgments about the targets of assessments.

Also involved in the process of assessment are the functional and normative constraints that lie outside of the assessor but shape her domain of choices and influence. Rawls describes the "basic structure" of society, which is the system of rules and institutions that guide public life. For Rawls, these primarily include the major institutions of government and the legal framework. However, the basic structure might also include any entity or institution whose decisions and rules affect large segments of a population. The basic structure, in whatever form it takes, can also influence the formation of reputation by providing constraints and affordances for decision-making and setting the domains of major decisions, such as courtrooms and marketplaces.

2.7 THE ROLE OF THE SUBJECT

In the prior section, the work of describing reputation as a process concentrates mainly on the work of assessors. This account is incomplete without acknowledging how subjects play a role in the formation of their reputations. In the performative account of reputation described by Goffman, we see that individuals can participate in the construction of their reputations. I know from experience how my behavior is likely to be viewed and judged by particular audiences. For example, I can anticipate the different impressions a choice to adorn my neck with a visible tattoo of an iron cross is likely to convey to different audiences. I am aware that, should I choose to get this tattoo and display it, it will affect those impressions. However, once the choice is made, it is up to assessors to respond with their assessment. The same choice can produce a range of

assessments. One assessor who admires tattoos and cares little about the historical or cultural relevance of the symbol might simply admire the mark. Another who associates that particular symbol with Nazi Germany and/or neo-Nazis is likely to suspect me of holding racist beliefs. Still another who finds tattoos foolish and those who get them misguided may simply think I am an ignoramus. In all three cases my choices matter but assessors hold the power of assessment. Yet, even given how the power of assessment in reputation seems to flow toward assessors, the subject does have influence. The question is, to what extent?

2.8 DEGREES OF INFLUENCE

The degree to which a subject can influence her reputation depends on at least three things. First, the subject must be aware they are being assessed. It is impossible to conform to a standard that one is not aware of, except by chance. Next, the standard must be comprehensible to the subject. The meaning of a reputational standard is contingent on cultural and situational factors. Third, the subject must be capable of meeting whatever standard prevails. Functional constraints, such as the availability of necessary resources for meeting the standards or the subject's prior preparation and capabilities are factors here.

Whether or which of these factors are present in a reputation context depends on the particular context at hand and the relationship between assessor and subject. When interacting with friends, colleagues, teammates, and others with whom the interaction is perceptible to all participants and there are common points of reference, such as shared culture, personal history, etc., it is possible to know one is being judged and to anticipate at least some of the features of that judgment. Also within such contexts, where participants enjoy roughly equal status, the assessment is bidirectional. Each assesses and is assessed by the others. However, there are many domains of life in which it is difficult or impossible to know who is paying attention. A detective who uses binoculars to peer

into the windows of a target's house without being noticed can form an assessment about an occupant without his knowledge. A mobile app developer who monitors how users interact with their devices outside the scope of the app's obvious functionality similarly gains the ability to profile their users without their knowledge.[34]

If a subject is aware he is being assessed, his ability to influence that assessment is contingent on having some idea about what matters to the assessor. In a job interview, the interviewer may be impressed by a job candidate's extensive vocabulary or may find certain word choices pompous. It might be revealed that app developers profile smartphone users based on how often they charge their battery completely, but it can remain unclear which charging behavior contributes to a "good" profile. Without being able to perceive what standard obtains in a given situation, accurately modulating one's behavior to fulfill it is unlikely.

The subject must also be capable of meeting the prevailing reputational standard of her assessor. Knowing that paying bills on time signals financial responsibility and low risk may lead me to prioritize doing so. However, being aware, even motivated to meet the expectations of other people does not imply that one can meet those standards. A laid-off single parent may be well aware of the reputational costs for not paying his bills on time and yet be unable do so due to his financial situation.

What degrees of influence suggest about reputation is that subjects are not entirely absent from the process. However, given the many ways in which the influence of subjects can be challenged by the constraints of information and capability, it would seem that the larger share of influence over reputation formation remains with assessors. As I have described it, reputation is a product of assessment in which the choices about what to notice, how to think about it, and who

[34] An interesting side-effect of these scenarios is that the ability of detectives and technologists to use their skills to form assessments of others for their clients contributes to their own reputations - as effective assessors!

to tell is the purview of assessors. This therefore places a great deal of reputational power in the hands of assessors. I investigate this aspect of reputation in the following section.

2.9 REPUTATION AND NORMATIVITY

In the previous section I argued that many types of reputational assessments are either individually subjective or are shaped by the choices of assessors about what to notice, how to think about it, and who to tell. Within this scope of reputation, assessors construct reputations, making a series of choices. I also briefly discussed how subjects influence their reputations. However, such influence is limited. It is the choices of assessors that play the most significant role. These are informed by systems of value, or what Rawls would call comprehensive doctrines. In this section I move from scaffolding definitions features of reputation to discussing what reputations do. I provide an overview of moral and social normativity and its link to reputation. My goal in this section is to set the stage for the development of a moral theory of reputation. I begin by reviewing how reputations are often informed by the prevailing values of assessors and their societies. Next, I describe the action-guiding nature of reputation to underscore the importance of understanding who holds the most power in reputational processes. Finally, I describe the coercive nature of reputation, particularly when under the control of powerful institutions, such as governments and economic actors who influence matters of basic justice.

2.9.1 Reputation and Values

First, by adopting a definition of reputation as a process, we accept that assessors make choices at every step of the reputation process. Those choices are informed by what an assessor values and are subject to their functional constraints. Integrating the concept of the "comprehensive doctrine" that informs people's conception of the good and guides many of their decisions, we can portray these assessor choices as based in a mix of cultural context, personal philosophy, and other aspects

of human politics that emerge from lives lived among others. This suggests that the standards by which we judge others are drawn from whatever matters to people—their values. Those values are the raw materials for the construction of the moral norms, or the obligations, we assign to members of the social context. Whether there exist truly universal moral norms, it is certainly true that many norms are neither universal nor inevitable.[35] They are constructed through social histories and prevailing culture and by arrangements of social power and influence. At their best, systems of value are moral; they promote social cohesion, fruitful and fairly distributed cooperation, and just punishment. At their worst, a system is imposed upon others who do not endorse it, with the result that benefits and losses are distributed in an unjust manner based on a concentration of power that lack legitimacy and moral worth.

2.9.2 The Normative Force of Reputation

Reputational information flows can appear facially to be either descriptive or normative, providing either an account of some observable act or an evaluative conclusion. Yet, whether presented as descriptive or normative, the *effect* of a reputational assessment is normative; it not only reports on a subject but is intended to direct the subject either toward a conception of goodness or away from badness. The normative effect is only as strong as the subject's desire or need to conform to the criteria of a given assessor.

From the perspective of an assessor, reputation is a means of sorting.[36] If I am a deciding on whom to hire to work in my restaurant, the reputation of various job applicants supports my

[35] It is not my intention to rehearse arguments for or against the existence of universal moral norms versus purely constructivist notions of normativity. It is sufficient for my purposes here to accept that some moral norms, such as prohibitions against unprovoked violence, may have universal appeal, while other norms, such as those that govern familial relationships, are socially constructed and have numerous variations.

[36] I am relying on a sense of this term offered by Barocas and Selbst (2016): "to provide a rational basis upon which to distinguish between individuals and to reliably confer to the individual the qualities possessed by those who seem statistically similar." Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671. 677.

choosing of *this* subject over *that* one. Reputational sorting has specific ends. A reputation is more than a list of preferences—as in the sorting performed for choosing targets for product marketing. Reputational sorts are structures of value intended to have effects on the person in relation to others. A reputation with a positive valence promotes selection, inclusion, or admiration based on a set of values present in the context. Negative reputational flows promote exclusion or disapproval. Whether the flow of reputational information produces a positive or negative impression is a function of the assessor's views, interests, and priorities, as well as the prevailing values of the social context. How one is perceived for the same acts and attitudes will vary considerably. To the extent that the subject of a reputation is aware of how they are perceived, desires to be perceived favorably by a particular assessor, and is capable of doing so, their reputation is action-guiding. A person will attempt to alter their behavior to meet the expectations of an assessor whose positive assessment they desire. Where primary social goods or allocations of severe sanctions are concerned, a person has even greater incentive to shape their behavior toward the priorities of their assessors.

2.9.3 Awareness and Conformity

Systems of reputation shape behavior. Members of a social or cultural context become aware of a prevailing norm, determine whether fulfilling that norm serves their interests, and choose how to respond.[37] This process helps to orient social actors within their social groups, guides them through behavioral uncertainty, and can support the achievement of their goals where those goals involve voluntary collaboration with others. The desire to be "liked" by another person is an incentive to be likable, whatever that means within the prevailing social context. The desire to be chosen for a

[37] Cristina Bicchieri, *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge University Press 2006).

job promotion is an incentive to perform one's job to a particular standard of excellence held by an employer. The reputational benefits and costs of norm adherence are action-guiding; we are incentivized to adhere to norms by the opportunities for selection and inclusion into the states of cooperation we desire, such as friendships, job promotions, and so on. Reputational subjects who fail to meet the prevailing expectations of assessors, based in a combination of personal and societal values, are likely to be tagged with a "bad reputation" and are less likely to be selected for cooperative ventures and opportunities. The threat of a bad reputation and the resulting check on access to desirable cooperative ventures and opportunities are powerful governors of behavior. If an employee would prefer to take two-hour lunches where the prevailing norm is 30 minutes, he must balance that desire with his desire to fulfill the expectations of those who have some control over something he desires: his employment. The opposite is also true. Good reputations afford access to desired ventures and opportunities. As a result, a positive reputational standing is a reward for subjects who meet normative expectations. A common norm in the American workplace is that working extra hours and weekends is viewed as demonstrative of one's commitment and dedication. Fulfilling this norm can lead to rewards such as raises and promotions, or at the very least, continued employment. Reputation subjects are motivated, to the extent they are able, to shape their behavior and the impressions they give off to meet an assessor's standards. This is the *normative force* of reputation. In the particular context of a given subject and his assessor, the subject's reputation is the playing field in which norms are expressed and enforced.

The aspects of normativity that are of most interest for moral evaluation are those that are involuntary and coercive, the contexts of normative force in which the subject has little choice in

the matter of what is expected. Philosopher Christine Korsgaard[38] argues that, in most cases, the prevailing norms that carry the most weight are those that are not chosen by individuals but are imposed upon them. While individuals may agree with and may actively work to uphold the normative standards by which they are judged, this engagement should not be confused with prior consent. One can consent to a set of norms. Someone who, as a free and rational adult, voluntarily joins a reclusive religious order that requires strict adherence to specific behavioral and spiritual practices can be said to consent to a set of norms. One who is born into the same religious order and must either conform to the order's code or be shunned by their family, cannot be said to consent to the norms.

This is also true of the basic structure of society, which is the underlying framework of a society described by Rawls. The basic structure consists of the foundational rules and institutions that guide political and social life. In the basic structure, the norms that are imposed take shape as rights, powers, and obligations. Our place within the basic structure of is not subject to actual consent. Society exists prior to our entry into it and remains a stable force throughout our lives. While it is possible to move from one society to another, doing so is only possible for some, and in any case, typically means trading one basic structure for another. As Rawls writes, we "enter the social world only at birth, leave it only by death."[39] Consequently, many of the normative standards that obtain in society, especially those that are backed by the power of law or some other persuasive force, are both involuntary and coercive. (I discuss this in more detail in Chapter 0.) Indeed, the coercive nature of normativity is the point. Social order is maintained by the sense that we *ought* to conform when faced with the soft power of social mores and the feeling that we *have*

[38] Christine M Korsgaard, *The Sources of Normativity* (Cambridge University Press 1996).
[39] Rawls, *Justice as Fairness* (n 33) 55.

to conform when faced with the harder power of institutional sanctions. It is perhaps for this reason that Korsgaard characterizes normative constructions as *obligations*, where obligation is understood to be "the imposition of value" on a subject, including a subject who may be reluctant or recalcitrant in their acceptance.[40] This contrasts with the view that moral norms emerge from social consensus, are easily recognized for the good they do, and are welcomed without dispute. For Korsgaard, the duty to adhere to norms is a claim upon us, and as such, we have a legitimate claim in return; we can demand a justificatory account of those norms.[41]

The normative force of reputation becomes important if we are to evaluate the legitimacy of reputational standards and practices that affect the basic justice interests of subjects. We ought to be able to use some standard to evaluate assessors and assessments that directly affect the material and existential destinies of others to ensure they are fair and reasonable, rather than arbitrary or malicious. Before attempting this evaluative work, I continue with scaffolding an understanding of reputation as a site of basic justice by drawing attention to the construction of norms. Specifically, I am interested in examining why a particular set of reputation-influencing norms prevails in a given context while others are deprecated.

2.9.4 Normativity and Social Power

The beneficial aspect of norms as expressed through reputational standards are their contribution to the social cohesion required for cooperation as well as group identity. Rough agreement about conversational etiquette in a professional setting or the use of hand signals when riding a bike are examples of normative standards that reduce friction between individuals, promote collaboration, and help to avoid conflict. The normative force of reputation encourages

[40] Korsgaard (n 38) 4.
[41] Korsgaard (n 38).

individual actors to conform to such expectations and may lead to desirable outcomes for all. Yet there are risks associated with the normative force of reputation. We often encounter powerful institutions who set the terms of normativity in many domains of life. Governments create laws and regulations that incentivize certain behaviors while discouraging others, backed by the power to reward and punish. Non-governmental institutions, such as the financial industry and, increasingly, the technology industry, also play a role in shaping human behavior by creating incentives and disincentives for particular ways of being and acting.

While the existing order of society may be welcome, unwelcome, or somewhere in between, the norms produced by this order, and therefore its reputational standards, can be coercive, setting up a conflict between the interests of assessors and subjects. While such conflicts are to be expected—agents hold different interests from each other and are as likely as not to pursue self-interest—there are contexts where such conflicts affect the basic justice concerns of the participants. Where basic justice concerns are at stake, such conflicts bear evaluation and, when we find them to be illegitimate or harmful, should be subject to revision.

2.10 ALGORITHMIC REPUTATION

So far in this chapter I have discussed reputation as feature of human relations, identifying the key participants as assessors and subjects. I narrowed the scope of this study to reputations assigned to human subjects and further narrowed the scope to relations in which an assessor possesses a significant level of economic, epistemic, or political power relative to the subject. I also broke down the mechanics of the reputation process into collection, analysis, and production and I discussed the constraints for subjects in exerting control over reputations. I now introduce the concept of *algorithmic reputation* as a digitally-mediated variant of reputation.

Algorithmic reputation is a term I will use to describe processes of producing reputational

assessments using digital technologies. It is shaped by a set of constraints, affordances, and dynamics that parallels those of reputation in general but is shaped by its production by and through information systems and its commodification as a set of information goods. Algorithmic reputation is made possible by the availability of very large, and growing, electronic data repositories about people, that are generated and routinely refreshed as a byproduct of digitally-mediated lives. Acted upon with the data modeling and analysis tools of artificial intelligence, these repositories provide the raw materials for generating predictive assessments. Such assessments are increasingly used to automate and scientize the evaluation of prospective employees, criminal defendants, insurance customers, romantic partners, tenants, and many others. Employment recruiting and screening is a particularly popular domain for algorithmic reputation technologies and techniques.[42] Approaches range from identifying prospective job candidates from diverse online sources to predicting "successful" job candidates by analyzing recorded data acquired during job interviews.[43] While employee recruiting and screening is a particularly rich area of research and development for algorithmic reputation, the market is robust and any opportunity to supplant familiar forms of reputational judging with an algorithmic variant seems likely to be pursued.

Despite some interesting differences, many aspects of algorithmic reputation coincide with reputation as I have so far described it in this chapter. Reputation in all its forms is an informational process of assessment and involves choices and constraints that operate on processes of collection, synthesis, and production. Whether produced by humans, machines, or through some combined effort, it reflects and reproduces human goals and worldviews.

---

[42] I conduct a case study of automated hiring systems in Chapter 5.

[43] Ifeoma Ajunwa, 'The Future of Work: Protecting Workers' Civil Rights in the Digital Age' (witness statement), Joint Hearing of Subcommittee on Civil Rights and Human Services, United States House of Representatives, 116th Congress, Washington, D.C., February 5, 2020 <https://edlabor.house.gov/imo/media/doc/AjunwaTestimony02052020.pdf> accessed 28 December 2020.

2.10.1 Collection

Collecting information for algorithmic processing is a byproduct of digitally-mediated lives. Virtually every communication, transaction, and interaction are recorded and stored. Data is collected from devices, platforms, and services, including clickstream data from websites and mobile apps, as well as the mining of existing data repositories that may include surveillance and personal data held by both government and commercial entities. [44] Government data on important life events (*e.g.* birth, death, marriage, divorce, arrest, bankruptcy) as well as the documented histories of retail transactions and financial activity has become trivial to acquire by data miners and other information industry players.[45] The quantity of the resulting information is large and, when aggregated, quite rich in detail. Unlike the limitations on human attention and memory, and the physical limitations of non-digital data storage, digital data storage is cheap and capacious. Social media data is an especially rich source of reputational data. The producers of reputational insights are keenly interested in how people are connected through social and professional ties based on highly developed theories about behavior similarities among peer groups and families.[46] Digitally-mediated subjects passively emit data about themselves through the full range of interactions with digital systems and as a byproduct of contemporary communications and commerce. Additionally, people are incentivized to actively provide reputational data. For instance, transactants rating each other on peer-to-peer commerce platforms, individuals submitting to lifestyle and activity monitoring in exchange for insurance discounts, and an

---

[44] Casey Johnston, 'Data Brokers Won't Even Tell the Government How It Uses, Sells Your Data' (*Ars Technica*, 21 December 2013) <http://arstechnica.com/business/2013/12/data-brokers-wont-even-tell-the-government-how-it-uses-sells-your-data/> accessed 21 July 2016.

[45] Natasha Singer, 'The Scoreboards Where You Can't See Your Score' *The New York Times* (27 December 2014) <http://www.nytimes.com/2014/12/28/technology/the-scoreboards-where-you-cant-see-your-score.html> accessed 29 December 2014.

[46] *cf.* FinTech Silicon Valley, *Video Interview with Brian Ley, CEO/Founder Alpharank* (2017) <https://www.youtube.com/watch?v=JJe9TlSM_8M> accessed 26 July 2018.

emerging requirement for job candidates to submit to increasingly invasive information disclosures and behavioral analysis in employment screening.

As with other forms of reputation, information system designers make choices and encounter constraints when collecting digital information. They are liberated form some human limitations of attention and memory by always-on and frequently invisible data collection practices coupled with seemingly infinite and inexpensive data storage. However, resembling human observers, they are functionally constrained: Data collected by one entity is not necessarily available to others; data collection is only as complete as available sensing technologies; and there is the risk of provoking resistance or regulation by going too far in violating prevailing information norms.[47] Behavioral data can lose some of its meaning when converted into data. Even with the affordances of data collection and storage, designers often make choices for the sake of efficiency about the perceived value and relevance of information for their purposes and those of their customers. So, while the collection of algorithmic reputation data involves a different inventory of limitations, there are parallel factors of choice and constraint that shape digital data collections.

2.10.2 Synthesis

Collected data has to be acted upon to provide meaning. Thousands of companies are now in the business of assembling data about individuals from disparate sources, combining it with information about others, and processing it using the immense and growing capabilities of analytic systems. Artificial intelligence technologies, including machine learning algorithms, rely on models built from data to portray the world. Algorithms compare a given subject with a model of similar subjects and render predictions. As legal and philosophy scholars Michael Veale and

[47] While it appears that the collective goal for the information industry is to gain access and collect as much information as technically possible about subjects, this effort is bounded by technical limitations; many companies recognize that the path to avoiding regulation is to step back from the far extremes of what is possible, for now.

Reuben Binns describe, machine learning algorithms "are designed to discriminate."[48] They distinguish patterns of data points from a larger mass and produce inferences. In question is whether the discrimination is harmful discrimination. A model of an ideal job candidate based on the profiles of prior hires may reproduce aspects of preexisting hiring discrimination.[49] The entry points for influences that confound data modeling are many and complex. For example, researchers found that most image recognition systems are trained on models that come primarily from the United States and Europe, skewing results toward those cultures, norms, and aesthetics.[50] The logics of analysis are only as precise as the considerations of designers, who may overlook important distinctions. Such oversights are revealed by the many reported problems of image search systems misidentifying people with darker skin tones as wild animals and text search results that emphasize portrayals of African American girls and women as pornographic models.[51] Of course, human synthesis of information is also vulnerable to oversights, biases, and discriminatory perspectives. The point here is that algorithmic processes are not immune to their own limitations and biases.

2.10.3 Production

The third stage of algorithmic reputation is production, where an algorithmically produced assessment is provided to an interested consumer. The technologies and data access required to construct algorithmic reputation limit its practice primarily to technology firms in pursuit of a

[48] Michael Veale and Reuben Binns, 'Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data' (2017) 4 Big Data & Society 2053951717743530, 2.

[49] Ifeoma Ajunwa, 'Automated Employment Discrimination' [2019] SSRN Electronic Journal <https://www.ssrn.com/abstract=3437631> accessed 13 March 2020.

[50] Shreya Shankar and others, 'No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World' [2017] arXiv:1711.08536 [stat] <http://arxiv.org/abs/1711.08536> accessed 24 July 2018.

[51] Noble (n 10).

market in reputational products and services. An example is Lenddo, a company that seeks to disrupt the consumer loan industry by offering financial risk assessments based on social media analysis[52] Another example is a tenant screening product that uses an applicant's online profile to create personality reports for landlords.[53] Governments are also a market for algorithmic reputation products. Many jurisdictions are experimenting with criminal defendant risk-assessment tools, such as PSA and SAFER, which support decisions in bail determinations.[54] As in other forms of reputation, algorithmic assessments are produced for an interested audience, which shapes the information provided. The interests of the employers, landlords, courts of law, and other target consumers drive the market and influence the entire chain of collecting, synthesizing, and producing information for an algorithmic reputation.

2.11 CHAPTER CONCLUSION

Reputation appears to function—often well—as a tool for decision-making in spite of its remarkable plasticity and suasiveness. Another view is that it is no accident that reputation is both useful and frail. Reputation derives its value, in part, from its very humanness; its uncertainty and mutability are part of how it functions either to the advantage or disadvantage of the assessor or subject. We know as a matter of lived experience that reputations operate along a spectrum of reliability and objectivity. We have known people to get jobs, win over romantic partners, attain political offices, and gain other opportunities that do not seem to clearly correspond with the

[52] Yanhao Wei and others, 'Credit Scoring with Social Network Data' (2015) 35 Marketing Science 234.

[53] Caitlin Dewey, 'Creepy Startup Will Help Landlords, Employers and Online Dates Strip-Mine Intimate Data from Your Facebook Page' *The Washington Post* (9 June 2016) <https://www.washingtonpost.com/news/the-intersect/wp/2016/06/09/creepy-startup-will-help-landlords-employers-and-online-dates-strip-mine-intimate-data-from-your-facebook-page/> accessed 11 June 2016.

[54] Kate Patrick, 'Arnold Foundation to Roll Out Pretrial Risk Assessment Tool Nationwide' (*InsideSources*, 4 September 2018) <https://www.insidesources.com/arnold-foundation-to-roll-out-pretrial-risk-assessment-tool-nationwide/> accessed 7 April 2019.

reputations we would assign them. This suggests that many forms of reputation are purely subjective; an "arrogant fool" for one person is an "imaginative visionary" for another. Consequently, we ought to pay attention to the method of construction for reputations, especially where reputational assessments affect the basic and primary interests of subjects. In matters where the allocation of primary social goods or severe sanctions are mediated by reputational assessment, the legitimacy of reputation processes and the authority of key assessors requires our considered moral judgment.

In this chapter, I have sought to develop a full account of reputation as a process practiced by individuals, small groups, and powerful institutions. I identified three definitions of reputation, including reputation as object, reputation as Goffmanian performance, and reputation as process to spotlight where we can most effectively set our normative gaze. Ultimately, I argued that a procedural account of reputation provides the best opportunity for normative evaluation.

I then turned attention to the mechanics of the reputation process, which includes processes of collection, synthesis, and production. Throughout this process, assessors make myriad choices about what to notice, how to think about it, and who to tell. I suggest that such choices are subject to constraints and influenced by the views and priorities of assessors, including their "comprehensive doctrines," or the frameworks of belief that guide them. I then discussed the ways that subjects can influence their reputations, taking note of the limitations on that influence, which include obstacles to understanding what matters to assessors and functional constraints on being capable of conforming to an assessor's standards.

I also discussed how reputation operates in the world as a "normative force" in which the behavior and attitudes of reputational subjects are shaped by the prevailing values of assessors and the background forces of their societies and varying to some degree by individual contexts.

Reputations, and their power to shape lives, are formed around the values of a society or social group with the result that the group's norms dictate the content of one's reputation. A simplified view is that conformity with the community's values results in a positive reputation, whereas a failure to conform results in a negative reputation. In practice, the formation of reputation is more complex, with individual agents aspiring to reputational profiles that suit a mix of local/specific and more general community commitments.

Finally, I introduced *algorithmic reputation* as a variant of reputation in which the tools of artificial intelligence act on the growing repositories of information about digitally mediated lives to produce predictive insights about personalities and behaviors. I described the many parallels between reputation as described previously and the algorithmic variant, particularly in its mechanics. As with other forms of reputation, the choices and flaws of humans and their messy social relations find expression in the collection, synthesis and production of algorithmic reputations.

CHAPTER 3 — RAWLSIAN JUSTICE

3.1 INTRODUCTION

In the previous chapter, I discuss how reputations are constructed, by humans and algorithms, under conditions of constraint and influence. In subsequent chapters, I further deconstruct reputation into its domains of practice. I also make normative claims about the role of reputation in the assignment of important goods, including income, liberty, powers, opportunities, and self-esteem. When reputational assessments are automated within processes that are inscrutable and/or unaccountable, asserting claims based in justice remains a worthy project. Lurking within these arguments and claims is a portrayal of reputation as a gatekeeper and also as a form of reasoning. Reputation leads to conclusions and decisions. When these are of limited significance to the subject's life, such as a choice about whether to invite someone to dinner, they do not merit serious consideration as philosophical questions. However, when important goods or severe sanctions are mediated by reputational assessments, reputation becomes a legitimate target of moral inquiry. Ultimately, I argue that the types of decisions being made indicate something about the reputational process employed. To get there, I will adopt an existing philosophical framework and model of society in which to situate the normative dimensions of my work. This provides a basis from which I eventually support a set of positions about what matters in reputation.

There are many theories of justice that may offer guidance for our reasoned moral judgment about the conditions and practitioners of reputational assessments in matters of fundamental justice. I have chosen to employ and modify the work of John Rawls.

In this chapter, I introduce aspects of the political philosophy of John Rawls and offer his work as a roadmap for evaluating systems of reputation. Rawls provides a detailed conception for the design of a society as a fair system of cooperation for mutual advantage. In his later work,

Rawls provides details about how this society would remain both fair and stable. While Rawls did not take on the specific concepts addressed in this dissertation, his work provides a number of useful concepts. They reflect many of the ideals inherent in the aspirational discourse guiding contemporary democracies, including a strategy for constructing a fair society and including the requirements for an open and inclusive political system. Rawls did not specifically consider information systems and practices in his work and yet I see evidence that a full conception of justice in contemporary society is deficient without reference to the influence of technology. In this project, I attempt to remedy this by enfolding an analysis of reputational profiling as an information practice within a Rawlsian conception of society.

Because Rawls's theory is extensive, I provide a summary of several key concepts from his work that will frame my own arguments about the moral and political status of reputation, particularly algorithmic reputation. I offer some friendly amendments to certain aspects of the work that suit both the subject of this dissertation and the conditions of contemporary societies. The philosophy of John Rawls is not uncontroversial, nor can we say with any certainty that it is the best or only framework for correctly identifying the principles of justice. In adopting the work of Rawls, I begin with the assumption that his approach is sufficient and set aside for now an in-depth defense of his work where it conflicts with competing frameworks. Even with this caveat, I include some critiques specifically leveled at his work and argue for making use of his approach with some qualifications.

3.2 JUSTICE AS FAIRNESS

In his major works, John Rawls offers a strategy for establishing his vision of an ideal society developed through an enhanced contractarian regime of procedural justice. Rawls's stated intent was to make a fresh attempt at applying moral philosophy to the project of democratic pluralism

by focusing on the achievement of mutually beneficial cooperation. According to his biographer, Rawls's marquee political theory, "justice as fairness," was intended as a moral conception of justice that could demonstrate how human beings, working together, could overcome the evils of the world.[55] Rawls sought to update the social contract doctrines found in the philosophies of John Locke, Jean-Jacques Rousseau, and Immanuel Kant as a response to the perfectionism, intuitionism, and utilitarianism he saw as the dominant strains in classical and contemporary political theory.[56] By adapting Kantian constructivism to his political philosophy, Rawls follows Kant in promoting human freedom and equality as the foundations of a just society.[57] Rawls builds upon Kant's view of the person as a public figure to construct a society that is not is only based on freedom and equality but is a society that sustains and promotes those values through the commitments and inclinations of its members. This recursive model, of a society *constructed by* free and equal people with the result that it *produces* free and equal people, is fundamental to Rawls's belief about what is required for a society marked by enduring fairness and stability over time.[58]

Rawls recognized that people are guided by their individual conceptions of their good, which include their fundamental aims and "highest-order interests," that both direct and regulate their conduct.[59] Yet, by prioritizing general, public, principles that support a free and fair system of cooperation, the people could generate the principles and structures of a society that all its members would endorse and commit themselves to upholding. In contrast, according to Rawls, utilitarianism threatens long-term collective ends by placing value primarily on individual satisfaction without

---

[55] Samuel Richardgar Freeman, *Rawls* (Reprinted, Routledge 2010).

[56] Rawls, *A Theory of Justice* (n 29) Preface.

[57] See, for example, Kant's discussion of the connection between freedom of thought and freedom of action with human dignity in "What is Enlightenment" (1784).

[58] Rawls, *A Theory of Justice* (n 29) §§ 6, 7.

[59] John Rawls, 'Kantian Constructivism in Moral Theory' (1980) 77 The Journal of Philosophy 515, 525.

attending to instilling a "sense of justice" in the members of a society that inspires them to work toward common justice.[60] Intuitionism fails to offer sufficient strategies for stability and the resolution of conflicting principles. A distinguishing feature of Rawls's theory is its emphasis on *cooperation,* reflecting Rawls's core belief in the sociability of people and their willingness to commit to the welfare of a society that is fair and will ultimately reward them for seeing their own interests embedded in those of everyone else. For Rawls, utilitarianism and intuitionism, which emphasize the efficient *coordination* of largely self-interested people over mutual cooperation, would each fail to provide the stability necessary to fulfill the goal of long-term stability and to set the conditions for attaining individual benefits from collective endeavors.

3.2.1 The Two Moral Powers: Reasonableness and Rationality

Following Kant, Rawls argues from the position that arriving at the first principles of justice requires the reasoned moral reflection of persons. The principles are not foreordained. Rather, a conception of justice is a construction by persons who are both *rational and reasonable*. People are rational in that they have a conception of their good. Rawls develops the concept of the "moral person" in possession of the *two moral powers,* as a threshold requirement for reaching rational and reasonable engagement with society and the development of principles of justice.[61]

So, what are these moral persons? First, they are *rational*. For Rawls this means that they have a conception of their own good, that they are interested in pursuing a plan of life. Rawls focuses on that aspect of rationality that includes the capacity and desire to achieve the necessary conditions of a good life, which we can label "rational eudaimonism." For Rawls this means those

[60] Rawls, *A Theory of Justice* (n 29) 476. Note that utilitarianism would likely not provide the moral basis for the first principles of justice Rawls has in mind. Utilitarianism allows for the sacrifice of some for the good of others, which seems an unlikely choice by a rationally self-interested participant in the original position from behind the veil of ignorance.

[61] The moral person conception is controversial, as discussed in 3.10.

things that "are generally necessary as social conditions and an all-purpose means to enable human beings to realize and exercise their moral powers and to pursue their final ends."[62] Evidence of this rationality includes having identified projects and plans for achieving them, including whatever tendencies and practices are required to convince others to support those projects as needed. In later work, Rawls is careful to revise this view of rationality to include the pursuit of the good of others, including persons, communities, places, and institutions.[63]

Reasonableness is distinguished from rationality as the inclination to cooperate with others in a manner that is fair, and which advances common aims. Because Rawls is convinced that rational persons would agree that society is a fair system of cooperation for mutual advantage, they will *reasonably* limit the pursuit of their own ends to the extent that they perceive that others will do so also. This is represented as one's participation in a *public* conception of justice where a reasonable person is "ready to propose principles and standards as fair terms of cooperation and to abide by them willingly, given the assurance that others will likely do so."[64] Here, Rawls makes reasonableness the foundation of one's commitment to society as a cooperative venture. Reasonableness, then, subordinates the rational by placing limits on the maximal pursuit of rational advantage in pursuit of mutually beneficial ends.[65]

Rationality as the basis of the moral person and reasonableness as a sort of check on their pursuit of self-interest appears to create a tension between self-interest and a shared vision of justice. For Rawls, conceiving of society as a fair system of cooperation requires that those possessing the "two moral powers" would see themselves reflected in this construction. There is a

---

[62] Rawls, 'Kantian Constructivism in Moral Theory' (n 59) 526.
[63] Rawls, *Political Liberalism* (n 14).
[64] ibid 49.
[65] Rawls, 'Kantian Constructivism in Moral Theory' (n 59).

risk that *reasonableness* appears to be a form of altruism. Yet, Rawls holds rationality and reasonableness to be complimentary and essential features of the person. By acknowledging the role of rational inclinations in concert with reasonable conceptions of justice, Rawls attempts to account for the entire person as both self-interested and committed to the requirements of cooperation.

3.2.2 Stages of Justice

Rawls attempts to provide a complete account of political justice by proposing the design of an ideal society constructed to achieve mutually beneficial cooperation. He does so by predicting the content of a just society constructed through just procedures followed a means for employing additional just procedures to preserve it. In *A Theory of Justice* Rawls proposes to set up the conditions for this society as emerging from a set of negotiations among people who will represent the interests of all through a set of agreements reached in an initial condition which is fair. This first stage of his exposition produces the core principles of justice and sets up the basic institutions that regulate them. In his later works, Rawls provides a set of concepts and strategies to ensure the enduring stability of this society following from its initial principles. In sections 3.3 to 3.7, I discuss some of the key elements of the first stage of Rawls's theory. Beginning in section 3.8, I discuss elements of the second stage, which are among the most important to my arguments about reputation.

3.3 THE WELL-ORDERED SOCIETY AND A PUBLIC CONCEPTION OF JUSTICE

Rawls proposed that the ideal form of mutual cooperation would take shape as a "well-ordered society."[66] Here I provide the elements of a well-ordered society as described by Rawls. First, it is

[66] Rawls, *A Theory of Justice* (n 29) 4.

effectively regulated by a public conception of justice. It is public in these three ways: First, members know and accept the same conception of justice, and they know that everyone else knows and accepts it.[67] Second, the basic structure of society including its central elements of political organization—rules and formal systems—are both visible to everyone and understood to promote justice. Third, members of a well-ordered society have a "sense of justice" that enables them to be the agents of the public forms of justice that the society provides.[68] In addition to the public conception, the well-ordered society is one in which members consider themselves and others as free and equal, by which we mean everyone accepts and knows that others likewise accept the same first principles of rights and justice.[69] This is not to say that everyone agrees on every feature of society and its government. It is to argue that there exist first principles that all can agree upon. Accepting the pluralistic nature of society, Rawls presumes that each member is guided by individual "conceptions of the good," and while many may share the same conception, there are presumed to be conflicts. However, because the society Rawls envisions is a cooperative venture for mutual advantage, members must agree on its essential features, such as systems that fairly allocate roles and responsibilities, claims and benefits, and other aspects of basic justice. Additionally, members of society recognize each other as entitled to make claims on institutions in pursuit of their own ends. Finally, the well-ordered society is a society that is *stable.*[70]

Rather than setting down the details of this venture, Rawls offers a procedural account; an account designed to set up the terms of agreement between parties with both conflicting and

---

[67] This may appear to be a tall order that demands a certain amount of faith in others. In Part III of *Theory of Justice* and in his later works, Rawls develops arguments about moral psychology in which members develop a "sense of justice." Over generations, both the evidence of fair cooperation and the way it shapes psychological development helps to ensure that this reciprocal arrangement is both rational and reasonable.

[68] Rawls, *Justice as Fairness* (n. 33) 8–9.

[69] Rawls, 'Kantian Constructivism in Moral Theory' (n 59).

[70] ibid. Stability as a feature of a well-ordered society, and its requirements, is discussed further in 3.7.2

overlapping aims and interests. Those agreements, properly arrived at, set the conditions through which societal members conduct themselves and interact. Central to Rawls's aim is to achieve a system that is both fair and *stable* over time by providing for decisions that are reasonable and rational, and that produce outcomes that most can agree are "fair." Fairness is not something set down by Rawls but is a feature of the processes that produce agreements and institutions.

3.4 PROCEDURAL JUSTICE

A just society in this view is one characterized by just *processes* for developing its rules and institutions and also for applying those over time. Stability is an important goal for Rawls and key to that stability is attention to the processes that both construct and implement a well-ordered society, processes that are themselves fair. Participants in a just society are more likely to endorse a society's political culture when it is exemplified by a procedural fairness that guarantees the impartial application of rules and free and public deliberation. In such a culture, members recognize that society promotes their "plan of life" through fruitful cooperation.[71]

Rawls's target then is to establish what he labels "pure procedural justice," which are the methods for producing a fair system of cooperation that is mutually agreeable and impartial. The concept of pure procedural justice produces desirable outcomes using processes that are characterized as either "fair" or "correct," and are, in either case, achievable.[72] Here, it is important to note Rawls's distinction between a conception of justice focused on just outcomes (substantive justice) and his preferred conception that focuses on just processes (procedural justice). How we arrive at a decision (such as the distribution of a scarce resource) is the work of justice. The achievement of just outcomes is predicted to occur, but its forms are not preordained. In this way

[71] Rawls, *Justice as Fairness* (n 33) 35.
[72] Rawls, *A Theory of Justice* (n 29) 85–86.

Rawls distinguishes his ethical framework from consequentialist theories that place the majority of evaluative weight on results. Rawls instead presumes that focusing on outcomes produces a system of justice that may achieve short-term benefits but would lack stability; conflicting conceptions of the good would threaten overarching principles that ensure long-term endorsement of the system "from one generation to the next." Following Kant's logic, we cannot reliably predict consequences, so it is not rational to arrange a system of justice based only on an evaluation of outcomes. However, with the properly chosen first principles and processes the outcomes are most likely to be just.

Procedural justice, or the methods of achieving a just result, are presented by Rawls in three possible forms; perfect, imperfect, and pure. In all cases, procedural justice describes a method of achieving a just result. Perfect procedural justice is described using the analogy of dividing a cake among many people. It is presumed, in advance, that the fairest division of the cake is an equal-size piece for all. A procedure for achieving that aim is then designed to ensure that outcome. Rawls rejects perfect procedural justice as unworkable. Perfect procedural justice requires the existence of an antecedent criteria from which to judge the action. However, following Kant, Rawls rejects the "moral facts" required to produce such criteria. Rather, Rawls argues that the only moral fact is the procedure for arriving at the principles of justice.[73]

Imperfect procedural justice describes systems that, while aimed toward justice, are predicted to produce many unjust outcomes over time. Rawls employs the analogy of a criminal trial in which the outcome cannot be reliably predicted, regardless of the defendant's guilt or innocence, and yet it frequently results in innocents being convicted and the guilty going free.[74] While this

[73] Rawls, 'Kantian Constructivism in Moral Theory' (n 59).

[74] Rawls's characterization is not a particularly accurate view of the American justice system, which, in its ideal form, favors the rights of the accused over those of accusers to limit the risk of convicting the innocent.

system might produce only a very low level of injustice and may be considered the best available approach to the problem at hand, it is not ideal because of its many failures.

Pure procedural justice is the preferred approach. Pure procedural justice does not require an independent criterion of justice and yet it is predicted to produce just outcomes. Again, Rawls argues in line with Kant that we cannot bet on outcomes. All we can say for certain is that rational and reasonable people, under ideal conditions, would conceive of a procedure to produce just outcomes. Whether they are successful is measured by those outcomes, but the outcomes do not design the procedure. While this may seem illogical, I believe what Rawls is suggesting here is that, even while not appealing to consequentialism, procedures and outcomes are tightly coupled. If the procedure is fair, just outcomes result by following it. "What is just is defined by the outcome of the procedure itself."[75] Arriving at such a procedure is no simple matter.[76] Failing to design it such that it incorporates enough details about the world, human behavior, economic conditions, etc. is unlikely to produce anything better than imperfect procedural justice. However, if we accept that pure procedural justice is indeed the means to achieving fairness, the key challenge is the design of such a process.

3.5 THE ORIGINAL POSITION AND THE VEIL OF IGNORANCE

Rawls's famous thought experiment, "the original position," is offered as an exemplar of pure procedural justice. It is "pure procedural justice as the highest level."[77] Justice as fairness is essentially a social contract account of society; participants agree to give up some of their freedoms

[75] Rawls, 'Kantian Constructivism in Moral Theory' (n 59) 523.

[76] A case that is used to trouble the concept of pure procedural justice is offered by Robert Nozick in *Anarchy, State, and Utopia,* (Basic Books 1974), known as the 'Wilt Chamberlain' case in which Nozick constructs a scenario that appears to follow a just procedure but produces a result that Rawls would likely not consider just. Whatever else we might learn from this case, it demonstrates that the requirements for constructing a process of pure procedural justice, if seeking Rawlsian justice, are significant.

[77] Rawls, 'Kantian Constructivism in Moral Theory' (n 59) 523.

and join with others for mutual protection and collective benefit. The original position is Rawls's meticulously designed forum for negotiating the terms of this contract, where moral persons representing the whole of society enact a just system of rules and principles from an initial situation which is fair. The work of the original position is to design the "basic structure" of society (discussed in 3.7), and also to produce a set of principles and initial distributions of life essentials. The key institutions and offices of society are also designed and allocated.[78]

Pure procedural justice is achieved in the original position by placing entrants to the negotiations in the original position behind a "veil of ignorance" that shields them from knowing much about themselves—their abilities, assets, challenges, and their individual conceptions of the good—and thereby unable to organize society to favor their specific, rather than general self-interest.[79] Rawls assumes that, lacking self-knowledge from behind the veil of ignorance, negotiators in the original position would see their own self-interest served by ensuring that the least well-off were guaranteed basic liberties and other advantages.

In the original position, the apparent tension between the rationality and reasonableness of persons is on display. Rawls is committed to a political philosophy motivated by sociability and cooperation, yet self-interest remains a guiding motivation that promotes fairness. Despite a reputation (largely unearned) as a theorist concerned more with collective than individual ends, Rawls is a champion of the individual as possessor of Kantian rationality. Self-interest plays a significant role in Rawls's vision of what motivates people toward fairness, which is not particularly altruistic or collectivist. Instead, persons are rational eudaemonists who seek a minimum allocation of primary goods as a rational end. From behind the veil of ignorance people

[78] Rawls, 1971/2005 (n 56), §4.
[79] Rawls, 1971/2005 (n 56), §24.

seek these goods and share similar fears about deprivation. When confronted with the uncertainty imposed by this condition and the risk of being among the less fortunate, rational actors would propose to minimize the possibility of deprivation for all, hedging their bets, as it were, against this risk. While Rawls argues elsewhere that members of a just society are motivated to promote justice and that motivation is part of what creates the conditions of fairness,[80] the original position is distinct; it produces principles of justice likely to have collective benefits, but they are arrived at through the pursuit of individual ends.

The original position is offered as the prime example of pure procedural justice. Freed from those features of self-interest that could disadvantage others, participants are free (or self-motivated) to develop a political conception of justice that is most likely to, first, provide adequately for the wellbeing of the least well-off, and second, to be acceptable to all, regardless of the individual plans of life each member ultimately chooses for oneself. While Rawls's target here is to emphasize the importance of the process, he also makes specific predictions about its outcomes.

3.6 PRIMARY GOODS AND THE DISTRIBUTIVE PARADIGM

The fair distribution of those things that people value, and that could theoretically be granted or claimed by society, has animated political theory for centuries. The tradition of liberal political theory has been especially concerned with distributive justice, particularly the redistribution of goods that can concentrate into relatively few hands. Rawls is strongly associated with distributive justice, or the distribution of primary goods, although in Rawlsian theory, the distribution of goods plays a relatively minor role. The distribution of goods matters to Rawls mainly as instrumental to members of society being able to realize their more fundamental rights, such as the basic liberties,

[80] *cf*. Rawls, 1971/2005 (n 56), §72–76, 86.

freedoms, and material requirements that promote participation in a fair system of cooperation and, as emphasized in his later work, in a functioning democracy. However, despite my belief that the distributive justice paradigm is overemphasized as a feature of Rawlsian theory, this aspect needs at least a mention because Rawls is so closely associated with it.

Rawls describes "primary social goods" as those that a society can bestow upon its members, such as rights, liberties, powers, opportunities, wealth, and self-respect. These social goods are distinguished from other goods a person may desire, pursue, or simply have by virtue of birth or fortune. These could include "natural goods," such as health, vigor, intelligence, and imagination, which Rawls holds are not directly distributable by society. Rawls focuses primarily on the distribution of goods as a function of properly established institutions, the "basic structure" (see 3.7) that ensures that the opportunities and constraints designed in the original position are articulated and enforced for everyone's benefit. In his later work, Rawls specifically indicates the relationship between the distribution of "primary social goods" (things of material value) and a society based on a liberal conception of justice; a liberal conception of justice protects basic rights and also ensures that people have the material means "to make effective use of those rights."[81]

Rawls distances himself from natural rights theorists who would be more likely to argue that certain goods are deserved or held as a feature of their humanity. Under a natural rights paradigm, certain fundamental rights, for example, are not given but attach to the person *a priori*. John Locke, for example, holds that persons have a natural right to their own "perfect freedom," a portion of which they grant to the state in exchange for protection.[82] Rawls instead situates freedom as a

[81] Rawls, *Political Liberalism* (n 14) 157. Rawls identifies primary *social* goods as those things that are connected to the basic structure of society. A useful example of a primary social good in the US context is public education. It is considered a requirement for participation in society and is allocated by public institutions.

[82] John Locke, *Second Treatise of Government: An Essay Concerning the True Original, Extent and End of Civil Government* (John Wiley & Sons 2014) §87, 95.

function of one's participation in human society: "Whether men are free is determined by the rights and duties established by the major institutions of society."[83] This follows from the social contract paradigm; we exchange something for the order that an organized society provides. The major tensions among contract theories rest on just how much one must surrender to an external authority.

Some contractarians hold that the realm of goods subject to distribution, except by the individual who holds them, is quite small. Robert Nozick is among those who argue against notions of distributive justice at the hands of the state or other dominating entity, arguing that meddling in the distributions that arise from the free choices of rational persons is a threat to human freedom. In particular, the free exchange of property among those who justly acquired it promotes a just society through the realization of individual freedom and intersecting self-interest.[84] Meanwhile, Elizabeth Anderson captures the work from another dimension of distributive justice where the distribution (or redistribution) of things of value ought to occur to ensure democratic equality. Anderson seeks to articulate the role that distributive justice can play in producing a society in which people stand in equal relation to each other without resorting to punishment or pity. Departing from conceptions focused on the distribution of goods as the goal of a fair society, Anderson sees distributive justice as the means to fairness and the realization of a system of governance conducted by equals. To accomplish that, Anderson's distributive justice "guarantees all law-abiding citizens effective access to the social conditions of their freedom at all times. It justifies the distributions required to secure this guarantee by appealing to the obligations of citizens in a democratic state."[85]

The conception of distributive justice discussed by Rawls more closely mirrors that of

---

[83] Rawls, *A Theory of Justice* (n 29) 63.
[84] Nozick (n 76).
[85] Elizabeth S Anderson, 'What Is the Point of Equality?' (1999) 109 Ethics 287, 289.

Anderson than Nozick. For Rawls, distributions are the work that just institutions do to ensure equality and fairness. The work of his theory is the design of these institutions, whose role in the lives of societal members is to regulate the opportunities and constraints that construct the "background justice" responsible for ensuring that conflicting and overlapping claims and interests can be adjudicated and resolved in a manner that would be endorsed by all.

3.7 THE BASIC STRUCTURE AS THE PRIMARY SUBJECT OF JUSTICE

Recall that a well-ordered society envisioned by Rawls is governed by a public conception of justice and that an element of that conception are the rules and institutions that coordinate and regulate society to ensure its fairness and stability. This is the "basic structure" of society and it is the first subject of justice. Its design is the main product of the work of the original position. The description of the basic structure evolved throughout Rawls's major works and can be summarized as "a public system of rules"[86] and "the way in which the major social institutions fit together into one system."[87] The basic structure forms the basis of procedures and institutions governing public life in a pluralistic democratic society. For example, a frequently mentioned element is a society's political constitution which sets down the core principles that will guide succeeding rules and decisions. Rawls is otherwise imprecise in how he describes the basic structure, but what emerges is a combination of legal, normative, and other structural constraints and affordances that operate in the background of daily life, including the formation of rules and practices that emerge in public life and within public institutions. The economic system, the political system, and the dominant institutions are all candidate elements of the basic structure.

In the original position, agents design the basic structure but also principles to regulate it, not

[86] Rawls, *A Theory of Justice* (n 29) 84.
[87] Rawls, *Political Liberalism* (n 14) 258.

assuming that it will come together and be sustained without intentional design or ongoing reflection. The stated purpose of the basic structure contains both principles that shape its institutions, and also systems of adjustment that ensure its fairness over time. The principles are general rules to govern all future transactions and distributions that will occur in society. For example, the basic structure is likely to include the design of the institutions through which fundamental rules of social engagement are constrained and afforded. The basic structure also includes regulations designed to adjust for the inequalities in people's prospects that arise over time and that threaten "background justice." The adjustments provided by regulation ensure that the system of cooperation set up by the basic structure remains fair. By including both principles and adjustment strategies, Rawls tasks the basic structure with not only being the subject of pure procedural justice but also an origin for applying just procedures to the task of adjudicating future claims. In short, the basic structure that emerges from the original position is envisioned as a fair system that also contains the necessary elements to sustain it.

3.7.1 Two Principles of Justice

If this description of the basic structure sounds vague it is because Rawls does not elaborate on many of its specifications. Presumably, Rawls was more committed to the process of creating the basic structure than in anticipating its form. However, Rawls does anticipate that participants in the original position would produce two key principles to be carried out within the basic structure to guide all future transactions: the liberty principle and the difference principle.

The liberty principle is fairly straightforward. It reads "each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others."[88] My interpretation of this principle is that everyone is entitled to those liberties that can be enjoyed by everyone.

[88] Rawls, *A Theory of Justice* (n 29) 60.

Similarly, no one is entitled to a liberty that cannot be similarly enjoyed by everyone. In an ideal democracy, eligible citizens enjoy the freedom to participate in their government by voting. Conversely, the freedom to own other human beings is incompatible with the rights of people who are owned, and so is not permitted. Example freedoms offered by Rawls include most of the freedoms guaranteed by the U.S. Constitution, particularly the Bill of Rights, along with property rights and freedom from arbitrary arrest and seizure.

The difference principle states that "social and economic inequalities are to be arranged so that they are both (a) reasonably expected to be to everyone's advantage, and (b) attached to positions and offices open to all."[89] Rawls expands on the meaning of these deceptively simple provisions over several sections of *A Theory of Justice*. The text of the first half of the difference principle forms the basis for what has been characterized as his "maximin" approach to justice; Any advantage that improves the situation for the most well-off must also improve things—*maximize* advantages—for the least well-off. While unequal distributions of goods (including opportunities) are to be expected, they are "just if and only if they work as part of a scheme which improves the expectations of the least advantaged members of society."[90] For Rawls, participants in the original position would choose the difference principle to ensure all members a decent life regardless of their position in society. This scheme roughly mimics the adage "a rising tide lifts all boats," except that, for Rawls, the tide is not an anonymous, arbitrary force. It is the function of "background justice" to ensure that all boats rise, starting with the smallest. This reflects Rawls's rejection of distributions that occur through "natural" means, at least for certain types of goods (more on this below).

[89] ibid.
[90] ibid 75.

The second half of the difference principle is also subtle. First, it seems that Rawls is only arguing that everyone should have access to the same opportunities, such as political offices. However, Rawls expands on the meaning of this principle to further emphasize the necessity of promoting fairness not only at the moment of the original position but also over time. He does this by attacking what he labels the "natural rights principle," which is the idea that certain liberties, including the liberty to acquire and hold property, are inalienable because they arise naturally from the expression of our talents and labor to which we hold a natural right. Rawls resists natural rights doctrines as the primary basis for morality because it may lead to injustice, thereby undermining presumed fairness. For example, a liberal theorist who endorses natural rights, such as Nozick, would hold that distributions of goods acquired and exchanged among individuals without force, fraud, or deception are necessarily fair.[91] Rawls rejects this approach on moral grounds. Rawls instead argues that "natural and social contingencies" that create unequal distributions of resources, which may include "accident and good fortune," are too arbitrary to have moral weight.[92] Furthermore, the conception of the person as *reasonable* means that cooperation is a scheme in which each person benefits with others. Even while some people may work harder or be more talented in what they do, acting in such a way as to deny a minimum share of the benefits to the unlucky does not fulfill the conception of reasonableness.[93] Recall also that an overarching goal for justice as fairness is *stability.* Among the risks to stability indicated by Rawls, particularly in his later works, is progressive concentration of wealth into few hands over time, leading to

[91] Nozick (n 76).

[92] Rawls, *A Theory of Justice* (n 29) 72.

[93] Note that Rawls does not argue that those who simply refuse to contribute, even if they could, are eligible for the benefits of the difference principle. Indeed, with his emphasis on cooperation as the basis of society, it is unclear what Rawls would prescribe for the indolent. Similarly, as discussed in a subsequent section, Rawls limits direct application of the principles of justice to "moral persons," setting aside those who cannot contribute. As a result, and not controversially, such persons are not portrayed as direct beneficiaries of the difference principle.

political domination and other destabilizing effects on society. The basic structure is intended to include provisions that could prevent this. These include equal access to education and the ideal of public reason.[94]

3.7.2 Fairness and Stability

When a society is regulated by the basic structure, including a combination of rules and institutions developed fairly, then just transactions that take place within that society should lead to just outcomes. Here again, Rawls departs from Nozick by demanding a robust and complex regulatory state to ensure that background justice obtains. Where Nozick would be satisfied with an initial condition that was fair to determine the fairness of subsequent transactions, Rawls emphasizes instead that results of nominally fair transactions over time might still create conditions that are unfair because transactants cannot reliably predict the effects of their transactions. "Individuals and associations cannot comprehend the ramifications of their particular actions viewed collectively, nor can they be expected to foresee future circumstances that shape and transform present tendencies."[95] The basic structure is intended to produce and maintain background justice and thereby ensure that not only are particular transactions fair, but also that the conditions that give rise to transactions have also been arrived at under the right circumstances.

There are two features of this construction. The "future circumstances" language clearly aligns with Kant's rejection of consequentialism in favor of the constructivism expressed by the categorical imperative. A Kantian agent cannot predict with certainty the consequences of her actions and should therefore focus on choosing rational processes for deciding action for the best chance of projecting their intentions.[96] Second, Rawls suggests that transactions that might have

[94] Rawls, *Justice as Fairness* (n 33).
[95] Rawls, *Political Liberalism* (n14) 268.
[96] Herbert James Paton, *Groundwork of the Metaphysics of Morals* (Harper & Rowe 1964).

been fair at one point in time cannot be presumed to be so at another point in time. This elides with Rawls's emphasis on "stability" as a cornerstone to a just society. Justice as fairness is not a fixed societal framework but one that is dynamic and subject to adjustment and compensation over time. Background justice is not the assumption that a single set of properly constructed processes will produce a legacy of fairness over generations. Background justice is achieved by a basic structure that includes processes for revisiting and revising its content over time.

This accords with the Rawlsian conception of justice as an ongoing site of negotiation rather than a fixed set of declarations. For Rawls, even if the initial conditions of the basic structure are fair, and even if the individual agreements that take place within the constraints and affordances of the basic structure are ostensibly fair, justice demands that the basic structure include institutions and processes for promoting residual fairness over time:

> Even though the initial state may have been just, and subsequent social conditions may also have been just for some time, the accumulated results of many separate and seemingly fair agreements entered into by individuals and associations are likely over an extended period of time to undermine the background conditions required for free and fair agreements. Very considerable wealth and property may accumulate in few hands, and these concentrations are likely to undermine fair equality of opportunity, the fair value of political liberties, and so on.[97]

Notably, having been somewhat oblique in about which harms he was concerned about in his earlier works, in his later work, Rawls begins to indicate the kinds of unfair transactions he is concerned about. This represents a slight deviation from his apparent desire to emphasize the procedural aspects of justice rather than attempting to regulate outcomes. Here, Rawls indicates that concentrations of wealth are a particular problem for justice as fairness and a specific threat to the two principles of justice.

This evolution in Rawls's sense of fairness begins to accord with that of some of his critics.

[97] Rawls, *Justice as Fairness* (n 33) 53.

For example, G.A. Cohen argues that we must attend to the outcomes of any system of justice rather than remaining overly focused on procedures or initial conditions. Cohen argues that while societal members may find the terms of particular transactions free and fair, they should concern themselves as well with what occurs beyond the parties to the transaction itself or what happens as a result of aggregate transactions. For Cohen, factors other than the fairness of a particular transaction have significant normative weight in determining a just and fair process.[98] Elizabeth Anderson similarly argues that egalitarianism justice should be characterized by a concern for human dignity and the role of society in dismantling oppressive hierarchies rather than concerning itself purely with distributive rules and processes.[99] Rawls's model of society can usefully explain the moral framework of reputation while its critics offer clues about how that model, as reflected in real life, is a potential tool of social hierarchy and oppression.

3.8 POLITICAL CONCEPTION OF JUSTICE AND COMPREHENSIVE DOCTRINES

The well-ordered society envisioned by Rawls, once it has been set up, requires a second stage of conceptualization to include concepts and strategies to ensure its stability and endurance over time. Stability is challenged by the "reasonable pluralism" of this society;[100] society members are not monolithic and hold conflicting conceptions of the good based in religious, philosophical, and other worldviews. Rawls argues that the stability of a fair system of cooperation requires making space for this diversity of conceptions. However, society members hold two views of the world; an individual concept of the good informed by a "comprehensive doctrine," and a political view that emerges from the social contract and concerns the regulation of public life. Because a

[98] G.A. Cohen, *Self-Ownership, Freedom, and Equality* (Cambridge University Press ; Maison des sciences de l'homme 1995).

[99] Anderson (n 85).

[100] Rawls, *Political Liberalism* (n 14) 144.

multiplicity of comprehensive doctrines is assumed in a well-ordered society, reasonable pluralism is an essential aspect of justice as fairness. In embracing reasonable pluralism, Rawls departs from theories of justice that endorse a single reasonable and rational good, including that found in the works of Plato and Aristotle, in the Christian ethics espoused by Augustine and Aquinas, and in the classic utilitarianism of Bentham. These ethics endorse a view of their being one single good and provide strategies for identifying it. For example, utilitarianism identifies a single good as one that emerges from the calculation of net utility. Instead, Rawls argues that there will always be multiple conceptions of the good and thus multiple comprehensive doctrines which must somehow function together.

A comprehensive doctrine is a rational conception of the good consisting of religious, philosophical, and moral views that are held by individuals.[101] A comprehensive doctrine may be shared with others, but in a society of reasonable pluralism there are many such doctrines and no single one legitimately governs the lives of everyone. The plurality of comprehensive doctrines cannot be too much in conflict if a society is to be stable enough to fulfill the aims of mutually beneficial cooperation. Social stability occurs through the "overlapping consensus" of comprehensive doctrines in which societal members find they have enough in common with each other to jointly pursue their essential interests.[102] Overlapping consensus is not merely a compromise among competing interests in which no one is particularly satisfied. Overlapping consensus is indeed a *consensus;* members of a well-ordered society endorse its values because there is space in their *reasonable* comprehensive doctrines for the overlap necessary to bind them to one another. For example, a reasonable religious doctrine may lead religious members of a

[101] ibid 140.
[102] ibid 134.

society to reject incest while non-religious members committed to the teachings of biology may reject incest with similar fervor because of the risks of incestual parentage. The result is the same; incest would not be endorsed by a society with these overlapping but distinct comprehensive doctrines. Here we can see how even if political power were to be dramatically redistributed in this society, so long as an overlapping consensus continued to endorse certain principles, the society would continue to endorse them as well.

Many features of the basic structure can be sustained by this overlapping consensus, but overlapping consensus alone is not sufficient to account for every situation faced by a well-ordered society. While the original position may provide a basis for future agreements, there are likely to be issues that require deliberation and the reach of a power greater than simple agreement among members of society to achieve. To achieve this larger government as basic structure requires that a type of pure procedural justice would need to extend beyond the original position into the life of a society to adapt and adjust it to meet the conditions that arise.

3.9 PUBLIC REASON

Public reason is an important concept in Rawls's later work, which is concerned with promoting the stability of a well-ordered society over time. Central to stability is the task of ensuring that the application of political power be accepted as *legitimate* by members of society. Public reason is presented as the requirement that the rules that regulate the lives of members of a well-ordered society be justifiable to them so that they will see them as legitimate and thereby continue to endorse the society.[103]

As I stated in section 3.8, members of a well-ordered society hold, in addition to a comprehensive view, a *political* view of the world. This gives rise to a political conception of

[103] Edward N Zalta and others, 'Stanford Encyclopedia of Philosophy' 29.

society. The political conception offered by Rawls concerns those aspects of social organization that directly link to the basic structure of society. That means the core institutions and principles construct the scope of people's liberty and equality. Important features of this political conception are that it operates separately from those domains of life governed by comprehensive doctrines, and also that it requires a specific and elevated form of inquiry and deliberation. These distinctions are important because of the nature and effect of political *power*.

Recall that Rawls builds his theory on Kant who sought to idealize human dignity as directly tied to a person's ability to express their will in the public sphere. Whereas comprehensive doctrines apply to personal views and the rules of associations that are generally perceived as voluntary, such as membership in a church or one's commitment to a particular philosophical position, another doctrinal construction is required to guide public life. Our relationships to certain types of institutions are *not* voluntary, such as the relation to the state and the elements of the basic structure of society. As Rawls frequently reminds us, the basic structure includes "institutions we enter only by birth and exit only by death."[104] These institutions exercise *political* power over society's members and that power is coercive; it is backed by the state and its ability to mete out punishments and other elements of the enforcement of its laws.

3.9.1 Principle of Legitimacy

A key problem for the political conception of the person in the type of democratic society envisioned by Rawls is the *legitimacy* of this coercive form of power, particularly given Rawls's assumption that this society is *pluralistic,* and members have many, conflicting, conceptions of the good. In a just and pluralistic society, the rules and institutions that guide people's actions and opportunities are enacted for reasons, but the nature of those reasons differs according to who is

[104] Rawls, *Political Liberalism* (n 14) 136.

enacting them and the scope and effects of their application. Here, Rawls offers an important distinction; that between "public" and "non-public" reason.[105]

Non-public reasons are sufficient for constructing rules based in the comprehensive doctrines of either individuals or their voluntary associations.[106] Such rules are only binding to members of voluntary associations or for guiding individual pursuits. While it may be sufficient to establish rules or to dictate people's choices within associations based on non-public reasons, such reasons would not be seen as legitimate to guide the application of political power over those who do not subscribe to the comprehensive doctrine of a particular individual or association. Non-public reasons are likely to come into unresolvable conflict with many members of a pluralistic society of free and equal persons. Therefore, Rawls argues that legitimate state authority based in reciprocal agreement is achieved only when its rules and institutions are subject to processes of open inquiry and justification that produces reasons that everyone can minimally endorse.[107] Arguments from public reason are open to deliberation and debate. They speak across the diversity of comprehensive doctrines to appeal to the overlapping consensus of those doctrines where the shared endorsement of the basic structure lies.

Recall that Rawls accepts that, under reasonable pluralism, there is a multitude of comprehensive doctrines that coexist among reasonable people, but there is no single comprehensive doctrine that is appropriate to guide a deliberative democratic society.[108] Any non-

---

[105] ibid 213.

[106] Rawls argues that there are also other forms of reason, such as social reason and domestic reason, that guide people's small-group, family, and personal lives. For Rawls, these domains are not subject to either public or non-public reason as he constructs them (see Rawls, *Political Liberalism* (n 14), p.220 footnote). While he claims that principles of equal citizenship apply in these domains, it is unclear what ethics or conception of justice is supposed to guide their reasoning and some critics have suggested that this construction leaves many people and situations dangerously exposed to unaccountable oppression and harm.

[107] John Rawls, *Justice as Fairness: A Restatement* (n 33) §9.

[108] Rawls, *Political Liberalism* (n 14).

public association, such as a church, a company, or a club may enact rules that members must follow to continue their association, but those rules are understood to apply only to the members of the association and cannot be reasonably applied to non-members. For example, free and equal peoples could not accept a single religious conception to dictate the conduct of all people, though many may be guided in their daily lives by that conception. Rawls argues that reasonable people who desire to cooperate for mutual advantage can accept a diversity of beliefs, and even have them enter into public discourse, so long as they are not coercive to those with different comprehensive doctrines. So, while one's religious doctrine may be offered as a justification for a position in public discourse, it can only be *offered to* the deliberative process; a process in which other views are also heard and considered and either accepted or rejected in democratic deliberations.

Public reason, meanwhile, is the moral basis of a democratic society based in egalitarian principles. By being independent of comprehensive doctrines, public reason is an attempt to get at a *neutral* conception of the good, one that is moral because it demonstrates respect for society members as deserving of equal consideration, rather than being subject to the will of others. Public reason is the "reason of its citizens, of those sharing the status of equal citizenship."[109] Society members have a moral duty to explain to one another the principles and policies that are to be imposed upon them as a cooperative body in a procedure of fair and open deliberation. This is because such principles and policies have a direct relationship to the degree of liberty that the basic structure provides. Where the stakes are high, public reason must prevail, because "whether men are free is determined by the rights and duties established by the major institutions of society."[110] While non-public reason is sufficient to justify the terms and conditions of voluntary associations,

[109] ibid 217.
[110] Rawls, *A Theory of Justice* (n 29) 63.

public reason is essential wherever matters of basic liberty are at stake.

3.9.2 Coercion

A key argument from Rawls about the necessity of public reason is based in the coercive power of a well-ordered society through its basic structure of rules and institutions. The elements of the basic structure have totalizing effects they have on people's lives. While it is a public system of rules that applies to everyone and that everyone agrees to be bound by them, the basic structure *governs.* Unlike the power of a club or a church in a democratic society, state power, even if applied through a system of rules developed through a process of pure procedural justice, is inescapable. We cannot evade a government's power without leaving its territory, an onerous demand. Leaving may even be prevented by a government through its authority to restrict the movement of citizens and control its borders. Where free movement is not prevented by an authority, leaving one's home, social networks, family, etc., are strong incentives to remain in place. Even for those who are able to pack up and leave, there are few choices but to enter another governed territory, becoming subject to the coercive authority of that government.

While the power of some associations can be extensive (*e.g.* religious orders, oppressive family structures, professional associations), they are, in theory at least, voluntary and we can exit them.[111]

The social contract outlined by Rawls is hypothetical in any case. We employ the device of the contract as a thought experiment that permits reasoned reflection about our actual conditions. Rawls acknowledges that, in reality, we have simply appeared in the state as it is without having

[111] The voluntariness of many associations that appear voluntary is by no means certain. A religious order may appear to be voluntary, for example, but living in a town dominated by a megachurch in which virtually every resident is a member would make participation much less voluntary than it might seem to be from a distance. However, while we can find similar examples of involuntariness in seemingly voluntary associations, there are certainly many that are indeed voluntary.

negotiated or chosen its terms. This too indicates the involuntary nature of state authority. So why should members of a democratic society accept this? Rawls answers that for this relationship to be just and democratic, it must meet a standard of legitimacy. Legitimacy in a Rawlsian society emerges from two locations. First, if the society has been constructed based on pure procedural justice and based in just principles, its members should view the society as reflecting their will and fulfilling their aims. Second, when the members of this society are able to justify to one another the rules and decisions that have the force of law based in reasoning that appeals to the most widely shared conceptions of the good, the exercise of this power does not feel oppressive. Rawls frames political power under these circumstances as reflecting the will of the body politic rather than acting upon them: "… in a democracy political power, which is always coercive power, is the power of the public, that is, of free and equal citizens as a collective body."[112]

The coercive nature of political power provides some guidance for differentiating the appropriate scope of such power. Consider a voluntary relationship, such as one's relationship to a car rental company, chosen among many competitors. A company can impose its own rules on customers and demand compliance, but only up to a point. Failing to return cars in good condition can lead the company to block a customer from future rentals, but it cannot prevent *all* rentals with other companies or prevent the customer from driving cars privately owned. Contrast this with the power of an institution from the basic structure; a court of law can suspend driver's license for failing to pay fines and prevent her from (legally) driving any car anywhere. The totalizing reach of the basic structure into the lives of people could be oppressive unless the principles and policies in effect can appeal to the broadest base of democratic ideals shared by citizens, without exclusive fealty to any particular comprehensive doctrine shared by only a portion. So long as society

[112] Rawls, *Political Liberalism* (n 14) 216.

members agree that there should be an authority that regulates drivers and that authority employs public reasons in applying its regulatory might, the exercise of this power is legitimate. This construction of legitimacy in the exercise of coercive political power maintains the stability of the basic structure and its institutions over time.

3.9.3 Limits to the Domain of Public Reason

A key difficulty in the idea of public reason is defining its scope of application. Because of reasonable pluralism in Rawlsian society, there are many domains of life in which people are satisfied by rules and decisions that are non-public. Devout Jews are satisfied with allowing scripture to guide their dietary habits and rest days and, in a pluralistic society, are not too concerned with enforcing their restrictions on others.[113] If Jewish dietary laws and sabbath days were enforced using coercive political power, non-Jews and non-observant Jews would not see this power as legitimate. Here, too, we see that legitimacy emerges not only from the construction of the basic structure and its ongoing public reason but from the choices in where such power should or should be employed.

Rawls seeks to answer this challenge by limiting the domain of public reason to "constitutional essentials" and "matters of basic justice."[114] Recall that a key element of the basic structure of society is that it contains a constitution to organize its political system and specify the "equal basic rights and liberties of citizens."[115] The creation of that constitution is envisioned to take place in the original position, but like all elements of the basic structure, it is subject to adaptations and adjustments and also interpretations of its provisions to ensure that is produces

[113] In practice, this may only be generally true. In certain neighborhoods in New York City, for example, businesses that sell non-kosher foods or that operate on the sabbath could face significant resistance. Presumably, the neighbors are fine with such activities taking place elsewhere and generally endorse democracy, but not in their neighborhood.

[114] Rawls, *Political Liberalism* (n 14) 215.

[115] ibid 228.

fairness over time. By example, Rawls offers the Supreme Court as an example of a body that employs public reason. Because the Supreme Court is tasked specifically with interpreting the constitution of a democratic society with significant effect on the basic rights and freedoms of its citizens, its deliberations must be justifiable to those citizens through public reason. Supreme Court deliberations are therefore constitutional essentials.

Constitutional essentials that require public reason include those pertaining to the general structure of government and the political process as well as the equal basic rights and liberties of citizens. However, what counts as a constitutional essential is quite narrow. While decisions about certain basic rights, such as whether to grant freedom of movement to citizens, would be subject to public reason, fair equality of opportunity is not a constitutional essential. It is, however, a matter of basic justice.

Basic justice is given secondary priority for the application of public reason. It concerns matters associated with distributive justice, such as the allocation of important resources, including income, wealth, and opportunities. However, Rawls also limits exactly how far into the realm of basic justice public reason should apply, preferring to leave many allocative questions to the less rigorous standards of the legislative process, a venue in which compromises and suboptimal outcomes are to be expected (imperfect procedural justice). Rawls justifies this lowered priority because matters of distribution are an often-controversial feature of political society. Deciding how to allocate funding for science education programs is one example. While there may be a common good achieved by allocating such funds through a process of public reason that produces the ideal outcome rather than a compromise that pleases various comprehensive doctrines, Rawls appeals to a sort of efficiency, or what the philosopher Jonathan Quong labels a "priority argument" [116].

[116] Jonathan Quong, 'The Scope of Public Reason' (2004) 52 Political Studies 233, 235.

Rawls wants public reason to be required only in a sufficiently limited number of domains to ensure that matters of "fundamental justice" are attended to.[117] If public reason is required for too large a domain of inquiry and deliberation, there is a chance that constitutional essentials and the most fundamental matters of basic justice will not be resolved. This might be especially true if public reason is attempted in matters where universal agreement is impossible, such as in questions about how much to tax wealthy families. It is unlikely to find universal, or even near-universal agreement on matters of taxation; attempts to employ the rigors of public reason to such deliberations would likely go nowhere.

Even with these conditions, Rawls is at pains to specify exactly where public reason *does* apply, giving only a narrow scope while gesturing toward a larger domain: "my aim is to consider first the strongest case where the political questions concern the most fundamental matters. … Should they hold here, we can then proceed to other cases"[118]. Rawls never gets to a point of explaining these other cases, but in often repeating the importance of reasonable pluralism and the many and conflicting conceptions of the good in society, I conclude that, in addition to his efficiency goals and concerns about reaching any sort of full agreement on some matters, Rawls is seeking to tread carefully on the domains of life he labels as non-political, such as the family, religious institutions, and to some extent, the institution of private property. The latter are particularly important to Rawlsian society because religious freedom is considered a core value of liberty.

[117] Rawls, *Justice as Fairness* (n 33) 152.
[118] (Rawls, 1993/2005 (n 106) 215)

3.10 WIDENING THE SCOPE OF PUBLIC REASON

This narrow scope for public reason appears important to the structure of Rawlsian theory but it is not clear that we must accept Rawls's limitations even while employing a Rawlsian conception of justice. It seems plausible that limiting public reason to Rawls's narrow accounts of constitutional essentials and matters of basic justice might lead to situations that threaten the kind of egalitarian fairness Rawls endorsed. I suggest that rather than hew to doctrinaire categories and limit public reason to them, we should instead consider what constitutes a set of fundamental matters worth preserving by applying the rigor of public reason. If the goal of public reason is to assure that society interjects its most democratic processes in matters that affect people's fundamental rights and interests, then we may find some beyond what qualifies as constitutional essentials and matters of basic justice as described by Rawls. One way to approach this is to consider public reason as the appropriate approach to the allocation of primary goods, and similarly, severe sanctions. For Rawls, most goods are allocated by the institutions of the basic structure, which he appears to limit as composed of state entities and doctrines. However, in contemporary society, private institutions play an increasing role in the distribution of primary goods and sanctions. Where firms do the work of distribution of such goods and bads, why should we exclude them from our conception of the basic structure of society? If they can be considered thus, they become subject to the principles of justice arrived at in the original position, and ultimately, to the domain of public reason in pursuit of stability.

Quong[119] similarly offers a more expansive view of the domain of public reason, finding Rawls's desire to limit public reason to constitutional essentials and matters of basic justice too

[119] Quong, 'Scope of Public Reason' (n 116) 215.

strict.[120] Instead, Quong proposes "the broad view" of public reason, which holds that "public reason ought to be applied, whenever possible, to all political decisions where citizens exercise coercive power over one another," rejecting Rawls's limitations as overly strict and likely to result in domination in a variety of political matters.[121] Quong argues that Rawls's conception of public reason, which he labels "the narrow view," does not provide the basis for a just society. Where Rawls seeks to limit the scope of public reason to demonstrate respect for the reasonable pluralism of comprehensive doctrines within society, Quong cautions that there are many decisions that may appear to be beyond the scope of public reason and yet are based in values we can confidently claim as political values, meaning they are not particularly subjective or expressive only of a particular comprehensive doctrine. Quong argues that there are many decisions that might at first appear to fall outside the realm of public reason but may be better resolved that way.[122] For example, a town may be struggling with a decision about whether to allocate funding for its second sports stadium or its first aquarium. This would at first appear to be a matter of competing desires and cultural values, and it partly is. However, upon closer inspection, there may be a decision based in evidence that appeals to everyone. Many may prefer either a sports stadium or an aquarium for non-public reasons, such as a strong passion for hockey over marine biology. However, it is possible to recognize that the non-public values both have merit while at the same time considering more objective evidence of collective benefit, such as if there is convincing evidence that one of the options is likely to produce greater economic benefits or require less public investment. While there is no guarantee that any particular evidence will convince everyone, upon

[120] Public reason may seem another Rawlsian concept that demands too much of us. However, recall that Rawlsian theory is *ideal* theory. We cannot say if public reason leading to general agreement is truly possible. What matters is that our reasoned reflections can hypothesize public reason as a means to decision-making and then pursue equilibrium with achievable conditions.

[121] Quong, 'Scope of Public Reason' (n 116) 234.

[122] ibid.

careful deliberation in good faith by the townspeople, it may appear a more logical and objective choice to favor the aquarium through a process that is based in public reason.

I take this point further and suggest that another reason for broadening the scope of public reason is that of necessity. In the ideal society envisioned by Rawls, many actors capable of exerting coercive power over members of society are absent. While government agencies, courts, and regulators are referred to in various ways, powerful and influential private actors, such as multinational corporations and industries, can have extraordinary and inescapable power over people's lives, are not. Rawls, by seeking to limit the domain of public reason sought to ensure that matters of fundamental justice would be attended to. But Rawls also argued that matters of fundamental justice are directly tied to the coercive nature of political power, which must be held to be legitimate in a just society. As I argue in subsequent sections, there are matters of fundamental justice at stake in the exercise of power of entities that Rawls did not include in his political conception, among them, the power to make assessments about a person's worth that follow them through their lives and affect their material and existential well-being.

To push on the boundaries of Rawls's limits to public reason, I suggest we expand the definition of "political power" to include non-governmental power that is coercive and inescapable. As Rawls tells us, where political power is to be applied by and for people who recognize each other as deserving of equal consideration, it must be democratically legitimate power. That can include forms of power not discussed by Rawls but present in many domains of life, including those mediated by information systems and practices. From whatever source it emerges, power that is coercive is by definition oppressive when it is applied to the lives and choices of people and groups through rules, decisions, and institutional structures, and it must be at all times justifiable and explainable to those affected by them in a manner that allows for public

deliberation, debate, and collective decision-making.

3.11 RAWLS'S IDEAL THEORY AND ITS DISCONTENTS

In this project, I employ the political theory of John Rawls as a means of examining the justice implications of reputational systems, including digitally-mediated consumer/citizen profiling and decision-making systems. I acknowledge the considerable criticism Rawls's work has attracted since its inception. This includes general critiques of egalitarian social contract theory from both within and without the domain of liberal political theory, and specific critiques from critical race, disability, and feminist perspectives. In particular four core aspects of Rawls's theory seem to be targets of critics: First, is Rawls's idea of the "well-ordered society," which is described in 3.3. Next is the "original position," which I describe in 3.5. Third is the distribution of "primary social goods," which I describe in 3.6. And fourth, the "difference principle," which I describe in 3.7. I briefly summarize the major critiques here and discuss why I employ aspects of Rawlsian theory in spite of these critiques and the theory's limitations.

3.11.1 The Problem of Ideal Theory

The theory of justice that Rawls developed and refined through his life is a theory built of model conceptions that describe the building blocks and core attributes of an ideal society.[123] Justice as fairness is not intended to reflect the world as it is. Rawlsian theory assumes that we, as rational and reasonable, engage in "reflective equilibrium" to essentially compare our conditions with those that would be arrived at under ideal conditions that are fair. To the extent we are able, we are to adjust our present conditions and resolve its conflicts so they reflect these ideals. However, justice as fairness is not a *theory of change* to apply to existing conditions but rather a

[123] Anna Hoffmann, 'Google Books as Infrastructure of in/Justice: Towards a Sociotechnical Account of Rawlsian Justice, Information, and Technology' (University of Wisconsin 2014) 27.

theory of ideals. It provides us with a framework to instruct our thinking and aim it toward a commitment to fair procedures. For this model to be clear, Rawls uses an artificially constrained model of the world—leaving out some of its real and confounding features in order to focus on perfectly arranged scenarios that can foreground the theory. As a result, Rawls's theory is not offered as a solution to our lived political and social problems per se.

Some of Rawls's prominent critics have argued that a theory of justice that is not wedded to actual lived experience is unlikely to be able to answer basic questions about existing forms of injustice. For example, critical race theorist and prominent Rawls critic Charles Mills argues that Rawls's theory cannot the address the roots and causes of contemporary injustice, such as the legacies of colonialism and slavery in the western world, because these would never enter into the ideal society constructed by Rawls through which he articulates his theory of justice.[124] As Mills argues, Rawls constructs an ideal society that is apparently free of war, occupation, and profound poverty, which are conditions that cannot be assumed in the founding of a real society. By design, inconvenient historical events like slavery, genocide, and oppressive distributions of fundamental resources are not discussed, even though the closest living approximation to Rawls's well-ordered society, the United States, has been shaped by these things.[125] Iris Marion Young similarly argues that a pursuit of justice that makes no attempt to confront, let alone resolve, the historical forces that produced current conditions is unlikely to offer a meaningful diagnosis of those conditions.[126]

This criticism flows into the original position and the veil of ignorance, which are designed to create a negotiating environment in which participants know so little about themselves that they cannot bargain for their own narrow interests and instead single-mindedly focused on constructing

[124] Charles W Mills, 'Decolonizing Western Political Philosophy' (2015) 37 New Political Science 1.

[125] While it is possible that historical contingencies might be accounted for in the original position, Rawls offers no evidence that such considerations would be among those allowed into consciousness by the veil of ignorance.

[126] Iris Marion Young, *Justice and the Politics of Difference* (Princeton University Press 1990).

a system for cooperating for mutual advantage. The original position is intended to show us what a depersonalized vision of fairness could achieve. But there are problems with a vision in which suffering of any kind is intentionally obscured from the negotiations that give rise to a society. At Rawls's idealized bargaining table, participants are unaware of external military threats, despotic rulers, mob violence, natural or manmade crises, impending starvation, brutal subjugation, or any other significant traumas or constraints that have been instrumental to the formation of every country that now exists. The wealth and military might of the most prominent countries in the world, including those whose democratic principles most closely resemble those that Rawls believes would be present in a well-ordered society, are the historical result of long periods of violence and dispossession. While it is not necessarily true that a society resembling the one envisioned by Rawls could not come into existence without such violence, it is difficult to identify where else the necessary raw materials to build a new society without external interference would come from. Mills argues that Rawls scaffolds an entire political theory on a society that not only does not exist, but *could not* exist.[127] While I believe Rawls's intent was to describe an ideal society only as a reference point for adapting existing societies toward greater fairness, in erasing actual historical reality, he risks aiming the work of justice at a target too ethereal to even approach, much less actually reach.

While this critique highlights important weaknesses in Rawls's original project, it does not convince me that an attempt to use his work to analyze the political justice concerns of contemporary society is a fool's errand. Not all of Rawls's work is limited to ideal conditions, nor is all of his work focused on the distribution (and implied *re*distribution) of goods. In this work, I rely heavily on Rawls's *Political Liberalism*, a text written more than two decades after his *Theory*

[127] Charles W Mills, 'Rawls on Race/Race in Rawls' (2009) 47 The Southern Journal of Philosophy 161.

of Justice. As stated by Freeman[128] the goals of *Political Liberalism* are twofold. One is the concept of "stability" in a just society. Rawls wanted to explore further how a society produced using the procedures and principles of justice as fairness could continue over time to meet the needs and goals of its citizens. The second goal is identifying the sources of "legitimacy" for such a system. The sources of legitimacy for a society include the systems of accountability and responsiveness that enable citizens to feel recognized and respected by a society's governing structures. This goal is interesting because strategies to establish legitimacy can be applied to existing democratic societies that are imperfect yet aiming toward fairness.

Another useful criticism is not leveled at ideal theory per se, or of the work of John Rawls, but emerges from the economist Elinor Ostrom[129] in her examination of the problem of managing natural resources held in common ("common pool resources"). Ostrom describes various proposals that have been put forward to solve the "tragedy of the commons" and similarly described challenges to productive cooperation among people in the absence of either an external authority or robust private property regimes. Ostrom levels a strong critique against the argument that only the regulatory might of an external authority, such as a government, would prevent cooperators from defecting on agreements to share a resource in pursuit of their own self-interest, Ostrom argues that an external authority, no matter how well-intentioned, cannot be assumed to operate with complete mastery or sufficient reach to successfully enforce its rules. Ostrom writes that such view predicts that unified authorities will operate in in the field as they have been designed to do—determining the best policies based on valid scientific theories and adequate information. Implementation of these policies without error is assumed. Monitoring and sanctioning activities

[128] Freeman, *Rawls* (n 55).

[129] Elinor Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press 1990).

are viewed as routine and non-problematic. Ostrom argues that this is unlikely to be so. Authorities rarely, if ever, have perfect information or perform perfectly.[130] This critique can be applied to Rawlsian theory. Rawls frequently prescribe solutions or procedures based in presumptions about their perfect execution once chosen. For example, if we adopt the difference principle, we are assuming that some societal body can accurately identify whose share of goods ought to be adjusted and is capable of making those adjustments accurately and completely. Based on Ostrom's critique, this seems rather implausible.

Additionally, Rawls's conception of the original position and the basic structure of society is vulnerable to this criticism. In the original position, participants lack fundamental information about themselves, but it is implied that they possess complete information about the resources and constraints of the world their society will inhabit. And providing they possess this knowledge they are also assumed to make no errors in applying it to the construction of basic institutions and the distribution of primary social goods. From Ostrom's more jaundiced perspective, failures of information or execution are likely to lead to unjust outcomes no matter how thoughtfully constructed. This suggests that Rawlsian concepts, such as pure procedural justice, only guarantee just outcomes if practiced by omniscient and all-powerful agents. Similarly, even the rigors of public reason do not guarantee that decisions concerning the political realm will be carried out as intended and enforced as required.

3.11.2 Liberal and Illiberal Critiques

Rawls is frequently associated with a type of soft socialism due to his commitment to distributive justice, the fair distribution of fundamental material and immaterial goods as a requirement for a just society. The distributive aspects of Rawlsian justice are a relatively minor feature of Rawls's

[130] ibid 22.

theory and yet receives a great deal of attention. Critiques of distributive justice based in egalitarian values typically come from critics on the political "right," who generally portray any state action to distribute, or redistribute goods as an unacceptable intervention into the lives of free people. They ask how justice can be achieved by seizing control over goods currently in the hands of those who acquired them, particularly where their acquisition can be portrayed as fair or related to the natural rights of persons.

3.11.3 Critical and Identitarian Perspectives

Rawls has also come under fire from his left by critical race, disability, feminist, and other perspectives. Generally speaking, these critiques are unified in rejecting Rawls's erasure of identity as a requirement for constructing a well-ordered society. Identity, including racial, gender, sexual preference, ethnicity, abledness, and other distinguishing and unchanging human aspects, define not only the person whom they describe but also the particular forms of injustice they face. Mills argues that African Americans experience injustice not simply because they live in societies marked by unjust conditions but because they are African American. Changing the material conditions of society, such as the initial distribution of wealth and political offices, may not be sufficient to overcome the cultural sources of racism, which could diminish the fairness of a society over time no matter how well-ordered. Perhaps even more problematic is Rawls's apparent choice to leave disabled people entirely out of his society.

While envisioning a society constructed from a place of pure procedural justice can show us how our own societies fail to meet that bar, Rawls could have done more to bridge the realm of the ideal to the tangible. In particular, Rawls's failure to address the historical roots of contemporary injustice leaves out important facts about the nature of justice, which leaves us with a conception that lacks any wayfinding to achieve it. As Mills complains about *The Theory of*

Justice, "here is a 600-page book on social justice ... in which no answers are given about the correction of the injustices of the past."[131]

3.11.4 Abledness Critique

Another important critique of Rawls's work is his inattention to the justice claims and interests of those whose cognitive and physical abilities do not meet specific criteria for inclusion as claimants. I refer to this area of criticism as the *abledness critique* because it addresses the perception that Rawls prioritized the application of theory to the rights and interests of those whose cognitive and physical abilities fall within a range considered "able" and "normal" over others who are labeled "disabled" or "abnormal." This criticism of Rawls emerges from certain choices and challenges Rawls acknowledged in his work. First, is Rawls's conception of society. Rawls conceives of society as a fair system of cooperation for mutual advantage. This implies that the primary relationship members of society have to each other is based in assumptions about how they can contribute to each other's projects and goals. A member of society who cannot directly contribute in some tangible way to the project of mutually advantageous cooperation appears to have a diminished, indirect claim to justice in such a society. Justice for people who are not presumed to be reciprocal cooperators is granted through an *extension* of justice rather than an initial grant based upon reciprocal dessert.

Next, Rawls decides early on that a key requirement for the political conception of justice, in which people negotiate the fair terms of cooperation, is that they have the *capacity* to participate in political culture. This accords with Rawls's specificity about who qualifies as a "moral person" in his theory. The moral person must have a capacity for a sense of justice and a capacity to form

[131] Mills, 'Decolonizing Western Political Philosophy' (n 124) 19.

a conception of the good.[132] Those who lack or lose those capacities would appear not to qualify as moral persons in Rawlsian society. In his later work, Rawls responded to criticism about his narrow scope of inclusion and was unable to remedy it. Instead, he argued that people with impairments are of course part of society, but their inclusion as political agents must be set aside in order "to achieve a clear and uncluttered view of … the fundamental question of political justice. …"[133] Here, it seems Rawls argues that it would be inefficient to do anything but postpone consideration of impaired persons in favor of getting down to the business of setting the terms of cooperation with those considered able and willing to do the work. Disability rights advocates have pointed out that this view is offensive to and dismissive of those who do not meet Rawls's criteria. They appear as "clutter" that obscures the ideal path toward justice.

As the philosopher Martha Nussbaum[134] argues, the problem of including disabled people in a conception of justice is not unique to Rawls and plagues contractarian views of society dating to Thomas Hobbes, the seventeenth century English philosopher. This is because in most accounts, the terms of society are typically worked out by the same people who experience it, or (as in Rawls's later work) by representatives of those people. Contracting parties are presumed to be at least roughly equal to each other, not only morally, but possessing a "rough equality of powers and resources."[135] While Nussbaum suggests that the veil of ignorance is a better strategy than most in attempting to strip away most of the hierarchical influences, she argues that it still only admits those who enjoy a very similar standard of natural advantages. It would certainly be difficult for someone whose cognitive profile is outside of a given range to be shielded from experiencing

[132] Rawls, 1971/2005 (n 56), § 77.
[133] Rawls, *Political Liberalism* (n 14) 20.
[134] Martha Craven Nussbaum, *Frontiers of Justice: Disability, Nationality, Species Membership* (First Harvard University Press paperback edition, The Belknap Press of Harvard University Press 2007).
[135] ibid 29.

their difference in natural capacities, or potentially of even being capable of participating. That would disqualify them from being considered legitimate users of the veil of ignorance and thereby block their eligibility to participate in the original position. Following the traditional contractarian requirements for participation, this also blocks their full consideration as clients of the social contract.

One reason Rawls's exclusion of disability from his conception of justice matters is because disabled people are especially vulnerable to *in*justice. Nussbaum points out that people with cognitive and physical impairments have only recently been understood to deserve more than the most minimal of care. Even more recently they have come to be recognized as being capable of making personal and political choices. Even the relatively enlightened view that accords this grant of basic dignity struggles against a trenchant social order that generally excludes disabled people from recognition as full members of society. Disabled people are frequently discounted and set aside as "clutter" or worse. Writing about the experience of oppression based in physical characteristics, Iris Marion Young writes, "when the dominant culture defines some groups as different, as the Other, the members of those groups are imprisoned in their bodies."[136] Disabled people are often conceived through labels that suggest sickness, disfigurement, revulsion, and so on, rather than with labels that suggest citizenship, cooperation, normalcy—the requirements for full recognition in the social contract. While Rawls recognized that disabled people ought to receive justice as an extension of the justice granted to others, Nussbaum argues that the additional step required to extend justice equates with a failure to treat equally those to whom justice is extended to the same extent as those who are the primary recipients of justice.

[136] Young (n 126) 123.

3.11.5 Rawlsian Theory as Aspirational Construct

These critiques, taken together, suggest that Rawls's work demonstrates the challenge of applying ideal theories of justice to real world problems. However, there is still value in employing aspects of Rawls's work to point a way forward for a political philosophy of technology while remaining open to modifications that may remedy its most acute shortcomings. To assume this stance is to argue that a reasonable goal for political philosophy is to identify a starting point for theory that corresponds with the common aims of society's members, remaining hopeful that there are indeed such commonalities. As Rawls himself argued, the point is to identify and articulate "shared notions and principles thought to be already latent in common sense …" and to propose "certain conceptions and principles congenial to its most essential convictions and historical traditions."[137] Rawls offers at least enough material to begin this work. The model conceptions society's basic elements and commitments have particular value for the work of this dissertation. When considered as *aspirational,* or "congenial" to use Rawls term, rather than certain or final, Rawls's notions of decision-making and basic justice resonate with the discourse surrounding existing practices of democratic governance. Aspirations, even while imperfect, are instructive as a way forward from an existing condition, so long as there is a commitment to revisit and refine any particular project of justice. This accords with Rawls's starting point of "pure procedural justice." Everything else in justice as fairness springs from this concept. But even here, recall that Rawls does not commit to an exact specification for the procedure. Instead, he argues that pure procedural justice cannot be prescribed in advance; instead the procedure is judged from the perspective of ends: "The essential feature of pure procedural justice … is that what is just is specified by the procedure, whatever it may be."[138] In other words, even the procedure itself is

[137] Rawls, 'Kantian Constructivism in Moral Theory' (n 59) 518.
[138] Rawls, *Political Liberalism* (n 14) 73.

uncertain unless we are satisfied that its outcomes are just. Pure procedural justice is a design concept from which he predicts a likely path. Everything else is constructed from this. Viewed this way, Rawls's decision not to commit to any particular procedure for achieving justice or for an *a priori* set of moral facts, the core of his proposal seems rather modest. It is the predictions that are bold. The thought experiment of the original position and the principles of justice appear to be monolithic, but they are predictions about what might occur with the right aims in sight. Perhaps Rawls too was merely being aspirational.

By the later part of his career, Rawls confronted the many criticisms of his work with a bit of resignation, acknowledging its limitations by stating frankly that there are some problems "on which justice as fairness may fail."[139] As Rawls argues, his theory is probably inadequate in many areas but could still attend to many important questions of justice. I conclude that even if justice as fairness cannot cover every case of injustice, rather than discard it entirely, we have the option to identify those areas of weakness and augment the theory with revisions and other approaches to achieve desirable ends. So, while I agree that a commitment to justice in the here and now requires that we take the critiques of Rawls seriously, I suspect some of his detractors overstate his commitments and are thereby too quick to discount the value of his work for addressing real-world issues of injustice.

The admission that no single theory or approach may be sufficient to solve all problems of political justice is reflected in the actual conduct of existing societies. Often, in the discourse of American democracy, the articulation of ideals about fair decision-making are employed to demand greater accountability and transparency from policymakers. Interpretation and reinterpretation of the law occurs throughout all three branches of government. Regulatory

[139] ibid 21.

agencies interpret the law in rulemaking; courts parse meaning from ambiguous provisions, and policymakers revisit prior legislation to address unforeseen or undesirable outcomes. While the results are certainly uneven and often imperfect, repeatedly referring to notions of fairness in such processes may actually promote fairer decisions. The systems of government in contemporary democracies are shaped and motivated by certain ideals (*e.g.* free expression, one person one vote) that, whether or to what extent they are realized in practice, are powerful motivators for policymaking and for envisioning social goals. Indeed, the act of comparison between what is and what could be is something Rawls had in mind. While I do not attempt to defend Rawls to his critics, I believe some of his model conceptions are sufficiently robust to scaffold the fundamental arguments I make about the fair uses of reputational systems.

CHAPTER 4 — INSTITUTIONAL REPUTATION

4.1 CHAPTER INTRODUCTION

In Chapter 2, I offered various of definitions of reputation and settled on a portrayal of reputation as procedural. I also offered a perspective on the degree of influence subjects have over their reputations and to the extent to which a reputational process can be said to be objective. Finally, I argued that reputation has "normative force;" it is both an *expression* and an *enforcer* of values. As such, reputational assessments reflect a set of views and priorities; assessors make myriad contextual choices about what matters and what doesn't based in their personal conceptions of the good and informed by the prevailing values of their society. In Chapter 0, I outlined some concepts from John Rawls's political philosophy that I will use to argue that there are aspects of reputation that are matters of political and social justice. In particular, I described how Rawls's theory of justice conceives of a society constructed under ideal conditions by people who *rationally* pursue their own ends while being *reasonably* committed to a society conceived as a system of cooperation that is fair. I described the core principles of such as a society, its basic structure, and the distributive requirements for such a society. I also discussed how in the reasonably pluralistic society envisioned by Rawls there are many individual conceptions of the good, or comprehensive doctrines, that guide people in their lives. However, in order for people to universally agree on the rules and institutions that will guide everyone's lives, a fairer approach than relying on any particular comprehensive doctrine is to employ a *political* conception of justice. Where matters of fundamental justice are concerned, we ought to employ the Rawlsian concept of *public reason*. I suggest that Rawls offers an overly impoverished account of the kinds of decisions that should be subject to public reason, applying it primarily to a subset of decisions taken by state actors.

However, non-state actors also engage in decision-making that impacts the fundamental justice concerns of subjects.

In this chapter, I continue to break down reputation by attempting to narrow the scope of reputation that is subject to a political conception. I argue that there is a class of reputation contexts that are *institutional,* and that pertain to public life and affect the existential and material well-being of subjects to a high degree. I attempt to trace the origins of this class of reputation in contemporary, networked society. First, I seek to narrow the scope of inquiry to institutional reputation. I begin with a descriptive account of this scope of reputation and then move to an account that begins to build a foundation for a normative inquiry into this type of reputation by identifying its role in involuntary and coercive relations between assessors and subjects. I anticipate the work of a subsequent chapter by identifying four values of interest in institutional reputation: autonomy, democracy, justice, and privacy. I follow the normative discussion with an exploration of credit reports and scoring, which exemplifies institutional reputation.

4.2 THE SCOPE OF INSTITUTIONAL REPUTATION

I begin by offering a provisional definition of what I will call *institutional reputation*, which is expanded upon in subsequent sections of this chapter. The term *institution* has many meanings. Here, I use it to describe forms of assessment emerging from the highly organized and concerted work of governmental and non-governmental entities. In general, I defined institutions as collections of entities that wield significant social, economic, epistemic, or political power over individuals or groups. In addition to government, which is constructed of many institutions, I include such non-governmental entities as the systems and actors in control of the financial credit reporting and rating system in the United States as well as other influential, collaborative systems of assigning individual worth. I also include collections of public and/or commercial actors whose

decision-making patterns affect large numbers of private citizens and groups. For example, employers, landlords, insurance companies, religious orders, and labor unions may be institutional actors. Generally speaking, members are only institutions to the extent that they act according to epistemic frameworks shared with others, such as "U.S. health insurance companies," who can be said to operate quite similarly. General Motors Inc. is an institution that acts in concert with other members of the institution "automobile manufacturers." Meanwhile, a local brewpub is only institutional to the extent that its employment practices resemble those of most other restaurants. There are two institutional foci for this dissertation project. First is the institution of employers, by which I mean private and public firms that recruit and hire many job candidates. The second institution is that of technology firms who collect and process information about individual people to produce insights about them to paying customers, such as employers.

Non-institutional actors also produce influential, deeply felt assessments. A key difference is the extent of their influence and its impact on any individual life or group. A group of friends or co-workers can affect the lives of its members, even placing some at risk of harm, but us unlikely to affect the very large number of people who fit into a class of people affected by institutions, such as "job candidates." The reputational context of non-institutional assessment may feel important to the affected person, the scope of the effect is more limited than other contexts that are institutional.

Another contribution to the definition of an institution is the "basic structure" of society described by John Rawls. As discussed in 3.7, the basic structure of society is the systems, doctrines, and distributive frameworks through which governments and other collective entities articulate the terms of mutual cooperation. While people and social connections come and go, often fading over time, institutions, such as governments and many non-governmental institutions,

are enduring and likely to persist long after our own lifetimes.

4.2.1 The Development of Institutional Reputation

The origins of reputation are likely as old as human society, however; reputation as an institutional phenomenon most likely developed alongside the complex social institutions that emerged in more recent history. Participatory democracy and the emergence of complex, multi-actor economic systems are some likely drivers that expanded reputation from a local practice tied to social relationships to something organized to transcend geographic and kinship boundaries.

I look to relatively recent history concerning the institutional practice of reputation to draw conclusions about the origins of this reputational form. It is not my intention to offer a comprehensive anthropological or sociological account of institutional reputation, which is a project that could be a dissertation in itself. Instead, I offer a light-touch, speculative history of institutional reputation to expand upon the relatively scant and ahistorical account of the basic structure described by Rawls to lay a foundation for analyzing the most recent developments into which we have some view. I adapt common narratives of social evolution found in works by a range of authors including Karl Marx, Sigmund Freud, Herbert Marcuse, and Émile Durkheim.[140] A synopsis of this narrative is that early humans, recognizing the benefits of cooperating with others in pursuit shared goals, like acquiring food, chose to band together to work toward mutual survival. Successful groups produced generations of offspring and thereby developed into larger groups. Attempts to make coordinated labor efficient led to the specialization of skills and a need to keep track of who had them. Informal systems of mutual assessment may have emerged to

[140] See for example the discussion of Marxist dialectical materialism in Alexander Rosenberg, *Philosophy of Social Science* (4th ed, Westview Press 2012). It is noted that various "social progress" theories of social development, including those summarized by Rosenberg, are contested. Sociologist Anthony Giddens, for example, emphatically rejects the notion that societies evolved in a stepwise logical fashion. My goal here is to suggest that the development of societies, whether linear and evolutionary or haphazard and the result of domination, provide a basis for a form of reputation that are fundamental to the institutional structures that currently exist.

support this. Such evaluations might be private or conveyed to others through narratives, such as "she is a good hunter" or "he is violent." The common narrative suggests that the human societies that survived into modernity developed settled agrarianism as their system of social and economic organization, and in some locations, this eventually produced systems of mass production. As population size competed with the smooth coordination of people by consensus or tribal order, societies developed systems of organized government, trade, and additional divisions of labor.

Durkheim[141] argues that by following this developmental path, people's reasons for cooperating—the basis of their "solidarity"—also changed. When groups are small and people are well-known to each other, their bond is what Durkheim describes as "mechanical," emerging from strong ties such as kinship.[142] In larger societies, people remain dependent on others, but their bonds are based in practical, more impersonal considerations, leading to what Durkheim labels "organic solidarity."[143] The ties that bind in conditions of organic solidarity are weaker than mechanical solidarity and may require external pressures and third-party collective action to maintain. One approach to this is the establishment of social hierarchies that establish a division of labor to prioritize efficiency, using social pressure and coercion where necessary to preserve stability. Such systems develop logics of rational organization and collective psychologies shaped by political and economic leaders. Among these logics are systems of institutional reputation, reputational systems that attempt to set aside kinship and instead employ systems of reasoning, deduction, and verifiability that appear to be less influenced by personal preferences.

Among the forces that coincide with organic societies are the technologies of long-distance

[141] Émile Durkheim, *Émile Durkheim on The Division of Labor in Society* (George Simpson tr, Macmillan 1933).
[142] ibid.
[143] ibid.

transportation and communication including cargo and human transport by ship.[144] Shipping enabled the creation of networks of communication and trade extending well beyond the immediate or adjacent borders of a trader's own region. Successful trade, like other forms of cooperation, requires traders to have some sense of their trading partners in the form of reputational assessment and norms-enforcement to prevent chaos and loss. This is particularly true where traders engage in repeated transactions that take place under conditions of uncertainty about the intentions of trading partners. Trading at a distance placed pressure on existing forms of reputation based in the mechanical social bonds emerging from kinship, direct knowledge, and regular contact. It might be sufficient in a market stall to maintain casual relations with other locals, but dealing with transactants who are strangers and with whom regular, direct interaction is unlikely, demands a different reputational process.

One such process, which appears to foreground the development of institutional forms of reputation, is documented by the economist Avner Greif[145] who describes an eleventh century reputation system used for long distance market transactions.[146] The Maghribi traders were Jewish migrants from the Arabian Peninsula who settled in North Africa and elsewhere in the Mediterranean where they participated in the region's shipping trade. Despite being geographically separated, the Maghribi established a system for coordinating and policing their trade. The traders sold goods individually along shipping routes using non-Maghribi overseas agents who accepted shipments and then sold them, returning the proceeds (minus commissions) to the traders.

---

[144] Avner Greif, 'Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders' (1989) 49 The Journal of Economic History 857.

[145] ibid; Avner Greif, 'Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition' (1993) 83 The American Economic Review 525.

[146] Greif's account is contested, most notably by Jeremy Edwards and Sheilagh Ogilvie's 1989 reappraisal entitled "Contract enforcement, institutions, and social capital: the Maghribi traders reappraised." Rather than take a side in this controversy, I suggest that Greif's account is, at minimum, *descriptive* of a form of reputation that has developed among concerted commercial actors and has an interesting resemblance to contemporary institutional practice.

Opportunity for cheating by the agents was obvious given their distance from the traders and the absence of enforceable international laws. To mitigate the risk of dishonest agents, the coalition of traders made a pact to jointly ostracize any agent who cheated another trader[147]. The Maghribi also held each other to account, ostracizing any trader from the coalition who continued doing business with an overseas agent believed to have cheated another Maghribi.[148] Some combination of the financial benefits and the Maghribi's pact made it attractive for the overseas agents to deal with the Maghribi and to do so honestly. The Maghribi traders case offers an example of reputation's normative force in action. For the overseas agents, maintaining a good reputation with the network would have had direct material benefits; the reputational assessment of a single trader was not only correlative with continued trade with that trader, it could also set the stage for trading opportunities with other members of the guild. The pact worked. According to evidence collected by Greif, the system functioned for generations and appears to have functioned effectively in disincentivizing cheating. Hence the Maghribi traders constructed a reputational system whose collective enforcement served to incentivize the honest dealings of distant partners.

4.2.2 Reputation and the Modern Information Economy

As seen in the Maghribi case, transportation technologies that could rapidly move goods and people from one place to another made it possible for traders to operate well beyond their local surroundings. However, as described by the communications theorist James Carey,[149] it was the emergence of an information technology, the telegraph, that dramatically altered both trade and custom where it was adopted. The telegraph may have also been a crucial building block for

[147] Greif 'Reputation and Coalitions in Medieval Trade' (n 144).
[148] Greif 'Contract Enforceability and Economic Institutions in Early Trade' (n 145).
[149] James W Carey, *Communication as Culture: Essays on Media and Society* (Unwin Hyman 1989).

modern forms of institutional reputation. I describe this by adapting Carey's work to portray the unique economic history of the United States as an information economy.

European colonizers to what became the United States, elaborating on the models of European countries such as The Netherlands and Great Britain, organized the U.S. economy as a decentralized capitalist country, rejecting feudal and collectivist models. Because the United States ultimately organized as a single country with a very large geographical expanse, the establishment of a truly national economy was challenged by distance. Even with riverboats, canal systems, and railroads, by the time of the U.S. Civil War in the mid-nineteenth century, the United States stretched across thousands of miles requiring days or weeks of travel to get from one place to another. This meant that people and goods moved slowly. This also affected communication; messages could only travel as fast as transportation systems could physically move them.[150]

Carey, speaking of the telegraph, describes how a single information technology emerging in the second half of the nineteenth century erased time and distance, creating new possibilities for trade while disrupting the social fabric of the time:

> "… this instrument altered the spatial and temporal boundaries of human interaction, brought into existence new forms of language as well as new conceptual systems...older forms of language and writing declined, traditional social interactions waned, and the pattern of city-state capitalism that dominated the first half of the nineteenth century was broken up."[151]

The telegraph had a combination of functional constraints and teleological aspects that shaped the kinds of communication that flowed across its wires. Telegraphy was expensive and tied to physical infrastructures that were precarious and required upkeep. The work of converting text to the dots and dashes that could be carried over the wires was a specialized skill and took time to perform. As a result, telegraphy was a communication form available primarily to elites and to

[150] ibid.
[151] ibid 204.

those who happened to live along the railroad routes where its wires were strung and where they could be most easily maintained. Even for these users, the cost of messaging incentivized brevity and concision in communication. Telegraphy ushered a form of communication that was especially terse and stripped down in a style that was essentially the *absence* of style or flourish—in stark contrast to the florid style of communication found in written communications of the day.[152]

The lack of flourish and extensive detail lent itself to a type of reduced narrative that relied on assumptions about the shared worldviews of sender and receiver. Telegrams were best understood by people who shared culture and language norms and who were prepared to breakdown and rebuild meaning from relatively small amounts of information. Reporting on particular people through telegraphy meant relying on perceived expectations about social and moral norms and specific understandings of the social order. In short, they were perfectly suited for constructing reputational profiles. Reputational profiles, after all, are reductionist accounts situated in culture. They can become meaningless when taken too far from their contexts of use.

As described by Carey, the reductionist features of telegraphy had effects on numerous areas of life, including journalism, which had been dominated by a newspaper industry often operated by political ideologues whose forceful opinions were reflected in their coverage. The telegraph enabled local newspapers to publish more recent national and international news, but the stories arrived written in the stripped-down language of the telegraph. This contributed to a form of news writing that did not appear to reflect any particular personality or point of view and may have ushered in the modern idea of "objective" news reporting.

An early adopter of telegraphy was the banking industry. Prior to electronic communication, moving money from one place to another was a complex and high-risk task. Travelers and

[152] Daniel Walker Howe, 'American Victorianism as a Culture' (1975) 27 American Quarterly 507.

businesses might use *circular letters of credit*, a type of banking instrument dating to the Renaissance, which could be used to withdraw funds at banks participating in credit networks.[153] This arrangement, which functioned in the United States until the 1970s, provided travelers and businesses access to funds held by distant institutions without having to transport currency themselves. However, even with such instruments, banking information moved slowly and could be outdated by the time it arrived at an institution where a banking decision needed to be made. With the telegraph, information could be transmitted instantly from one bank to another, opening opportunities for complex financial networks, interbank guarantees, and the mobility of investment capital. Functional barriers prevented regular folks from interfering in these communications, since it required access to the equipment and the ability to speak the language of telegraphy. The telegraph also paved the way for one of the most prominent forms of institutional reputation, credit reporting, which is discussed in 4.5.

The brief histories I have offered of trade and communications networks, particularly the dawn of electronic "instantaneous" networks and their implications for business, suggest that institutional forms of reputation have a parallel history with that of information technologies. This history also suggests that the intertwined development of trade and communications technologies contribute to the power and reach of institutions, including the power to overcome prior constraints of time, distance, and scale. The power of institutions, expressed through their ability to judge others, go to some normative considerations discussed in the next section.

4.3 DEVELOPING A NORMATIVE ACCOUNT OF INSTITUTIONAL REPUTATION

In this section, I lay the groundwork for a normative account of institutional reputation, which

[153] Kent McKeever and Boriana Ditcheva, 'The Circular Letter of Credit' (2006) <http://library.law.columbia.edu/CircularLetterOfCredit/> accessed 20 December 2019.

will be developed in subsequent chapters. In a normative account of some feature or practice, it is common to identify specific moral values that are affected and to describe why those values matter, for example, because of their connection to human flourishing. Moral values are values that, when affected by some actor or practice, reflect on the goodness or badness of those actors or practices.[154] Because the goal of this dissertation is to theorize reputation as an information practice, I follow contemporary philosophers whose work ties together philosophy and human rights, such as Philip Brey,[155] in identifying four key values that are frequently associated with technology. These are autonomy, democracy, justice, and privacy. Institutional reputation has implications for all four. In chapter 6, I argue that each of these values is implicated in a particular class of reputation related to institutions. I do not take up the full normative argument here but instead use this space to outline the epistemic assumptions of institutional reputation, which may cause problems for arguments in favor of letting institutions choose entirely on their own how they may affect these values when producing assessments.

In the remainder of this section, I refine my account of institutional reputation, first by associating institutional reputation with forms of knowledge described by other scholars, including French philosopher and sociologist Jean-François Lyotard and the ancient Greek scholar Aristotle. I argue that institutional reputation gets its power by being associated with an objective, "scientific" view of knowledge, but that we can also view such reputations as situated in more contingent, "narrative" forms of knowledge. I then argue that this distinction matters because institutional reputations are both *coercive* and, for all intents and purposes, *inescapable.*

What we call *knowledge* is often associated with factive, consistent, and uniformly applied

[154] Philip Brey, 'Disclosive Computer Ethics' (2000) 30 SIGCAS Comput. Soc. 10.
[155] ibid.

information. Yet, there is another type of knowledge that is contingent on individual experience and perceptions of the world. Its scope of application is more limited, but it is nevertheless used to guide some of what human beings think and do. This latter type of knowledge may be shared by groups of individuals who share a cultural milieu but is not necessarily shared by others. Lyotard labels the factive and consistent form of knowledge *scientific knowledge* and contingent, context-dependent knowledge *narrative knowledge.*[156] Narrative knowledge is most familiar to us as it pertains to the realm of the social. It is a form of knowledge situated in custom and social context, which lies closer to aesthetics than to the articulations of facts, even if facts are present. For Lyotard, narrative knowledge resonates deeply within us; it is not necessarily tied to formal learning. It is socially and culturally situated, less about knowing *what* and more about *knowing how* and *knowing of.*[157] Recall that reputation is presented here not merely as a collection of facts but as the synthesized interpretation of information. A reputation may contain factual information but its integration into human understanding relies on the belief systems of assessors and the consumers of a reputational assessment. Narrative knowledge reflects synthesis. It is embedded in, and expressive of culture. It mainly falls into the subjective and personal and is considered insufficient for contemporary science or other formal systems of knowledge. But narrative knowledge is a form a knowledge that is contextually meaningful. It is an expression of situated human reason and is an important component of human interaction, relationship building, and experience. When producing and exchanging narrative knowledge, we offer opinions, beliefs, tastes, and other perspectives on the world. These are generally accepted by others as subjective, being authoritative only to the extent that others have some reason to trust or cherish our views

[156] Jean-François Lyotard, *The Postmodern Condition: A Report on Knowledge* (Geoff Bennington and Brian Massumi trs, University of Minnesota Press 1984).
[157] ibid.

without relying on them as final truth. Narrative knowledge guides action. Narrative knowledge, while contingent, is nevertheless part of human rationality. Human beings integrate the facts and beliefs available to or held by them into plans and actions.

Narrative reputation bears some correspondence with Aristotle's depiction of the virtue of prudence. In the Nicomachean Ethics, Aristotle identifies one type of intellectual pursuit as a form of rationality that is expressed as informed opinion, driven by desires, and moderated by reason. Prudence is the action guiding application of reason that converts feelings and opinions into moral actions.[158] On this view, narrative knowledge reflects the rational process of prudence. When practiced in a reputational assessment received from another entity, our prudence dictates how reasonable that judgment is. We do not presume that any particular opinion about a person is the final, invariable word. Similarly, the application of prudence for Aristotle admits of variability, but it is not the same as capriciousness. Prudential action is thoughtful action shaped by moral evaluation. The variability of prudence does make its actions unreliable so much as clearly situated in our intellectual capacity for moral action. In a reputational assessment based in narrative knowledge, if we view the assessor as sufficiently prudent, the assessment will be useful and find consensus among those who practice similarly prudent judgment.

There also appears to be an echo of narrative knowledge in the work of John Rawls. Recall that the Rawlsian concept of a "comprehensive doctrine" is a set of ideals and worldview based in philosophy, religion, and other beliefs.[159] A comprehensive doctrine, and the judgments that it motivates, do not require universal acceptance by others to be nevertheless guiding or instructive for the person as an expression of their conception of the good. Similarly, Lyotard's narrative

[158] Aristotle, *Nicomachean Ethics* (JEC Welldon tr, Hackett Pub Co 1999).
[159] Rawls, *Political Liberalism* (n 14).

knowledge derives its appeal and function from its internal consistency and relationship to the knower. Narrative knowledge, then, is firmly embedded in human relations, including histories, cultures, and codes.

A second category of knowledge is what Lyotard terms *scientific* knowledge. This is knowledge that is considered factive and objective. The distinction between scientific knowledge from other forms of knowledge is an ancient one. Aristotle set aside scientific reasoning from other forms for its "demonstrable" and eternal character. Scientific knowledge is perceived as invariable and can be taught and taken up by others with a consistent sameness.[160] Similarly, Lyotard distinguishes scientific knowledge from narrative knowledge through its relationship with truth claims and the deliberative processes of both arriving at those claims and convincing others of their veracity.[161] Lyotard's scientific knowledge is structured knowledge that lends itself to formal proofs based on stepwise articulation.

Scientific knowledge is the form of knowledge typically demanded in public debate over matters of shared importance and is central to large scale voluntary cooperation. For example, scientific knowledge is the requisite form for launching a spaceship into outer space. Many people have to collaborate on the design and construction of the spaceship and still others have to agree to put resources (*e.g*. money) at risk in the venture. This requires a certain amount of legitimization of the knowledge involved, which is likely to include the well-established, public, and repeatedly endorsed understandings of aeronautical science. Collaborators must demonstrate to each other that their contributed knowledge is based in accepted facts and that their work will promote the shared goal of launching the spaceship into outer space. They must also convince their sponsors

[160] Aristotle, *Nicomachean Ethics* (n 158) Book VI.
[161] Lyotard, *The Postmodern Condition* (n 156).

that resources are being expended in a reasonable way on a venture that is likely to succeed. Everyone involved in the venture must agree in principle that the knowledge applied is sufficiently factual and uniformly compelling. Individual theories or beliefs that cannot be endorsed by the entire collaborative team won't be accepted as contributive to the goal of launching the spaceship. A single extremely wealthy person who believes that spaceships can be built from wood and launched with prayer will have difficulty finding skillful contributors willing to work from this vision. The wealthy person might succeed in building something that looks like a spaceship, but the likelihood of this spaceship flying into outer space is very small.

When institutions engage in reputational assessments, there is a tacit assumption that their assessments are based in scientific knowledge. This is persuasive to some extent. Institutional practices are often standardized; the same standard is applied to every case based on some theory or prior evidence about the likely outcome. However, what is easily overlooked is the culturally contingent nature of the institution itself. A court of law is not a fact of nature but rather a body designed to manage behavior in accordance with existing laws and associated procedures of legal practice. In a society that is both reflective and democratic, the practices and procedures of a court of law evolve to produce the best expression of justice as it is understood by society. There are many possible variations. Legal procedure is a set of interpretations that are not inevitable; the law could be interpreted in other ways. Laws themselves are contingent on prevailing political wills and desires. While under ideal conditions, laws and their application would follow an inevitable path toward a universal justice, in practice, each law, especially the sanctions for violation, are contingent on the desires and priorities of the people who make them. The justice they provide is imperfect culturally shaped. We see evidence of this when comparing laws between countries. The laws of one country may provide for same-sex marriage while the laws of another condemn

homosexuals to death. The uniform application of such laws does not make their basis of knowledge entirely scientific. They transmit narrative forms of knowledge as well. The identification of scientific, and therefore fully generalizable systems of justice is the work of political philosophy, but there is likely no existing practice that is not, at least partly, based in narrative forms of knowledge.

4.4 THE POWER OF INSTITUTIONAL REPUTATION

Perhaps the most important distinction for institutional reputation is the degree of voluntariness of the association between the subject and the assessor. Outside of institutional contexts, reputation is informed by social ties, direct experience, emotional response, and unarticulated preferences at least as often as established fact or claims of truth. Many reputational contexts are generally understood to be subjective, meaning that they reflect a set of values from the perspective of singular or shared conception of acceptable and unacceptable behavior. In practice, we expect people to endorse different values and to make judgments about others based on those values. Someone who views non-monogamous relationships as sinful or reckless is likely to assess the reputations of people who practice polyamory differently than someone who does not hold this view. Such a judgment is not assumed to be objective; it is typically understood to reflect a set of preferences, even if those preferences are deeply felt by many people. A view about polyamory does not require an appeal to fact or objective reality; one's individual conception of what is *good* is core to the shaping of an assessment here.

The reason why we should care about the subjective content of reputational assessments is their impact on the lives of subjects and the possibility for subjects to assert their agency in a reputation context. Where the stakes are relatively low, like the formation or maintenance of a social tie, the possibility that an assessment is based in narrow beliefs is both expected and

negotiable.

4.4.1 Inescapable

Voluntariness is a key indicator of institutional vs. non-institutional reputation contexts. While our social bonds are important, it is often the case that we can leave them and find others. Institutions are entities from whose influence and rules we cannot escape, at least not without suffering a great deal for it. For example, the authority of one's national government is not a choice. As Rawls describes our relationship to the state, "we enter only by birth and exit only by death."[162] We are bound by the laws and the political system in the place where we live.[163] While we may defy or circumvent prevailing constraints to some extent, the only way to truly escape them is to leave the country or other jurisdiction for another that will accept us. But even when moving elsewhere is feasible, we become subject to the new jurisdiction's inescapable laws and systems. These systems may not feel at all oppressive, or not so to all society members. A political system that encodes fundamental rights and freedoms into its structure, offers protection from various types of harm, and provides options for participating in governmental affairs may be both morally and experientially acceptable. However, regardless of the rights, freedoms, and protections offered by them, governments are extremely durable and powerful. Their power resides with a ruling class and the options of the governed are constrained by that power.

4.4.2 Coercive

Another distinguishing feature of institutional reputation is the effect of an assessment on the subject. Institutional reputational assessments have significant material, physical, or existential

[162] Rawls 'Kantian Constructivism in Moral Theory' (n 59) 136.

[163] This is not to say that people cannot act other than how law or custom dictates. Here, being "bound" means they risk *consequences*. I am able to drive a car in excess of the posted speed limit, but I risk the consequence of a speeding ticket if I do so. In this way, I am bound by applicable traffic laws.

effects on the subject. In contrast, if a relative declares a person to be “no good,” that may cause interpersonal problems with other family members but does not necessarily result in lasting harm to the person’s life chances. If an officer of the court makes a similar determination as part of a sentencing report in a criminal matter, that can have consequences on the subject’s physical freedom, such as leading to a longer prison sentence. A similar assessment by a prospective landlord where affordable housing is in short supply can have significant material consequences for someone in need of housing.

The coercive nature of institutional reputation is most obviously reflected in the relationship between citizen and state, where the state’s power is extensive. Non-governmental actors can also be coercive in society. If a company or set of companies is tasked with making evaluations of consumers that are used in dozens of life domains, the relationship between subjects and these companies may also cross the threshold of coercion. In the next section, I describe credit reporting and scoring as a form of institutional reputation.

4.5 The Case of Credit Reporting

In this section, I consider credit reports and scores as a form of institutional reputation and subject them to a brief normative inquiry. This inquiry is intended to be only a sketch that foreshadows a more substantive moral evaluation of reputation in chapter 5. I begin with a short history of credit reports and scores and then discuss some of the presumptions about their usefulness, fairness, and status for moral evaluation.

Credit reports and scores have been widely used for decades in the United States by financial institutions and other potential creditors to evaluate customers for various types of financial

access.[164] More recently, they have been used for other purposes such as tenancy and employment screening.[165] Credit reporting generally describes any method of collecting and sharing information about individuals for the purpose of assessing their value and risk as customers. Credit reporting, which emerged in a more or less familiar form in the late eighteenth century as a means of evaluating banking customers, meets the definition of an institutional form of reputation. First, it is a form of assessment practiced by an institution (the collection of banks and other businesses involved in lending money). This institution is enduring, having existed in more or less its present form for generations. Next, participation in the credit system for people in the United States is involuntary. One does not have to consent to be evaluated as a credit risk and to be the subject of credit reports and scores. Simply doing business in the usual ways, setting up utility accounts, subscribing to a cell phone plan or cable subscription, taking out student loans, etc., results in information about subjects becoming part of the credit system. Avoiding being included in this system is virtually impossible for people in the U.S. Finally, the credit system is coercive. An acceptable credit rating is required for most credit cards, which are required for many common types of purchases. Renting a car, for example, is extremely difficult in the absence of a credit card. Similarly, many types of subscriptions and accounts that have become important to mainstream life, including internet service and cell phones, involve a check of one's credit. Still further, credit ratings are used in a range of relationships; they are commonly used to screen rental applicants, job candidates, insurance customers, and in other domains.[166] Note that I have not yet

[164] Martha Poon, 'Scorecards as Devices for Consumer Credit: The Case of Fair, Isaac & Company Incorporated' (2007) 55 The Sociological Review 284.

[165] Justina Victor and others, 'SHRM Survey Findings: Background Checking–The Use of Credit Background Checks in Hiring Decisions' (Society for Human Resource Management, 19 July 2012) <https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Pages/creditbackgroundchecks.aspx>.

[166] Lisa Rice and Deidre Swesnik, 'Discriminatory Effects of Credit Scoring on Communities of Color Symposium: Credit Scoring and Credit Reporting' (2013) 46 Suffolk University Law Review 935.

argued a position about whether any of this is good or bad. I have only so far argued that credit ratings and scores are institutional, enduring, involuntary, and coercive. However, I shall engage with a key defense of credit reports and scores, which is that they are believed to be objective assessments based in carefully constructed statistical methods that accurately reflect a person's risk and work. The history of and current practice with these instruments suggest that this assertion is questionable.

Credit reporting tools, like other forms of reputation, are intended to solve the problem of information asymmetry between parties, such as between financial lenders and prospective borrowers. Reputation has always played a role in financial transactions, but whereas it may have once been customary for bankers to know or to meet with prospective borrowers and to correspond directly with merchants and other business contacts about them, the development of the credit report was intended to create a more reliable and uniform standard. The move to standardize is typically a move to shift the basis of knowledge from narrative, and therefore contingent, to scientific knowledge, based in an appeal to fact and processes of deductive reasoning.

Credit reporting appears to be effective in carrying out a particular task (financial risk assessment) on behalf of particular actors (loan officers, employers, landlords, etc.). However, the evidence that credit reporting is robustly *scientific* is not entirely convincing. Credit reports are shaped by a number of factors that cast their objectivity into doubt. First, the history of credit reports does not offer confidence. Prior to federal regulation, credit reports were assembled by hand using subjective, and often discriminatory standards. As the legal scholar Frank Pasquale explains, as late as the 1960s, "innuendo percolated into reports filed by untrained 'investigators.' They included attributes like messiness, poorly kept yards, and 'effeminate gestures.'"[167] From

[167] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015) 22.

this description, it appears that at least some of the content of credit reports was based in the situated opinions of the investigators, which included discriminatory bias. The Fair Credit Reporting Act (FCRA) of 1970 addressed some of the most egregious problems and compelled credit reporting agencies to make reports available for inspection to individuals.[168] The FCRA also offers a means for consumers to contest the content of reports about them. Notably, however, the FCRA did not entirely remove subjectivity from credit assessments, nor does it ensure accuracy in practice. First, while certain details regarding employment, medical care, and a few other categories are limited by the statute, the scope of information eligible for inclusion in a credit report is largely left to the credit reporting agencies to decide. In practice, reports of payment delinquency are generally accepted by credit agencies from any furnisher who wishes to provide them. This potentially opens the way for malicious actors to intentionally tarnish credit profiles. Disputes between landlords and tenants, for example, have been known to produce false claims that become the victim's burden to address.[169] It is also increasingly common for consumers to be the victims of identity theft and scams that result in tarnished credit.[170] A related risk is that of romantic partners who open revolving credit accounts in the other persons' name and run up debts as a form of partner-abuse.[171]

The FCRA provides some relief for some of the situations indicated above, but that relief is limited. The law's chief improvement over historical practice is that it provides for contestability—credit reports can be challenged under the FCRA by consumers. Incidents of fraud in addition to

---

[168] 15 U.S.C. § 1681 et seq.

[169] *cf.* Michael Katell, 'Adverse Detection: The Promise and Peril of Body-Worn Cameras' in Bryce Clayton Newell, Tjerk Timan and Bert-Jaap Koops (eds), *Surveillance, Privacy, and Public Space* (Taylor & Francis 2018).

[170] Rob Douglas, '2020 Identity Theft Statistics' *Consumer Affairs* (14 May 2020)<https://www.consumeraffairs.com/finance/identity-theft-statistics.html> accessed 5 March 2020.

[171] Ariane Lange, 'She Trusted Her Husband To Handle Her Money. It Cost Her More Than She Imagined' *BuzzFeed News* (7 January 2019) <https://www.buzzfeednews.com/article/arianelange/coerced-debt-financial-abuse-fix-credit-score> accessed 5 March 2020.

falsities and mistakes are subject to contestation. However, the discretion to update credit reports once challenged is granted wholly to the credit agencies who decide what to include, discard, or correct. As victims of abuse and theft have reported, having credit reports corrected is time-consuming, frustrating, and not at all guaranteed.[172] The accuracy of a credit report, then, appears to hinge on a combination of good fortune, vigilance, and the discretion of the credit agencies.

More complex to analyze are numeric credit *scores.* Credit scores were developed in the mid-twentieth century by the Fair Isaac Company (FICO), beginning first with "scorecards" derived from credit reports and eventually becoming the algorithmically produced three-digit FICO scores in common use today.[173] FICO scores are calculated using proprietary algorithms that are inaccessible to consumers and regulators. The scores cannot be directly challenged under the FCRA, which only covers credit *reports*. Credit subjects can attempt to modify their credit scores by examining their credit reports for outstanding debts, but the most they can do is to guess which entries matter for their FICO score and address those hoping for an uptick. I cannot say whether credit scores meet a standard of objectivity because we cannot evaluate them and must trust the producers of the scores at their word that they are composed using sound methods. A core reason the opacity of credit scores merits concern is because they are used by a wide range of opportunity gatekeepers, including landlords, employers, and dating apps.[174] While some empirical data exists that correlates credit scores to the likelihood a person will pay back a loan, the evidence is weaker that credit scores correlate reliably with other domains of life, such as employment or romance. And yet, credit scores govern access to many such opportunities.

An additional controversy is the fact that credit scores are statistically higher for white men

[172] ibid.

[173] Poon (n 164).

[174] Rice and Swesnik (n 166); Sabrina D Volpone and others, 'Exploring the Use of Credit Scores in Selection Processes: Beware of Adverse Impact' (2015) 30 Journal of Business and Psychology 357.

as compared to women and people of color. It is probable that credit scores do not automatically mark women and people of color with lower scores. At issue is that the lower scores for particular categories of people suggests the scores are proxies for forms of advantage and disadvantage that have their roots in social and political structures rather than merely reflecting personal failings. Illustrative examples are not difficult to conjure: Women are more likely than men to be saddled with children and deadbeat exes, leading to more opportunities to incur debt and obstacles to resolving those debts; African Americans are statistically less likely to inherit intergenerational wealth, making it harder to build reliable financial profiles that insulate against unforeseen hardships, leading to late payments; and so on. These conditions may be reflected in credit scores. Credit agencies have little incentive to act on concerns about these social factors so long as the scores function well enough as predictors of risk that decision-makers will continue to rely on them and pay for them. However, a holistic moral account of credit ratings and scores is one that accounts for, and privileges interests in addition to credit agencies and their customers. We can also ask whether the social factors affecting credit scores are important to the moral justifications that motivate larger questions of justice.

One approach to this question is to evaluate credit scoring for its effects on the values held by society. Some values are moral values, which generally means we care about them for how they pertain to the rightness or wrongness of actions or people's performance of actions. Other values are non-moral, meaning that they are cherished by people to varying degrees but carry little moral weight. An example of a non-moral value is that of profitability. While I may prefer that a particular business be profitable, either because it improves my own financial situation or because I like to buy its wares, I am not concerned with the profitability of all businesses as a moral matter. Similarly, I do not view unprofitable businesses as being "bad" simply because they fail to make

a profit, even if there are other facets of the business, such as corruption or unfair treatment of workers, that I may judge from a moral standpoint. In contrast, an example of a moral value I care about is that of helping others. I may derive a personal benefit from someone helping me or someone I care about, but I also value that others be helped when they can be because it is *morally right* to help others. If I believe someone should have helped someone else and failed to do so, I am inclined to feel that person was morally wrong for not helping.

The value of helping others can be entered into our consideration of credit reporting. The credit system generally rewards those whose labor is valued in society because those are the people most likely to be financially stable. In our society, there is a great deal of labor that involves helping others that is unpaid. Men are less likely than women to perform this labor, which includes caring for children, elders, and others who are unable to care for themselves. The burden of helping others competes with the economic efficiency of committing to a stable career that ensures financial stability. While having a steady income might be instrumental to helping others, it is not the same as actually helping others by actively caring for them. Credit reporting agencies are not in the business of caring about unpaid labor. Their calculations reward those who are in a position to be successful at financial stability. However, the interests of credit agencies are largely non-moral, more closely connected to values like efficiency and profitability. In a society committed to moral values, including the value of helping others, we should care at least as much about those who engage in unpaid labor to help others as we do for those who do much less of it. However, those who decline to perform an equal share of unpaid labor and instead focus primarily on paid labor are more likely to have higher credit scores. They are therefore able to reap the benefits of those scores while others who do valuable work are unrewarded. This is not to suggest that we cannot use credit scores for purposes where there is broad agreement about their usefulness. But we should

not permit credit scores to stand alone as a primary and authoritative measure of individual worth.

Taking the question of value a bit farther, we can also consider other values surfaced by the credit system. The value of democracy can be described as "rule by the people" and the ability of persons to influence the decisions that concern how they are governed.[175] While the value of democracy is most frequently associated with the relationship between citizens and government, credit scores also govern in many areas of life. They afford or prevent access to goods that are fundamentally important to human wellbeing, including jobs. Yet, the governing aspect of credit scoring affords no opportunity for democratic participation. The scoring method is hidden from most people and their scoring conclusions are not contestable. The choice to use a credit score to make a decision is also largely out of the hands of the affected person. A defense of credit scores as they are might privilege the interests of business over those of individuals when they come into conflict in pursuit of a larger social benefit. Richard Posner is among those who argue that the benefits to society provided by business activity are superior to individual interests in many domains.[176] A simple response to this argument is that the value of democracy does not vanish when confronted with the interests of business. What is not clear, and I do not attempt to take up here, is to what extent should we value democracy over the business of credit scoring and its uses. The point of this argument is to demonstrate that, as practiced, credit scoring can be evaluated for its connection to the value of democracy.

Similarly, the value of justice may also be implicated by credit scoring. Justice can be described as the demand that people not be advantaged or disadvantaged based on arbitrary facts.[177] Credit scoring evaluated through this value appears to unfairly disadvantage people based

---

[175] Brey (n 154).

[176] Richard A Posner, 'Privacy, Secrecy, and Reputation 1978 James McCormick Mitchell Lecture, The' (1978) 28 Buffalo Law Review 1.

[177] Brey (n 154).

on their gender and race classifications. Even if this is unintentional and merely actuarial, the end result is a condition of injustice for those affected. A solution that does not upend familiar ways of doing business is not obvious, but that does not prevent us from calling it a problem. Indeed, there is a history of contesting the actuarial facts of risk assessment when it appears to disfavor a particular category of person, in which advocates prioritized justice over the exigencies of the market. The work of moral theory in the realm of justice is often the work of choosing which interests should prevail when they come into conflict. For example, it was once acceptable to charge higher insurance rates for African Americans due to statistical claims about race and life expectancy. When arguing in favor of the first law to ban the practice, a Massachusetts legislator argued that we should be more concerned with the long-term prospects of African Americans than the short-term interests of insurance companies.[178] This was a moral argument and one that ultimately prevailed. Similarly, accepting that credit scores are good enough as they are because they are actuarily sound is indicative of a choice to favor the interests of the credit agencies and their customers over those of their subjects. That choice is available, but it is not the only choice. The normative issues presented by reputation, and a class of reputation in particular, is taken up in more detail in Chapter 6.

4.6 CHAPTER CONCLUSION

In this chapter I have introduced the concept of institutional reputation to narrow the scope of inquiry for this dissertation project. I described how institutional actors differ from others due to their outsized effects on the material and existential conditions of others' lives. I argued that institutional reputation is of interest to moral theory because of the scope of influence on individual

[178] Rodrigo Ochigame, 'The Long History of Algorithmic Fairness' [2020] *Phenomenal World* <https://phenomenalworld.org/analysis/long-history-algorithmic-fairness> accessed 5 March 2020.

lives and cherished values from institutional actors. I began a descriptive account of institutional reputation by discussing its epistemic basis, finding that it is often presumed to be *scientific* and objective but is prone to the more subjective features of *narrative* forms of knowledge. I provided a brief history of institutional reputation, connecting it to the development of long-distance trade and electronic communication. I then began a normative inquiry into institutional reputation to foreground a larger inquiry into the moral status of certain forms of reputation to be addressed in a subsequent chapter. Here, I further elaborated on the distinctive qualities of institutional reputation as both involuntary and coercive. In the final section of this chapter, I employed the case of credit reports and scores as a form of institutional reputation and posed questions about their presumption of accuracy and their value as arbiters of important life opportunities. I analyzed credit scoring for its reflection of the values of democracy and justice and found that credit scores create problems for moral theory when analyzed this way. In the next chapter, I further narrow the scope of inquiry to what I call "algorithmic reputation," a form of institutional reputation that is tied to computation. I then conduct a detailed case study of a particular instance of algorithmic reputation.

CHAPTER 5 — CASE STUDY OF HIREVUE VIDEO ASSESSMENT TECHNOLOGY

5.1 CHAPTER INTRODUCTION

I have offered an account of reputation as a human social process of judgment and decision-making that is characterized by the mechanics of observation, analysis, and production. I argued that the mechanics of reputation provide myriad opportunities for influence by an assessor's conception of the good, or what John Rawls described as one's "comprehensive doctrine." I have also argued that there is a form of reputation conducted either exclusively by or with the assistance of digital systems that is called "algorithmic reputation." I have also described a form of reputation that I labeled as *institutional,* which reflects the priorities and desires of organized entities, such as governments and large firms. I argued that decisions by institutional actors, including their assessments of others, have significant power over people's lives. Their influence is both *coercive,* by which I mean imposed and involuntary, and *inescapable,* by which I mean there are few (if any) avenues for avoiding their influence. I further argued that algorithmic reputation is typically an expression of institutional reputation because it is most often a practice of institutions such as government agencies and influential firms.

This chapter is an investigation into an algorithmic reputation product that is a member of a class of products known as automated hiring systems (AHSs).[179] AHSs automate the activities of employee sourcing, screening, interviewing, and selection using artificial intelligence technologies. I focus my investigation on an AHS marketed by the company HireVue, Inc. (henceforth "HireVue") which is prominent in the industry. My investigation specifically concerns

[179] Miranda Bogen and Aaron Rieke, 'Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias' (Upturn 2018) <https://www.upturn.org/reports/2018/hiring-algorithms> accessed 30 September 2019.

HireVue's candidate assessment system, which uses artificial intelligence technologies to evaluate audio, video, and interaction data collected during job interviews. HireVue claims to be able to gain insight into the cognitive abilities and psychological profile of job candidates which are analyzed with a machine-learning approach for predicting job success.[180] The company also claims that its technologies reduce discriminatory bias in the hiring pipeline.[181] These claims are inventoried, investigated, and morally evaluated.

The study in this chapter is presented as a Value Sensitive Design study.[182] Value Sensitive Design is a tripartite methodology that integrates conceptual, technical, and empirical investigations about a technical artifact or practice in an iterative and integrative whole. A key goal is the identification of human values implicated by a technology and its stakeholders.[183] Value Sensitive Design presumes that values and normative commitments can be embodied in the design of technologies and expressed through interactions between stakeholders and the technology.[184] I employ this framework as part of the project of this dissertation to develop an applied ethics of reputation systems.

The organization of this chapter is as follows. In section 5.2, I detail the method used in this chapter's investigations. I begin by describing the Value Sensitive Design framework and then drill down further into my specific investigative approach, a *critical case study*, which combines the tradition of case study research with critical approaches that surface the distribution of power

---

[180] Dr Nathan Mondragon, Clemens Aichholzer and Dr Kiki Leutner, 'The Next Generation of Assessments' (*HireVue* Whitepaper February 2019) < https://hrlens.org/wp-content/uploads/2019/11/The-Next-Generation-of-Assessments-HireVue-White-Paper.pdf> accessed 21 December 2020.

[181] 'Bias, AI Ethics, and the HireVue Approach' (*HireVue*) <https://www.hirevue.com/why-hirevue/ethical-ai> accessed 19 March 2020.

[182] Batya Friedman and David Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination* (MIT Press 2019). The Value Sensitive Design approach is described in more detail in Methods.

[183] ibid.

[184] Jeroen van den Hoven, 'Moral Methodology and Information Technology' in Kenneth Einar Himma and Herman T Tavani (eds), *The Handbook of Information and Computer Ethics* (John Wiley & Sons, Inc 2008).

implicated by the target technology. In section 5.3 I conduct the case study itself. I begin by providing an overview of HireVue's business segment and the company itself, including a general description of its assessment technologies. I then conduct a conceptual investigation of HireVue interview assessment technologies, which includes a stakeholder analysis to identify the direct and indirect stakeholders affected in some way by HireVue's technology. The conceptual investigation is followed by a technical investigation into HireVue candidate assessment technologies.

The case study involves an analysis of documents produced by the company that describe its products and an analysis of patents that provide insights into the underlying technologies. I close with a discussion in which the company's central claims about its products and its philosophical commitments (including its expressed values) are critically assessed.

5.2 METHODS

HireVue technologies are investigated as a Value Sensitive Design, Critical Case Study.

5.2.1 Value Sensitive Design

I ground the work of this chapter in the framework of Value Sensitive Design. The motivating ethic of Values Sensitive Design is the design of technology with human values in mind.[185] Value Sensitive Design offers a tripartite framework that integrates conceptual, technical, and empirical investigations of a particular technological artifact or practice.[186] Value Sensitive Design is an interactional approach that holds that values are not necessarily inscribed in a technology in a fixed way but emerge through a combination of the intentions of its designers and the goals of those

[185] Jessica Miller and others, 'Value tensions in design', *Proceedings of the 2007 International ACM Conference / Supporting Group Work (GROUP '07)* (2007).

[186] Batya Friedman, Peter H Kahn and Alan Borning, 'Value Sensitive Design and Information Systems' in Ping Zhang and Dennis F Galletta (eds), *Human-computer interaction and management information systems: foundations* (ME Sharpe 2006).

who use the technology.[187] Value Sensitive Design is associated with the "design turn" in applied ethics, which holds that many ethical issues can be addressed by attending to human values and normative assumptions proactively during the design of technologies that affect people's lives.[188] However, while usefully applied during or prior to design activities, the elements of a Value Sensitive Design study are also useful for taking stock retrospectively of technologies already in use. I adopt this latter approach for the work of this chapter. I identify evidence of the values implicated by existing HireVue technologies that emerge through interactions between the company's values, the perceived desires of customers, and likely effects on job candidates.

5.2.1.1 Conceptual, Technical, and Empirical Investigations

In a Value Sensitive Design study, a conceptual investigation identifies the theoretical grounding of the investigation. A feature of the conceptual investigation is the identification and declaration of a set of core human values implicated by an artifact or practice as a means of centering those values in the analysis of the artifact or practice. Values emerge from taking stock of whose interests are implicated by a technology or practice and an account of the likely benefits and harms. The values implicated by a technology may be expressed by features of the artifact itself or highlighted by the direct or anticipated experiences of stakeholders. Values may also be identified by the investigator for whom the study is an opportunity to develop or test a theory. Identified values are subjected to a philosophical investigation that surfaces definitions for the identified values and places them within a moral framework.

A particularly important feature of Value Sensitive Design is a *stakeholder analysis*.[189] Value Sensitive Design projects indicate two types of stakeholders. *Direct* stakeholders interact with a

[187] Friedman and Hendry, *Value Sensitive Design* (n 182).
[188] van den Hoven (n 184).
[189] Friedman and Hendry, *Value Sensitive Design* (n 182).

technical system or its outputs. These stakeholders may include a system's operators or the people upon whom the system is used. *Indirect* stakeholders are people or entities who do not directly interact with an artifact but are affected by its use.[190]

A technical investigation considers a particular artifact or practice from the perspective of its designed and engineered features. Technical investigations may concern any technical aspect, but frequently concerns elements of its design, focusing either on existing properties of a technology or proactively on a design plan.[191] A technological investigation might include design specifications, schematics, or the minutes of code review meetings. It may include examinations of code, physical features, or take the form of a usability study of an interface. As performed in this chapter, a study of patents relating to a technology appear to qualify as worthy subjects for a technical investigation.

Empirical investigations gather data from the human context in which a technology participates. Empirical investigations can be conducted on any human activity that can be measured, observed, or documented, using a wide variety of research strategies.[192] Empirical investigations can directly study human interaction with a technology or can include a study of a larger context into which the technology intervenes. For example, an investigation into red-light cameras used for the enforcement of traffic laws might include survey data collected from drivers who have received automated summonses. It might also include evaluations of data about the effects of red-light cameras on traffic safety in different locations.

This project includes conceptual and technical investigations. While I include some discussion about the business landscape in which HireVue participates and other types of evidence that could

---

[190] ibid.

[191] Alan Borning, Batya Friedman and Peter H Kahn, 'Designing for Human Values in an Urban Simulation System: Value Sensitive Design and Participatory Design' (2004).

[192] ibid.

inform an empirical inquiry, I generally leave a full empirical investigation to future work.

5.2.1.2 Moral and Nonmoral Values

Given the name of the method, what is meant by *values* needs to be unpacked. Within the scope of human relations, "value" has many meanings which include financial worth as well as other aspects of human experience and desire. Here, I focus on meanings of value that describe features of lived experience that people care about. Values in this range are things that people consider important in life.[193] However, as this is a project of moral inquiry, I focus on a narrower scope of values. Following Brey,[194] I suggest that some human values are *moral* values. Philosophical traditions differ on the meaning of moral values. Kantians, who favor a portrayal of morality as being connected to our duties and obligations, characterize acts and persons based on their *rightness.* Meanwhile, other traditions define values in terms of things that are unambiguously *good.* In either case, the goal is to identify values that are fundamental to the attainment of human flourishing, and potentially, toward the flourishing of non-human entities. What people merely prefer, or desire does not necessarily contain any moral content that can guide decision-making beyond suiting what pleases someone. This limited scope is particularly important to consider in matters that concern the flourishing of other beings. Moral values are morally guiding because they support generalizable evaluations that we can reasonably share about the actions or commitments of people and entities for praise or blame.[195] To further narrow the scope, I follow the information ethicist Adam Moore,[196] and focus on values that are *relational,* by which I mean they relate to assessments of goodness or rightness that are independent of personal desires but

[193] Friedman, Kahn and Borning (n 186).
[194] Brey, 'Disclosive Computer Ethics' (n. 154).
[195] ibid.
[196] Adam D Moore, *Privacy Rights: Moral and Legal Foundations* (Pennsylvania State University Press 2010).

relate to the fundamental requirements of wellbeing tied to the rationally determined needs of an entity. Relational values are moral because they sustain, promote, or further the life of the entity for whom it is described.[197] For example, *justice* is typically understood as citizens "having an equal claim to basic rights and liberties and equal opportunities."[198] It may also be defined such that advantages and disadvantages should not be intentionally allocated in a manner that is unfair or undeserved.[199] Justice is a relational moral value because the fair and deserved allocation of advantages and disadvantages is not merely a personal preference but something we should endorse for the good of others, and based on a perception about what is indeed good for them. Justice promotes the achievement of fruitful lives and communities and can be evaluated without focusing exclusively on what individuals or groups personally desire.

Nonmoral values describe other aspects of human experience that do not provide much insight into the moral worth of an action or commitment or the moral status of an actor. While we may individually find value in a particular act or actor unless that finding clearly generalizes beyond a particular context, individual worldview, or a shared conception of goodness, it carries little weight in our reasoned considerations about the fates of selves and others. Non-moral values provide little guidance toward the identification of evaluative criteria we can all share. An example of a nonmoral value is the happiness experienced when receiving a birthday card from a friend. People may appreciate receiving birthday cards but failing to receive one does not make the non-sender morally suspect. The receiver/non-receiver of the birthday card may assign great value to the act/non-act, but it would be unreasonable to generalize that value to others. Failing to send birthday cards to friends may point to other aspects of a person that are eligible for judgment on a moral

[197] ibid 38.
[198] Philip Brey, 'The Technological Construction of Social Power' (2008) 22 Social Epistemology 71, 2.
[199] Brey, 'Disclosive Computer Ethics' (n 154).

basis, but the specific act of sending/not-sending is not a moral act.

5.2.2 Critical Case Study

Beyond the requirements of the general framework, Value Sensitive Design does not dictate particular methods of inquiry. Rather, theorists and practitioners of Value Sensitive Design have evolved a number of different methods, such as Daisy Yoo et al.'s *value sensitive action-reflection model*[200] and Batya Friedman and David Hendry's *envisioning cards*.[201] I presume here that adapting other pre-existing methods for use within a Value Sensitive Design study is not unwelcome to the framework. I employ a critical case study approach to the investigation of HireVue video interview technology. A case study is an investigatory approach that seeks to explain some phenomenon within a particular context, using multiple forms of evidence.[202] According to research methodologist John Creswell, a case study is an inquiry that explores in depth a "program, event, activity, process, or one or more individuals …"[203] Another research methodologist, Robert Yin describes case study as the method of choice when a phenomenon is not readily distinguishable from its context.[204] The emphasis on context accords with the *interactional* stance of Value Sensitive Design in which the values of a technology or practice emerge from meanings assigned to it by designers and stakeholders rather than through the endogenous work of inscription by designers.[205] In the case of HireVue's interview software, the

[200] Daisy Yoo and others, 'A Value Sensitive Action-Reflection Model: Evolving a Co-Design Space with Stakeholder and Designer Prompts' (ACM 2013).

[201] Batya Friedman and David G Hendry, 'The Envisioning Cards: A Toolkit for Catalyzing Humanistic and Technical Imaginations' (ACM Press 2012) <http://dl.acm.org/citation.cfm?doid=2207676.2208562> accessed 24 January 2015.

[202] Robert K Yin, *Applications of Case Study Research* (2nd ed, Sage Publications 2003).

[203] John W Creswell, *Research Design : Qualitative, Quantitative, and Mixed Methods Approaches* (3rd ed., Los Angeles : Sage 2009) 13.

[204] Yin (n 202).

[205] Friedman and Hendry, *Value Sensitive Design* (n 182).

context is AI-supported employment recruiting and screening. The software is situated as both an intervention in the context of employment and a heroic means of solving seemingly intractable problems in this domain using technology. The inventors of the software and the spokespersons promoting it employ discursive strategies that indicate framing and beliefs about the challenges faced by employers and the potential of technology to address those challenges. Understanding the context in which this technology operates and how it is promoted to customers supports a rich understanding of the company's values and those of their target consumers. While an analysis of the software apart from its context is feasible, doing so risks stripping away important aspects of its meaning and some of the salience of its moral profile.

Taking this further, the type of case study I employ is a *critical case study,* in reference to the tradition of critical theory. Critical theory describes various approaches to political analyses that account for the distribution and exercise of power in society as exemplified by the philosophical and sociological works of Theodor Adorno, Max Horkheimer, Georg Lukács, and others. I adapt my critical case study approach by integrating the *instrumental case study* described by research methodologist Alison Pickard[206] with the critical theory of technology proposed by Brey.[207] Pickard states that case studies may be designed to either develop a theory or to exemplify a previously articulated theory. An instrumental case study is an example of the latter approach and reports on what an investigator has discovered for its salience within a theoretical framework. Reflecting on the application of critical theory to technology, Brey argues that one way to investigate a technological artifact or phenomenon is to do so in a way that foregrounds its role in relations of power. This critical case study, then, is one in which I report on the contextual meaning

[206] Alison Jane Pickard, *Research Methods in Information* (Facet 2007).
[207] Brey, 'The Technological Construction of Social Power' (n 198).

and discourse of the artifact and phenomenon of HireVue's technology, including the ways in which its use by institutional actors and underlying technologies respond to theories about the instantiation or intervention of power relations in society.

5.2.3 Thematic Analysis

For this case study, I studied publicly available documents about HireVue's products, including website documents, white papers, patents, news articles, and documents related to a regulatory inquiry. Because HireVue's products are proprietary and their code and design held in confidence by the company, access to the underlying models and data used in HireVue products was not feasible. Scholars have discussed the limitations of investigations into algorithmic systems and practices without access to the models or data used to construct them and have concluded that we can still analyze a company and its products by examining the documents that are available, such as a firm's publications and patents.[208] Indeed, science and technology studies scholar Rob Kitchin argues that finding alternative methods for analyzing algorithmic systems is not only possible but also urgently needed because of the way algorithmic systems increasingly govern aspects of people's lives.[209]

I developed an understanding of HireVue's products and claims using documents found on the company's website and public statements by company representatives in press accounts. These documents and statements make specific claims about product features and reveal aspects of the company's guiding philosophy. I also studied patents filed by the company by searching the database of the United States Patent and Trademark Office. These contain a mix of technical

---

[208] Manish Raghavan and others, 'Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices' (*Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) 469 <https://doi.org/10.1145/3351095.3372828> accessed 21 December 2020.

[209] Rob Kitchin, 'Thinking Critically about and Researching Algorithms' (2017) 20 Information, Communication & Society 14.

information, general claims about technological capabilities, and indications of the inventors' guiding philosophies. I refined these understandings with a review of prominent critiques and complaints about the HireVue's core products.

In analyzing these documents, I employed thematic analysis to surface and inventory themes. Thematic analysis, or the search for themes in a collection of expressions, is a common strategy in qualitative research methods. It is frequently applied to transcripts from interviews and focus groups where the insights offered by research subjects coalesce around discrete topic areas.[210] I employ thematic analysis on the publications and patent documents produced by HireVue, allowing these to "speak" for their authors. I borrow insights from sociologist Susan Leigh Star,[211] who describes an ethnographic approach in which consideration of multiple forms of evidence, including some that may not at first appear meaningful, can reveal a "master narrative," providing insights into politics and social structures that surround and shape a technology, practice, or entity. In the case of HireVue, discoverable insights from a thematic analysis include indications of the company's epistemic commitments, its assumptions about the predictive capabilities of artificial intelligence, and the reducibility of human psychology to machine detection and analysis.

5.2.4 Patent Discovery

I searched the United States Patent and Trademark Office website[212] and identified eleven patents assigned to HireVue by searching on the company name. In an initial exploratory analysis of the patent documents that were responsive to this search, I looked for terms that made reference to familiar labels and concepts relating to biometric identification, emotion detection, and algorithmic

[210] Creswell (n 203).

[211] Susan Leigh Star, 'The Ethnography of Infrastructure' (1999) 43 American Behavioral Scientist 377.

[212] US Patent & Trademarks Office, 'Search for Patents' <https://www.uspto.gov/patents-application-process/search-patents> accessed 21 December 2020.

processing. I then generated a list of keywords related to biometric assessment that were identified during the exploration, including "emotion," "face," "facial," "heart-rate," "mood," "personality," "sentiment," and "stress." I found that six of the eleven patents currently assigned to HireVue were responsive to searches that included at least one of the search terms. I used a text comparison tool[213] to determine which patents were substantially similar to each other based on their text content, dividing patents into affinity groups for further investigation. I identified four substantially similar patents concerning interview analysis technology whose text similarity ranged between 88 and 90 percent. I identified two other patents that, while substantially dissimilar from each other and the others in the set, have in common claims about the mitigation of bias in hiring activities through the detection of candidate demographics and class membership. I performed a close reading of these six responsive patents.

Patent documents follow an established format consisting of several sections. I consider three sections in particular, two of which are prioritized by patent experts for analysis.[214] First, I consider a patent's *claims,* which are explicit statements that explain what the invention does.[215] This is the legal description of the patent. Claims are carefully constructed so as to provide a basis for legal action to exclude others from performing the same processes or constructions without permission of the inventor. I also consider a patent's *specification,* which is a detailed account that provides an "educated reader"[216] a basis from which to make and use the invention. In addition to the two sections recommended for consideration, I also consider the *background* section of patent documents. Background sections are short narratives that contextualize the invention within a

[213] Small SEO Tools, 'Page Comparison Tool' <https://smallseotools.com/page-comparison/> accessed 21 December 2020.
[214] Roberta J Morris, 'Anatomy of a Patent' in Avery Goldstein (ed), *Patent Law for Scientists and Engineers* (Taylor & Francis 2005).
[215] ibid.
[216] ibid 16.

particular scope of practice. They may speak of a problem that the invention is intended to solve. They may include an overview of the business landscape or other environment in which the patent can make a contribution. My perception is that the background section of a patent is not required to offer evidence or pose a substantive argument for the patent reviewer. The background section offers a frame of reference with the likely intent to guide the investigator toward the conclusion that the invention is useful and necessary. In the case of patents belonging to HireVue, the themes of the background text support and clarify the themes of other public facing documents produced by the company.

5.3 CASE STUDY

5.3.1 The Business Domain of Automated Hiring

HireVue participates in a business segment called "human capital management," which the business literature describes a multibillion-dollar market in software and systems targeted to hiring managers, admissions officers, and other professionals tasked with choosing among candidates for employment, education, and other opportunities.[217] HireVue's products can be more narrowly described as "automated hiring systems" (AHSs).[218] This category includes data-driven technologies and automated decision systems that reduce the role of human evaluators in sourcing, interviewing, screening, and selecting employees in favor of software algorithms that automate

[217] Jason Cerrato and Jeff Freyermuth, 'Market Guide for Talent Acquisition Applications' (*Gartner*, 18 December 2018) <https://www.gartner.com/document/3896176?ref=solrAll&refval=243416472> accessed 11 March 2020.

[218] Javier Sánchez-Monedero, Lina Dencik and Lilian Edwards, 'What Does It Mean to "Solve" the Problem of Discrimination in Hiring?: Social, Technical and Legal Perspectives from the UK on Automated Hiring Systems', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) <http://dl.acm.org/doi/10.1145/3351095.3372849> accessed 29 January 2020.

aspects of these task.[219] AHS systems are seeing increasing use in the United States and elsewhere. It is estimated that 98 percent of Fortune 500 companies use some type of AHS.[220]

There are hundreds of companies marketing AHS products. These include companies with which the reader may be familiar, such as IBM, LinkedIn, and Oracle, as well as more recent entrants such as Entelo, Workday, and Yello. The many offerings of these companies include methods of sourcing job candidates by trawling resume sites and social media, virtual assistants and chatbots to automate some HR tasks, dedicated communications platforms to connect candidates and hiring teams, candidate-employer matching systems, and predictive analytics platforms.

Within this ecology are systems that support interview and screening processes by employing artificial intelligence to take over some of the tasks traditionally performed by hiring managers and other personnel. HireVue offers products in this category. The applications of AI in this space include identifying keywords and target phrases in resumes, locating and analyzing biographic information about candidates online, automatically constructing interview questions for particular roles, and deriving insights about job candidates from their behavioral interactions with recruiting systems. HireVue currently offers an AHS system consisting of a video-interview platform that automates unsupervised job interviews and performs subsequent screening using an analytic system that evaluates subjects based on their interview performance and interactions with the interview platform.

[219] Bogen and Rieke (n 179); Sánchez-Monedero, Dencik and Edwards (n 218). NOTE: The business literature also refers to this technology as "Talent Acquisition" technology (see n 214).

[220] Sánchez-Monedero, Dencik and Edwards (n 218).

5.3.2 Company Profile

HireVue was founded in 2004 and is based in the U.S. state of Utah. At the time of writing, it is currently a private company with approximately 350 employees. Some of HireVue's clients are the major corporations Ocean Spray, Unilever, Urban Outfitters, and Under Armour.[221] HireVue attracts attention for its novel mix of AI technologies in screening job candidates including the use of natural language processing (NLP), facial recognition, and machine learning methods for identifying and evaluating biometric and sentiment data from audio, video, and device interactions produced by job candidates.

5.3.3 Product Description

HireVue has produced technologies for conducting video interviews of job candidates since its inception. The company added artificial intelligence assessments to its video-interviewing system in 2014. As described in press coverage at the time, the system analyzes interaction data "to recommend candidates based on 15,000 interaction, behavioral and performance attributes."[222] The company's currently offered video interviewing system is an interactive tool that provides a means for job candidates to sit for unsupervised video interviews, which are retained on company servers and processed through an assessment engine that processes audio, video, and interaction data using either custom built or off-the-shelf models of "successful" employees. [223] I sketch out the HireVue interview and assessment process in **Figure** . Job candidates interact with the system by launching an app or accessing an online system where the interview is launched on demand.

---

[221] 'PitchBook Profile - HireVue' (*Pitchbook*, 3 February 2020) <https://my.pitchbook.com/profile/47555-65/company/profile#insights> accessed 17 March 2020.

[222] 'HireVue Rolls Out Data-Driven Candidate and Interviewer Recommendation Engine' [2014] *Professional Services Close-Up* accessed 16 March 2020.

[223] 'Video Interview Software | On-Demand Interview Technology' (*HireVue*) <https://www.hirevue.com/products/video-interviewing> accessed 26 March 2020.

The system automatically prompts questions and then records responses using the camera and microphone in the candidate's computer or mobile device. Based on the analysis of audio, video, and interaction data collected about the candidate, the candidate is assigned a score and/or is placed on an "achievement index."[224] The results are shared with hiring managers for human follow up. Job candidates are not provided their score by the system and receive feedback about their interview performance only at the discretion of the employer.

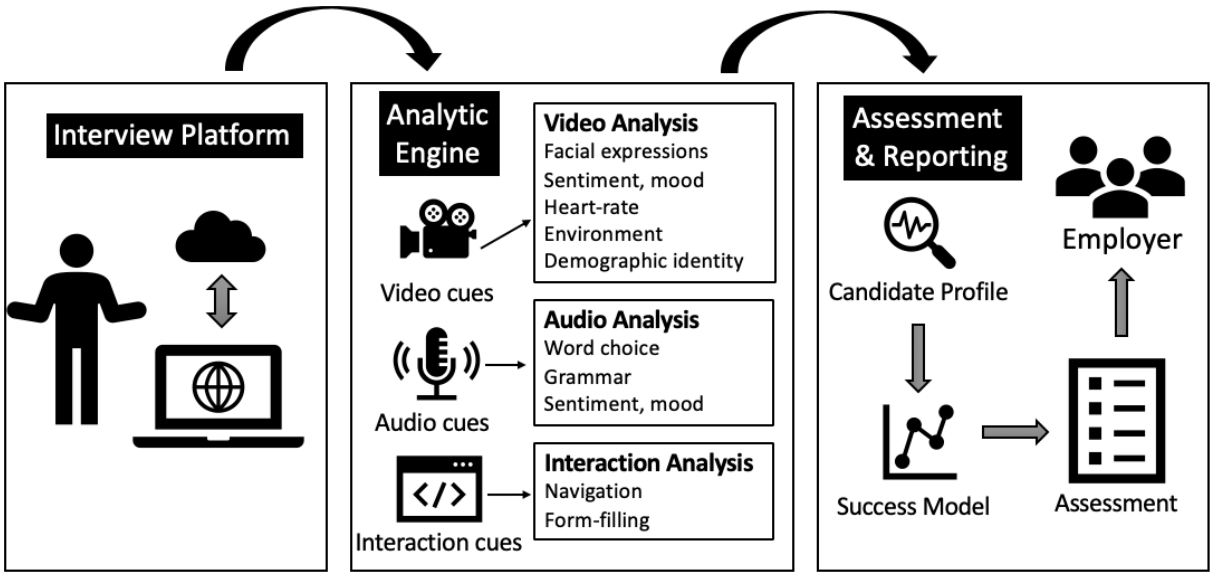

***Figure 5.1**. Interview platform interaction and analysis model. Job candidates interact with HireVue interview platform and respond to question prompts. Interview data is processed by analytic engine. Video, audio, and user interaction data is processed.*

5.3.4 Conceptual Investigation

I conduct a conceptual investigation into the assessment technology offered by HireVue. I begin with a stakeholder analysis and then examine publications produced by the company that describe how its products implicate stakeholder values.

[224] Loren Larsen and Benjamin Taylor, 'United States Patent: 9009045 – Model-Driven Candidate Sorting'.

5.3.4.1 Stakeholders

Based on an analysis of the business environment and profile of HireVue, as well as marketing documents produced by the company, I identified two direct stakeholder groups and two indirect stakeholder groups. The first direct stakeholder group consists of the customers of HireVue, which are employers who use HireVue products to support their hiring activities. Given the business segment in which HireVue participates and based on claims by the company, target customers are medium and large firms (as opposed to individual sole proprietors) that typically receive hundreds of job applications (or more) for a single position. Employers interact with the system by participating in the construction of interview question sets, inviting candidates for interviews, and consuming the predictive analyses produced from interview data. They may also participate in workplace research whose results are used to build data models used by the HireVue assessment engine. The second direct stakeholder group consists of job candidates who are assessed using the HireVue system. Job candidates interact with the system when they agree to be interviewed. Interactions include filling in forms, uploading documents, optionally downloading an interview app, and starting and stopping the interview recording. However, as I note below, including job candidates in this implies that the interests and values of job candidates are seriously taken into account in the design and implementation of this technology. As I suggest further on, this does not appear to be the case. The designers and engineers who produce HireVue's video-interview and analysis are also direct stakeholders.[225] While they are not the *users* of the system or directly affected by its use, my analysis in this case indicates that their values are strongly expressed in its design.

[225] Without the direct cooperation of HireVue personnel, an inventory of their stakeholder values must be inferred rather than collected. However, public statements by HireVue spokespersons, marketing content, and narratives found in patent documents indicate a set of values held by the company and expressed through their products.

Indirect stakeholders are typically defined as people who do not use a technology but are nevertheless affected by it. These include job seekers who have not been interviewed and analyzed by a HireVue system or other AHS. Their interests are potentially implicated by the knowledge they may be evaluated using an AHS system, as such systems become more ubiquitous. This awareness potentially shapes their thinking and preparation as they navigate the job market. While job-seekers who are interviewed using the HireVue system are listed above as direct stakeholders by virtue of their *use* of the system, they have much in common with indirect stakeholders who have yet to use the system. The interests of all job seekers are potentially affected by their awareness of the AHS employment environment and exposure to specific artifacts, such as HireVue, but this stakeholder group's capacity for influencing that environment is quite limited. As argued below, many job seekers can neither choose nor shape the effects of AHS systems in shaping their employment prospects.

5.3.4.2 Values

I identified a set of values to guide this investigation and discovered other values in documents produced by HireVue, in public statements made by company representatives, and in patent documents concerning HireVue's assessment technologies. Adapting an approach taken by Alan Borning and others, I distinguish between *explicit* values, which are values held by system designers and pursued in the artifacts they produce, and *stakeholder values* which are values of apparent importance to some, if not all, the stakeholders who either use the technology or are somehow affected by it.[226] Taking the distinction a step farther, I identify a set of values that reflect

[226] Alan Borning and others, 'Informing Public Deliberation: Value Sensitive Design of Indicators for a Large-Scale Urban Simulation' in Hans Gellersen and others (eds), *ECSCW 2005* (Springer-Verlag 2005). NOTE that Borning and others identified "explicit values" proactively as values to be integrated into a system whereas my approach is retrospective.

the *ethical stance*[227] of this author that I will refer to as "investigator values." First, an investigator value indicated by the focus of this dissertation is *reputation.* Job candidates have reputations based in their performances and documented histories. HireVue's perspective on this value is expressed by the production of candidate reputations by analyzing their behaviors using artificial intelligence and issuing assessments. This would appear to position HireVue's understanding of reputation as something that can be objectively ascertained. HireVue publications tout the capability of gaining deeper insights into the characters of job candidates, promoting their ability to produce a more detailed account of each candidate as a selling point. By claiming that their systems are "objective" and "scientifically validated,"[228] HireVue implies that reputations are static and scientifically measurable. HireVue signals that their technologies can surface hidden facets of job candidates by analyzing gestures, word choices, heart rate, and other external cues thought to shed light on personality traits. By positioning their technology as "truth-teller" about the inner lives of job candidates, HireVue signals that the subjects of its technology have immutable traits that can readily and accurately derived, and also, that the standards by which they evaluate job candidates conform to universal notions of cognitive fitness.

Company narratives, as expressed in marketing materials, patent documents, and public statements, suggest that the reputation data their products produce is more reliable than what is produced by other methods. HireVue emphasizes a strong endorsement of machine capabilities. For example, HireVue makes frequent reference to the unreliability of human evaluators and states that people trust artificial intelligence more than human-produced insights.[229] In response to these

---

[227] See Chapter 1.3, "Ethical Stance."

[228] Mondragon, Aichholzer and Leutner (n 180).

[229] Drew Harwell, 'A Face-Scanning Algorithm Increasingly Decides Whether You Deserve the Job' *Washington Post* (6 November 2019) <https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/> accessed 30 March 2020.

claims, my analysis places the explicit values expressed by HireVue in conversation with my investigator values that challenge the a view of reputation as static, objective, and interpretable by software.

Additional investigator values include *justice*, *fairness*, and *accountability*. Assessment processes are not always just, fair, or accountable and this is a generally accepted condition in many domains of daily life. However, as I argue elsewhere, there are some contexts in which the stakes are sufficiently high that we can demand—or at least approach—justice, fairness, and/or accountability in processes of reputation. I argue that employment merits consideration as one of these high-stakes domains. For the purpose of this study, I offer some brief definitions. For example, Brey defines justice as a state in which individuals are not advantaged or disadvantaged unfairly.[230] Elsewhere, Brey also defines justice as meaning that "citizens have an equal claim to basic rights and liberties and equal opportunities in the pursuit of social benefits."[231] Using these definitions, we can state as a starting point that justice and fairness equates to the correct or acceptable use of a technology. However, Brey's definitions gesture toward the need to consider a wider scope of impact on affected agents. For example, Anna Hoffmann argues that the pursuit of justice and fairness in technical systems also means addressing the social context in which a system comes into existence and performs its work.[232] On this view, even the correct use of a technology that carries forward unjust or unfair structures of power is neither just nor fair.

Understood through the lens of reputation, fairness can mean that reputational assessments are performed in a manner that meets a standard that would be endorsed by both an assessor and a subject (here, a job candidate). Additionally, the fairness of a reputational assessment can be

[230] Brey, 'Disclosive Computer Ethics' (n 154).
[231] Brey'The Technological Construction of Social Power' (n 198) 2.
[232] Anna Lauren Hoffmann, 'Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse' (2019) 22 Information, Communication & Society 900.

determined through comparing its procedure to that which would be adopted under the conditions of the original position as articulated by Rawls. This is not to say that fairness requires that subjects are always assessed exactly as they wish but rather that whatever system is used to assess them would be at least minimally acceptable to them if they had an opportunity to evaluate it for all participants under an initial condition that was fair. Fairness also demands that we look at the ecology in which some assessments are performed and evaluate the conditions of power that the assessment uncritically reproduces. Assuming a condition in which there are more job candidates than jobs, subjects agreeing to be assessed in a particular manner while under duress and offered no appealing alternatives is not a condition of fairness.[233] Justice also implies that a system of reputation ought to be accountable. When a system produces harms to either the assessor or the subject, there should be some means of adjudicating such harms. Justice also means that the accountability be meaningful and take not only individual actors to task but also examine the underlying social structures that provides some people or institutions with the power to make crucial decisions over people's lives.

HireVue publications provide several explicit stakeholder values to consider. These values are provided by HireVue as meaningful and important to HireVue's customers (employers). They also reflect HireVue's own commitments. There is brief mention of a value HireVue believes is important to job candidates ("convenience"). The most prominent stakeholder values expressed by HireVue include accuracy, bias mitigation, efficiency, and the supremacy of science/machines. An additional value described in HireVue patents is employee "fit" with a company.[234] The majority of values offered in HireVue documents are non-moral values, however, the values of justice and

[233] Under a different set of conditions, the meaning of fairness is likely to be different. Here I make certain assumptions about scarcity and the types of employment contexts in which HireVue's technology is most likely to be used.
[234] Larsen and Taylor, 'United States Patent' (n 224).

fairness lurk behind some of these. For example, accuracy is not generally considered a moral value, though it may be desirable to some stakeholders. HireVue touts accuracy, along with scientific validity and objectivity, as a form of fairness. As argued in their publications, the reputational assessments performed by HireVue technologies are fair because, for them, the main source of unfairness is human bias. Because their systems are designed to minimize human decision-making in the hiring pipeline, they therefore characterize their automated assessments as fair.

HireVue also makes a number of claims about its commitment to combatting discrimination in the hiring pipeline. The terminology used varies from one document to another but includes phrases such as "adverse impact mitigation" and "bias-mitigation."[235] For HireVue, discriminatory bias and adverse impact are perceived to be harms that should be avoided by all persons at all places in time, suggesting that this is a morally desirable aim. Fairness is the most obvious moral value here. If HireVue is correct in asserting that its technologies address discriminatory bias, then it would follow that they are claiming their system is fair. However, justice is also implicated in these claims. HireVue makes clear that the bias-mitigation features of its technology can limit liability for their customers who are at risk of employment discrimination claims. This portrayal of justice is arguably directed to only one dimension of the employer-candidate relationship. It implies a benefit primarily intended for employers; they feel confident they will be more immune to anti-discrimination claims. Anti-discrimination claims in HireVue documents, meanwhile, do not discuss how or if their technologies promote justice for job candidates who suspect employment discrimination. This is not to say that HireVue does not or cannot achieve positive net results for job candidates when compared with prior practices. It is to say that HireVue provides

[235] Mondragon, Aichholzer and Leutner (n 180).

no evidence for their claims and offers promises only to employers.

Table 5.1. *Inventory of stakeholder values. Includes mapping of stakeholder values to explicit values, stakeholder holding value, and sample text.*

STAKEHOLDER VALUES	EXPLICIT VALUES	STAKEHOLDER	SAMPLE LANGUAGE
ACCURACY	Fairness	Customers	HireVue augments recruiting teams' hiring decisions with accurate insight into every candidate's job-related competencies.[236]
ACCOUNTABILITY	Fairness, Justice	Candidates	*(No evidence of accountability as a value considered in HireVue statements or publications)*
BIAS REDUCTION	Fairness, Justice	Customers, HireVue	Our mission is…to use the technology to actively promote diversity and aid in the achievement of equal opportunity for everyone regardless of gender, ethnicity, age, or disability status.[237] … when candidates are chosen on the basis of gender, race, religion, ethnicity, sexual orientation, disability, or other categories that are protected to some degree by law, penalties may be imposed on entities for such practices.[238]
CONVENIENCE (SPEED)	(none)	Customers, Candidates	A single employee candidate can be very costly in terms of man-hours needed to evaluate and interact with the candidate before the candidate is hired.[239] Recruiters, hiring managers, and candidates gain time and convenience, reduce time to hire…[240]
EFFICIENCY	(none)	Customers	Understanding a comprehensive view of every candidate allows you to focus your time and resources on the candidates with the highest potential...[241]
FIT	(none)	Customers	It is important for an employer to find employees that "fit" open positions.[242]
OBJECTIVITY & SCIENTIFIC VALIDITY	Fairness	Customers	By combining video interviews with predictive, validated IO science and artificial intelligence (AI), HireVue augments human decision-making in the hiring process and delivers higher quality talent, faster.[243]
MACHINE SUPREMACY	Fairness	Customers	Humans are inconsistent by nature. They inject their subjectivity into the evaluations…AI can database [sic] what the human processes in an interview, without bias … And humans are now believing in machine decisions over human feedback.[244]

[236] 'Pre-Employment Testing Software | Online Gamified Assessments | HireVue' <https://www.hirevue.com/products/assessments> accessed 30 March 2020.

[237] 'Bias, AI Ethics, and the HireVue Approach' (n 181).

[238] Benjamin Taylor and Loren Larsen, 'United States Patent: 9652745 - Model-Driven Evaluator Bias Detection'.

[239] Larsen and Taylor, 'United States Patent' (n 224).

[240] 'Video Interview Software | On-Demand Interview Technology' (n223).

[241] ibid.

[242] Larsen and Taylor, 'United States Patent' (n 224).

[243] 'About HireVue | What We Do and How We Got Here' (*HireVue*) <https://www.hirevue.com/company/about-us> accessed 8 March 2020.

[244] Harwell (n 229).

As addressed in the Discussion section below, HireVue's statements about their commitment to fairness are insufficient assurances that their technologies actually *achieve* these moral goals. After an examination of the technical features of HireVue technologies, a discussion of HireVue claims is assessed against potential problems with the company's claims.

A value that is not expressed in HireVue publications and statements but reported as a concern by job candidates and scholars concerned with HireVue practices is the value of accountability. Accountability can be understood as a foundational value in its own right but also one that is instrumental to the values of justice and fairness. Accountability can be understood as a foundational value in its own right but also one that is instrumental to the values of justice and fairness. As a foundational value, accountability is defined by Moore and co-author Sean Martin as the normative condition in which "one is morally justified in sanctioning, punishing, praising, or benefiting someone."[245] While discussing accountability in technical systems, philosopher Helen Nissenbaum equates accountability with moral "blameworthiness."[246] Citing Ian Feinberg, Nissenbaum defines accountability in terms of the bearing of responsibility for actions that cause harm or contribute to harm, whether or not intended.[247] For accountability to be meaningful, agents must be able to enforce some sort of sanction.[248] Implied in Nissenbaum's depiction of accountability is the problem of identifying the responsible party and/or the source of the harm. Holding someone accountable is particularly challenging in the domain of technical systems and products because of the complexity of their construction and the complicated organizational structures involved in their production. These make assigning responsibility difficult. An

[245] Adam D Moore and Sean Martin, 'Privacy, Transparency, and the Prisoner's Dilemma' [2018] SSRN Electronic Journal 9 <https://www.ssrn.com/abstract=3212217> accessed 13 April 2020.
[246] Helen Nissenbaum, 'Accountability in a Computerized Society' (1996) 2 Science and Engineering Ethics 25, 27–28.
[247] ibid.
[248] Moore and Martin (n 245).

additional layer of challenge to accountability in technical systems is the secrecy demanded by technology firms to protect their proprietary work from competitors. Technology companies frequently invoke intellectual property protections and sector-specific laws, such as the Computer Fraud and Abuse Act, when challenged to reveal the finer details of how their systems work.[249] The challenges of identifying a party to bear accountability and in uncovering evidence make it especially difficult to hold technology firms accountable.

In the case of HireVue, job candidates who are assessed by the system have few options for accountability, though there are indications they desire it.[250] Job candidates who agree to be subjected to HireVue technologies are given very little information about how they are being assessed and are only given access to their scores at the employer's discretion. The opacity of the system for job candidates means that they are unable to hold employers or HireVue fully accountable, leaving them vulnerable to being rejected based on unreasonable or unlawful standards.

5.3.5 Technical Investigation

In the technical investigation, I analyze publications, public statements, and patents produced by HireVue to construct an understanding of what HireVue's technology does and how its work implicates human values.

According to HireVue publications and public statements, HireVue video interviewing products use artificial intelligence to assess job candidates by analyzing: i) speech content, ii) "intonation, inflection, and other audio cues," and iii) "the emotions a candidate portrays, particularly in relation to what is being said at the time."[251] Based on a close reading of four pages on the company's public-facing website, a company-produced white paper that discusses the use of AI in job candidate assessments, and statements by company officials in two recent news articles, I inventoried several product claims (see

[249] Ajunwa, 'Automated Employment Discrimination' (n 49).
[250] Harwell (n 229).
[251] Mondragon, Aichholzer and Leutner (n 180) 4.

Table 5.2).

Among the claims made by the company, HireVue collects data from interview subjects in the form of audio, video, and interaction "cues," which are then processed by artificial intelligence. HireVue claims that their system not only evaluates the verbal answers to interview questions but also evaluates candidates' cognitive abilities and emotional state by assessing the data. As expressed by HireVue, their systems evaluate the content of candidate utterances (presumably using natural language processing), as well as their intonation, inflection, and other detectable features from interview audio. HireVue also claims to detect the candidate's emotions from audio and video data.[252]

[252] ibid.

Table 5.2*. Inventory of claims found in HireVue publications/statements, including implicated values.*

CLAIM FEATURE	PRODUCT CLAIMS	IMPLICATED VALUES
INTERVIEWING PLATFORM	The HireVue video-interviewing platform enables hiring managers to arrange video interviews that can be conducted at the interviewee's convenience and on any device.[253]	Efficiency, Convenience
AUTOMATED ASSESSMENT	Audio, video, and platform interaction data is assessed using artificial intelligence to compare word choice, tone of voice, facial expressions, and other data with data models of successful job performers.[254] Recorded interviews provide thousands of data points, which contribute to an overall score for a job candidate.[255] Twenty nine percent of a person's score comes from "facial action units" while the remaining score comes from audio the words used and audio features, such as tone of voice.[256] "A standard 30 minute interview assessment can yield up to 500,000 data points."[257] Hiring managers receive a score for each job candidate assessed by the system.[258]	Machine supremacy
BIAS MITIGATION	HireVue products reduce human bias in recruiting and hiring.[259] HireVue practices and technologies advance the ability to "monitor, detect, and mitigate bias."[260] Data that "contributes to adverse impact" can be removed from the analysis.[261]	Fairness, Justice

5.3.6 Patent Analysis

In an analysis of patents assigned to HireVue, I identified claims and specifications that provide some insight into the technological features of the company's products. Without access to the source code or other protected elements of the products themselves, some speculation is required to make a direct connection between patents and products. However, the claims and background information provided in the company's published patents appear to correspond with claims made in the company's marketing materials. I proceed under the assumption that the correspondence is not coincidental.

---

[253] ibid 4.
[254] ibid.
[255] Mondragon, Aichholzer and Leutner (n 180).
[256] Harwell (n 229).
[257] ibid.
[258] ibid.
[259] Mondragon, Aichholzer and Leutner (n 180).
[260] 'Bias, AI Ethics, and the HireVue Approach' (n 181).
[261] Mondragon, Aichholzer and Leutner (n 180) 4.

5.3.6.1 Patents with Claims Regarding Assessment of Job Candidates

Four patents, 8751231, 8856000, 9009045, and 9305286,[262] filed by HireVue between December 2013 and February 2014, were responsive to the keywords "sentiment," "personality," "mood," "heart-rate," and "facial."[263] These four patents concern software capabilities for analyzing audio, video, and interaction data acquired from job candidates. In the claims sections for patents 8751231 and 8856000, the patents are described as chiefly concerned with processing "audio cues," while in patents 9009045 and 9305286, the claims use the more inclusive term "interview data." Sub-claims for all four patents include similar references to audio, video, and interaction data. Despite differences in the arrangement of text and some terminology used in the claims sections, all four patents use identical or nearly identical language in describing the processing of audio, video, and interaction data in their detailed description sections. Consequently, while there are differences between the patents, I consider all four patents to be sufficiently similar to faithfully summarize their claims together:

Table 5.3*. Summary of capabilities described in United States patents 8751231, 8856000, 9009045, and 9305286*

FUNCTION	DETAILS
AUDIO PROCESSING	Converts audio to text Produces an analysis of candidate's emotions from audio cues Calculates a "sentiment score" from the text Calculates statistics about word choice, grammar, vulgarity, and filler words
VIDEO PROCESSING	Detects subject's heart rate Identifies facial expressions (smiling, confusion, agitation) Identifies subject's environment (clutter, degree of privacy) Identifies movement of subject
INTERACTION PROCESSING	Analyzes form-filling behavior (errors) Analyzes other interactive behavior for performance
SUCCESS PREDICTION	Compares subject's audio, video, and interaction with model of successful employees

[262] Loren Larsen and Benjamin Taylor, 'United States Patent: 9305286 – Model-Driven Candidate Sorting'; Larsen and Taylor, 'United States Patent' (n 224); Loren Larsen and Benjamin Taylor, 'United States Patent: 8856000 – Model-Driven Candidate Sorting Based on Audio Cues'; Loren Larsen and Benjamin Taylor, 'United States Patent: 8751231 – Model-Driven Candidate Sorting Based on Audio Cues'.

[263] A keyword list was generated through an exploratory search of several patents. See the methods section of this chapter for more details.

As described in the patent documents, the technologies include analytic capabilities to process interview data, including recorded audio and video and data about the candidate's interaction with the interview platform. Audio data is interpreted by algorithms that identify words, evaluate word choices for grammar and vocabulary, repeated words, gaps between utterances, and sentiments revealed by words used. Audio data is also analyzed for "mood detection (*e.g.* aggression, distress, engagement, motivation, or nervousness), or the like." Video data is interpreted by algorithms that examine face and body features to predict the candidate's emotions. Video data can be magnified using "Eulerian video magnification" to detect candidate's heartrate.[264] Video data is also analyzed to identify characteristics of the physical space in which the candidate conducts the interview to determine "how cluttered is the background, how private is the environment."[265] User interaction with a user interface is analyzed based on performance factors including "proper form-filling," words typed per minute, and the speed with which the candidate navigates the platform.[266]

Each of the capabilities is described as being used to place and order candidates on an "achievement index" that rates candidates against an unspecified criterion of success for a given job classification.[267] Notably, the patents in my search provide a clear indication as to how voice, facial features, or interactions are associated with the moods, sentiments, or other psychological states. Statements by HireVue representatives and non-patent publications discuss the ability of HireVue to assist in developing custom-built data models that provide "benchmarks of success" based on an analysis of current employees.[268] HireVue publications also indicate that pre-built models are available based on primarily technology-related jobs (such as coding).[269] The level of

[264] Larsen and Taylor, 'United States Patent' (n 224).
[265] ibid.
[266] ibid.
[267] ibid.
[268] Mondragon, Aichholzer and Leutner (n 180); Harwell (n 229).
[269] Mondragon, Aichholzer and Leutner (n 180).

detail in the patents is not helpful in gaining a full understanding of these features or its basis in behavioral or cognitive science. HireVue publications make repeated reference to "scientific validity" and state that their data scientists work alongside industrial-organization psychologists.[270] However, the details of the testing methodologies and the nature of the collaboration with psychologists are not provided in any detail.

5.3.6.2 Patents Concerning Bias Detection and Mitigation

I also identified two other patents that responded to a keyword search for "facial" and "recognition." The first of these patents, 9652745, was filed in 2014 and describes a method for detecting bias in an evaluation process." Patent 10438135, filed 2016, is related to the first patent and extends the description of the machine learning techniques used to process interview data. Its most salient features are also found in claims and description of the first patent. These are summarized in **Table 5.4**. Summary of claims and specifications in United States patent 9652745.

Patent 9652745 is described as a "human bias detection tool." The system is claimed to be able to extract features from interview data to determine i) if the candidate is a member of a protected class (listed as race, gender, religion, ethnicity, sexual orientation, age, or socioeconomic status), and ii) if the predicted outcome of their evaluation (hired/not hired) fails to match the actual outcome based on a metric of disparate impact. These functions are summarized in the detailed description:

> The bias detection tool extracts characteristics of the evaluation candidates from the digital interview data, classifies the evaluation candidates based on the characteristics of the candidate extracted from the digital interview data, and determines whether the evaluation data indicates a bias of one or more evaluators of the set of evaluators with respect to one or more of the extracted characteristics.[271]

[270] 'Bias, AI Ethics, and the HireVue Approach' (n 181); Mondragon, Aichholzer and Leutner (n 180).
[271] Taylor and Larsen (n 238) col 7.

The patent indicates the use of both audio and facial recognition technologies. Facial recognition technologies include an "Eigenface" technique, an established facial recognition approach that employs multiple derivations of a visual image of a face to map its identifying characteristics.[272] The patent also mentions the use of an "Active Appearance Model," described as "a computer vision algorithm that includes a number of points that may be mapped onto the face to form a model of the face."[273] The patent also discusses voice analysis techniques that include the analysis of audio data to convert voice to text and to detect voice features including pitch, speech rate, accent, and other features. The patent describes a "supervised learning" system (a form of machine learning) for applying the detected visual and audio data to a model that indicates whether a candidate is a member of a protected class.

***Table 5.4**. Summary of claims and specifications in United States patent 9652745*

FUNCTION	DETAILS
AUDIO PROCESSING	Analyzes voice features to identify candidate demographics (race, gender, age)
VIDEO PROCESSING	Analyzes facial features to identify candidate demographics
PROTECTED CLASS MODELING AND PREDICTION	Builds a model using supervised learning to identify features of protected class candidates Compares candidate interview characteristics against a predictive model to determine likely outcome
DETECTION AND REPORTING	Analyzes interview data and company hiring practices to determine if there is disparate impact in hiring
REPORTING	Notifies decision-maker when disparate impact in hiring practices is detected

5.3.7 Discussion

Throughout their publications, public statements, and patents, HireVue makes a series of claims, both explicitly and implicitly, about the capabilities of their assessment system. In pursuit of a "master narrative" depicting HireVue's commitments,[274] I summarize these claims into three categories: First, HireVue claims that their use of artificial intelligence can provide insights into

[272] See https://en.wikipedia.org/wiki/Eigenface
[273] Taylor and Larsen (n 238) col 8.
[274] Star (n 211).

the personalities of job candidates who participate in interviews using their system. Next, HireVue claims that it can use these insights to determine the quality of "fit" for a job candidate based on models of employee success. Finally, HireVue claims that their system mitigates discriminatory bias in the hiring pipeline both by reducing the participation of human decision-makers in candidate screening and also by employing artificial intelligence to detect bias in hiring. I provide a lengthy treatment of the first claim and then briefly touch on the second and third.

5.3.7.1 Algorithmic Psychometrics

HireVue claims that through the use of artificial intelligence, their system can reliably interpret audio, video, and interaction data to provide insights into the personalities of job candidates. For example, HireVue claims they can detect the "mood" of a job candidate from an audio file. To support this claim, HireVue touts their use of both psychological and data science. They also claim that their methods are "scientifically validated and the results are "objective."[275] These claims coincide with an emerging trend that has been labeled *algorithmic psychometrics*, which joins together the methods and techniques of data science with a science that seeks to identify inner psychological states from external cues.[276] By constructing a model of human behavior mapped to psychological states—emotions and the like—data scientists employ methods to predict the state of an observed subject. For example, HireVue uses image magnification to detect the heartrate of interview subjects and compares rate increases and decreases with a model of other heartrates.[277] Joined with other data and models, HireVue claims they can predict a subject's nervousness or aggression.

Efforts to unite psychology and computer science is as old as modern electronic computing

[275] Mondragon, Aichholzer and Leutner (n 180).
[276] Stark (n. 3).
[277] Larsen and Taylor, 'United States Patent' (n 224).

but has become prominent with the emergence of the field of computational social science, which blends data science with research into human affairs. Computational social science employs big data sources and methods to corelate behavioral data with behavioral predictions. The emergence of this research domain follows the so-called "behavioral turn" in digital commerce in which behavioral data gathered from people interacting with information systems is used as the raw material for constructing novel information commodities.[278] HireVue participates in this marketplace by converting interview data into an assessment product that offers job candidate insights to prospective employers.

The use of algorithmic psychometrics by companies like HireVue for evaluating job candidates is not uncontroversial. Prominent critics argue that the evidence supporting claims that facial movements and other biometric cues can reliably predict employee productivity or success is unconvincing.[279] Other critics accuse HireVue of engaging in "pseudoscience" by claiming to understand the inner lives of job candidates from facial movements and the like.[280] Going beyond hyperbolic critiques, there is a long-running debate about the challenges of reliably and ethically assigning numerical scores to human experience. A key element in the debate are the epistemic assumptions inherent in the move to quantify and rationalize the full range of human experience. Here, a concern is that quantifying human action strips away important aspects of the actuality and context of the behavior.[281] This is not merely a problem of imprecision but one of epistemic commitment. HireVue demonstrates a strong commitment to the belief that artificial intelligence can act as a sort of truth-teller about human psychology. While history may show that HireVue's

[278] Anthony Nadler and Lee McGuigan, 'An Impulse to Exploit: The Behavioral Turn in Data-Driven Marketing' (2018) 35 Critical Studies in Media Communication 151.
[279] Harwell (n 229).
[280] ibid.
[281] Scott Timcke, 'The One-Dimensionality of Econometric Data: The Frankfurt School and the Critique of Quantification' 15.

claims are accurate, there are lessons from history that suggest caution in accepting them. HireVue is certainly not the first promoter of a seemingly powerful technology and its ability to reveal deeper meanings about test subjects. One notable example is the polygraph (a.k.a. "lie detector"), which was claimed by its champions to be ninety percent accurate in identifying falsehoods and was used by government agencies, employers, and others to gain insights into their subjects for decades.[282] However, while polygraph use peaked in the 1980s and is still used in a narrowly defined range of circumstances, it has been widely discredited. Studies of government agencies and industries that made widespread use of the polygraph as a norm found it to be no more reliable a means of identifying falsehood than asking an examiner to guess about a subject's truthfulness.[283] Furthermore, comparative research on businesses based on their adoption of the polygraph found that the use of polygraphs either did not benefit or actually cost businesses lost profit potential through overreliance on the technology. The use of the technology in criminal justice matters, once commonplace, was ultimately found to produce unacceptable rates of false positives. The polygraph is now banned for use in many domains of practice, including employment.[284]

The criticism of the polygraph and its ultimate fall from prominence fits a pattern that may be applicable to the current optimism about artificial intelligence used for similar purposes. Psychologists question the ability of any technology to quantify the essence of personalities and moods into mathematics. Too many aspects of human experience are irreducible to rational quantification, risking the production of profiles that are a distortion of reality. From a qualitative standpoint, imposing measurement on of human experience extracts that experience from its rich context. In the case of the polygraph, for example, victims of sexual abuse were often declared

[282] Walter Goodman, 'Lie Detectors Don't Lie' *New York Times* (New York NY USA, 1965) SM12.
[283] Kerry Segrave, *Lie Detectors: A Social History* (McFarland 2004).
[284] ibid.

untruthful during lie detector tests but had their claims validated upon further investigation.[285] The problem with the test is its reliance on a baseline of stress indicators and is ill-equipped to distinguish them for subjects in a heightened emotional state, such as that of abuse victims.[286] Quantification tests rely on theories of static relationships between psychological and physical traits that do not account for environmental or other factors. Tests like the polygraph do not embrace the complexity of social structures, political orientations, and the full range of somatic experience, which do not transfer smoothly into fixed numbers and then back again to inconstant and dynamic human reality.

Even the most mundane of measurements strips away or reifies critical features. Philosopher Georg Lukács illustrates this point using the example of the clock as a measure of human labor. We generally trust clocks to measure out time, yet while reliably measured, time has an important experiential aspect, as anyone who engages in high-intensity fitness regimes can attest. Employers value the time clock because it quantifies something that is most easily calculated for renumeration. Yet, in the workplace, an hour can pass quickly or excruciatingly slowly. It may make practical sense for an employer to measure out wages uniformly based on the passage of time, but something is left out, nonetheless. This example, while arguably simplistic, suggests that the act of valuing human behavior along scales of quantification risks overlooking important details about that behavior and erasing many contextual factors. Echoing this concern, science and technology studies scholar Luke Stark contests claims by health professionals embracing the digital mediation of human experience as requiring a number of "conceptual jumps" to conclude that digital systems can reliably and impartially measure human bodies and minds.[287] For Stark,

[285] ibid.
[286] ibid.
[287] Stark (n 3) 210.

algorithmic psychometrics forces the reimagination of human experience across distinct and incompatible measurement scales, from qualitative to quantitative and then back again.[288] Applied to HireVue, the claim to be able to quantify job candidates based on biometric cues dismisses the destructive effects of shifting analysis of human behavior from context-rich qualitative measurements to sterile quantification, and then projecting the results back onto the living subject by assigning a detected mental state. Even if a HireVue assessment is predictive of something that its customers find useful, the quantification makes it incomplete at best and potentially inaccurate for many candidates.

The use of algorithmic psychometrics by employers also affects the balance of epistemic power between employers and job candidates. In a typical relation between employer and job candidate, there is some degree of informational give and take in an assessment process. The job candidate seeks to balance between providing an abundance of information about herself to a prospective employer while still maintaining a domain of privacy to protect information that is discrediting, embarrassing, needlessly invasive, or some combination. Meanwhile, employers provide the minimal amount of insight into their selection methods to prevent candidates from "gaming the system." The affordances of digital systems, in which a great deal of information is available about both the professional and non-professional features of the lives of job candidates, grant an increasing amount of epistemic power to employers in this give and take.

Another criticism of HireVue is the information asymmetry built into their process. While their technologies provide a range of data points to employers, they offer far less information to job candidates. Under conditions of scarcity, in which there are fewer jobs than job candidates, there is an imbalance of power between employers and job candidates. The emergence of new

[288] Stark (n 3).

affordances that provide information-rich assessments—but only to employers—exacerbates this power dynamic. Job candidates are not only placed in a position of relative weakness in the assessment process, they have little recourse to contest the process or to learn from the assessment to improve their future prospects.[289] While the decision-processes of employers are often opaque, career counselors can conduct mock interviews and coach job seekers to improve their performance. However, the inaccessibility of HireVue's evaluative standards makes this difficult. Career counselors report that they do not know what advice to give, leaving candidates to simply make uninformed guesses in hope of satisfying the algorithm.[290] This lack of recourse for job candidates means that they are subject to system that makes decisions about them while lacking any route to accountability. Here we can ask, is it fair to be evaluated by a system whose evaluative criteria is i) hidden, ii) only provided to others, and iii) affects the distribution of a primary good? While there may be practical arguments in favor of a company like HireVue choosing not to make their system more knowable to those who are subjected to it, such arguments leave the interests of job candidates entirely ignored. Employers at least get some access to the products of the system, but they too are largely blocked from knowing the mechanics of the assessments.

There are two obvious responses to these criticisms, and I take up each in turn. First, even while critics cast doubt on what HireVue systems do from an epistemic standpoint, HireVue claims that their systems work, satisfying a large number of their customers. As evidence, HireVue points to the satisfaction of their customers and defends their practices as "scientifically validated."[291] For example, in one publication, HireVue defines validation as,

> the process by which a selection system is shown to reliably and consistently support valid inferences that relate to and predict job-related outcomes and

[289] Harwell (n 229).

[290] Rachel Metz, 'There's a New Obstacle to Landing a Job after College: Getting Approved by AI' *CNN* (15 January 2020) <https://www.cnn.com/2020/01/15/tech/ai-job-interview/index.html> accessed 15 January 2020.

[291] 'Pre-Employment Testing Software | Online Gamified Assessments | HireVue' (n 236).

> behaviors. A validated assessment will have evidence to support how job-related behaviors or characteristics are measured, what is measured, and how that measurement process relates to valued outcomes, such as on-the-job performance, core competency behaviors, and retention.[292]

HireVue follows up this definition by making frequent references to their own rigorous evaluation of their technologies with "full validation testing."[293] However, in company publications and public statements, I found no details about the content or results of that testing beyond their general claims of validity. As revealed in their publications and public statements, HireVue's testing is conducted entirely in-house by a "board of expert advisors."[294] Customer satisfaction may be one measure of success, and perhaps as far as HireVue and their employer-customers are concerned, that is enough. Yet, the customers may be operating with misplaced confidence. Polygraph tests were widely adopted by major employers for decades who bet their business success on their reliability only to eventually learn that they had believed in a chimera.

There are two other problems with the customer satisfaction defense. First, simply satisfying their customers only provides for the interests of customers. The value of justice is one in which the basic terms are those of which all reasonable participants can agree upon, even if they are disadvantaged by it from time to time. As such, a reasonable system of justice in hiring would at least acknowledge the justice claims of job candidates even if they are not handled in the same fashion as the claims of employers. While the question of justice in employment predates (and may outlive) HireVue, HireVue's practices appear to worsen what is already an arguably uneven playing field for job candidates operating under conditions of scarce jobs. The issue of inattention to job candidates becomes more acute when considering that government agencies also employ

[292] Mondragon, Aichholzer and Leutner (n 180) 2.
[293] ibid 4.
[294] Harwell (n 229).

AHS tools, such as those offered by HireVue.[295] Even if we decline to be concerned about prioritizing the interests of employers over candidates in private sector transactions, the standards ought to be higher for government agencies who are bound to the public interest and the values of participatory democracy.

A second problem with HireVue's defense is that we are in no position to know if their systems work as well as they claim. Instead, we are forced to trust what HireVue claims, which is made more difficult by the knowledge that HireVue's business prospects rest on projecting confidence in their systems. It is difficult to accept that automated hiring systems do what their manufacturers say when they are not evaluated by disinterested parties. This is not an impossible hurdle. Ajunwa argues that intellectual property and company secrets can be preserved through the validation of internal tests by certified external auditors.[296] Yet, HireVue does not currently submit to either external testing or validation of its internal testing by external auditors.[297]

This presents a problem for accountability and trust under conditions of information asymmetry. Adam Moore discusses this in his treatment of the bland assurances of government officials offered when accused of engaging in secretive and/or invasive actions for the sake of a stated noble goal. Moore labels such assurances the "just trust us" defense, which demands we place our faith in the stated intentions of an actor in a position of power.[298] The problem with this defense is one of moral hazard. Even an agent with noble aims is potentially corruptible when the consequences of acting against the interests of another are insignificant compared to their potential gains. Applied to HireVue, which has exclusive access to their system validation data, the company

[295] Kate Crawford and Jason Schultz, 'AI Sytems as State Actors' (2020) 119 Columbia Law Review 33.
[296] Ajunwa, 'Automated Employment Discrimination' (n 49).
[297] Harwell (n 229).
[298] Adam D Moore, 'Privacy, Security, and Government Surveillance: Wikileaks and the New Accountability' (2011) 25 Public Affairs Quarterly 141.

has incentives for distorting the truth in pursuit of profitability, especially if the truth about their systems is less flattering than their claims. This leaves us with a "just trust us" account of their systems, which is unsatisfying. So, while the reader may characterize some of the charges leveled against HireVue as theoretical or otherwise unproven here, the charges are also uncontested in the absence of qualified evidence about the company's products.

5.3.7.2 Cultural Fit

Another claim made by HireVue is that their assessments increase the potential for hired employee to "fit" in a company.[299] A risk of algorithmic psychometric systems is that, when working exactly as intended, their behavioral prediction can become behavioral prescriptions. Stark argues that the conversion of behavioral data into behavioral predictions potentially flows both ways.[300] Using the Facebook emotional contagion study as an example, Stark points to the pursuit of techniques that both gauge emotions and also attempt to influence them by projecting behavioral patterns back onto target subjects.[301] In the context of HireVue, cultural fit is transformed from a vaguely articulated ideal to a mechanistic inscription of norms. HireVue pursues a prediction of fit using their system to place job candidates on an "achievement index"[302] using an artificial intelligence technique known as supervised learning. First, HireVue conducts a workplace job analysis for the employer to build a data model based on current, successful employees.[303] The data model is trained by data scientists with measures of cultural fit and success derived from the workplace study.[304] From the model, HireVue constructs interview questions that

[299] Larsen and Taylor, 'United States Patent' (n 224) *background.*
[300] Stark (n 3).
[301] ibid.
[302] Larsen and Taylor, 'United States Patent' (n 262).
[303] 'Pre-Employment Testing Software | Online Gamified Assessments | HireVue' (n 236).
[304] Larsen and Taylor, 'United States Patent' (n 224).

are uniformly administered to job candidates. Resulting interview data is analyzed against the model and candidates are scored and/or placed on the index.[305]

Measures of cultural fit potentially reproduce undesirable hiring patterns. Writing about employment discrimination from automated systems, Ajunwa warns that machine learning algorithms potentially interpret the rather amorphous concept of "fit" as a strict rule, thereby rationalizing a concept with a degree of unintended certainty.[306] By example, in an infamous case, the major firm Amazon was compelled, after several years of development, to scrap an AI recruiting tool that systematically rejected qualified women candidates, particularly for jobs as software developers.[307] The system failure was attributed to its reliance on analyzing the resumes of successful hires over the previous decade, during which the majority of software developers hired into the company were men. Here, the system rationally evaluated "fit" as it was modeled in prior hiring practices, which, for whatever reason, under-selected women.[308] In automating hiring systems, reliance on a model of successful employees is risky because, in a data set that reflects a history of rewarding a particular social group and disfavoring others, there is a strong tendency for machine learning algorithms to reproduce the same pattern. As Ajunwa writes, "a company that tends to hire from a privileged and homogeneous community and then uses 'culture fit' as a factor in hiring decisions could end up methodically rejecting otherwise qualified candidates who come from more diverse backgrounds."[309]

HireVue responds to these charges in their stated commitment to diversity as well as claims

---

[305] Larsen and Taylor, 'United States Patent' (n 262).

[306] Ifeoma Ajunwa, 'The Paradox of Automation as Anti-Bias Intervention' [2020] Cardoza Law Review 55.

[307] Jeffrey Dastin, 'Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women' *Reuters* (10 October 2018) <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G> accessed 20 September 2019.

[308] The fact that Amazon decided to simply cancel the project rather than to learn from the errors and retool the algorithm suggests that the problem of cultural fit may be larger than a software patch.

[309] Ajunwa (n 49) 12.

about the power of their systems to help root out discrimination in hiring. As previously mentioned, HireVue emphasizes rigorous validation tests as proof that their systems are fair and not vulnerable to charges like those leveled by Ajunwa and others. However, without access to validation tests, we cannot say for sure how their system manages the problem of modeling, such as if it were confronted with a workplace as imbalanced as Amazon's software development division. We are faced with same the problem discussed for the defense of algorithmic psychometrics; we have to take HireVue at their word that their systems are fair, or at least fairer than the alternatives. Without more convincing evidence, the arguments of HireVue's critics are at least as strong as this defense by HireVue.

5.3.7.3 Mitigating Discriminatory Bias

The third category of claims made by HireVue is that the use of their system reduces or eliminates discriminatory bias from the hiring pipeline. The company claims this occurs in two ways. First, simply using their system reduces the participation of human decision-makers in the hiring pipeline. Human decision-makers are often guilty of allowing their implicit and explicit biases to shape their decisions. The bias in a human-made hiring decision may be obvious in some cases and very hard to detect in others. As one HireVue representative states, humans are the "ultimate black box."[310]

There are two objections to HireVue's approach to discrimination. First is an objection on functional grounds. In the two HireVue patents specifically concerned with discrimination, HireVue uses biometric identification methods to detect "a feature corresponding to at least one of age, gender, sexual orientation, disability, or race."[311] They then use statistical methods to evaluate

[310] Harwell (n 229).

[311] Loren Larsen and Benjamin Taylor, 'United States Patent: 10438135 – Performance Model Adverse Impact Correction'.

hiring patterns looking for evidence of discrimination. In addition to the use of audio data to detect these features, HireVue also uses facial recognition techniques that resemble those used in other commercial facial recognition technologies marketed to law enforcement among others.

The use of artificial intelligence to address hiring discrimination raises some concerns. First, detecting demographic categories and identities using biometric identification methods is notoriously inaccurate. HireVue's patents indicate that they use established facial recognition techniques.[312] They make no claims that their approach is significantly different from those in commercial use. Based on numerous studies, women and people of color are least likely to be correctly identified using established techniques.[313] Audio detection of identities fares little better. Recent research suggests that audio detection fails more often for African Americans by a significant margin.[314] This raises some concerns about the use of this technology to *identify* members who fit categories of protected employment classes, including the very same categories in which the technology is known to fail. Adding to this conundrum is the troubled history of assigning racial categories.[315] Even without technology, distinguishing between who does and does not fall into a particular race category is both politicized and contested and lacks a firm basis in science.[316] There is little evidence in HireVue's publications that they have grappled with the full complexity of race and somehow solved it using AI.

Identification using AI is similarly problematic for people who do not conform to gender norms. Decades of development of the machine learning systems that underlie facial recognition

---

[312] ibid.

[313] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', *Proceedings of Machine Learning Research* (2018).

[314] Allison Koenecke and others, 'Racial Disparities in Automated Speech Recognition' [2020] Proceedings of the National Academy of Sciences.

[315] Cheryl I Harris, 'Whiteness as Property' (1993) 106 Harvard Law Review 1707.

[316] Pilar Ossorio and Troy Duster, 'Race and Genetics: Controversies in Biomedical, Behavioral, and Forensic Sciences' (2005) 60 American Psychologist 115.

and other biometric techniques have operationalized gender as both binary and immutable, which essentially *erases* trans people from the possibility of being accurately identified by such systems for who they are.[317] The erasure continues with HireVue's technologies, which seek only evidence that a candidate meets an unstated gender criteria. Consequently, even while people with trans and non-binary gender identities are subject to discrimination in many walks of life, HireVue limits its efforts to those who fall into the current legal categories of discrimination, which do not include non-binary genders, and which only marginally protect trans women.

The socially situated bases of discrimination resist technological solutionism because they are not, in themselves, scientific issues. HireVue in its confidence in the power of AI, may be overpromising with a technological solutionist approach to a set of human problems that are far more complex and contested than their technologies can address.

An additional problem comes from HireVue's perception of the source of hiring discrimination. HireVue claims that because their system reduces human decision-making in the hiring pipeline, they remove human bias from the equation.[318] Equating the problem of racism with racist individuals is a familiar narrative that parallels the dominant legal and political approaches in the United States. As Hoffmann argues, this is the "bad actors" model of discrimination, which finds expression in a series of United States Supreme Court decisions since the passage of the Civil Rights Act.[319] The key limitation of this approach is that it fails to account for the systemic and structural features that create discriminatory outcomes. The legacy of slavery and Jim Crow, redlining of neighborhoods, uneven resource allocation for schools, and other

[317] Os Keyes, 'The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition' (2018) 2 Proceedings of the ACM on Human-Computer Interaction 1.
[318] 'Bias, AI Ethics, and the HireVue Approach' (n 181).
[319] Hoffmann, 'Where Fairness Fails' (n 232) 903–904.

policies that have generally resulted in greater access to opportunities for white people, are examples of forces of racism that are not tied to particular individuals. There is no particular individual to blame. As Hoffmann argues, when we focus on racism as a personal failing—a violation of established norms—we risk failing to consider how the norms themselves originate, how they become established, and whose interests they best serve. On this view, the complicated social history that has created conditions of discrimination are not stripped away with the removal of potentially racist hiring managers. Hiring discrimination is a much larger problem.

HireVue is emphatic in its marketing materials that it is deeply committed to the problem of hiring discrimination. They admirably claim to be committed to "actively promote diversity and aid in the achievement of equal opportunity."[320] HireVue also argues that their systems are an improvement on the alternative—relying on human decision-makers and their many biases. Presuming their systems do improve outcomes for historically discriminated job candidates, there is a risk that accepting a patchwork approach to a deeply complex social problem merely kicks the ball down the field, distracting us with a small gain achieved in a few workplaces from the much larger and much harder project of transformative social change. However, even the claims made about the systems are unsatisfying. We are given little reason to believe that HireVue has solved the well-theorized functional limits of detecting individuals who are vulnerable to biased hiring decisions with any precision. They provide no evidence other than their assurances. Given that their systems rely on technologies that have been shown to reproduce racism in their own right, we have legitimate reasons to doubt their claims. And while it is heartening to learn that HireVue partners with ethics and legal experts in pursuit of their antidiscrimination aims,[321] in the absence

[320] 'Bias, AI Ethics, and the HireVue Approach' (n 181).
[321] ibid.

of externally validated evidence, we have only HireVue's word as to the extent of these efforts.

5.4 CHAPTER CONCLUSION

The inscrutability of people's inner natures makes hiring challenging and potentially risky for employers. The promise of utilizing artificial intelligence to gain insights into the personalities of job candidates has produced a burgeoning market in automated hiring systems. HireVue, Inc. markets their AHS as the solution to hiring challenges that they view as originating in the fallibility of human decision-makers and the traditional, inefficient methods of screening candidates. In addition, discrimination in the hiring pipeline based on gender, race, age, and disability is a genuine problem that contributes to inequalities in opportunity and flourishing in contemporary societies. Here too, HireVue claims that their products are effective at combatting a problem that decades of scholarship, policymaking, and evolving hiring practices have yet to satisfactorily resolve. Yet, to gauge whether HireVue's technologies can solve these problems, they must be evaluated along with the company's core assumptions about the nature of the problem and its claims about the efficacy and desirability of its solutions.

In this chapter I have conducted a critical case study on the candidate screening technologies offered by HireVue. I sought to unpack the assumptions underlying HireVue technologies by employing elements of Value Sensitive Design to surface the human values implicated by the technologies. By analyzing HireVue patents, as well as publications and press accounts that discuss the technologies, I cataloged the company's claims about the features and functions of its technology. The application of artificial intelligence to generate insights about job candidates is characterized as *algorithmic psychometrics* by Stark.[322] I also identified a "master narrative" of HireVue technologies that includes the company's epistemic commitments and technological

[322] Stark (n 3).

claims. I investigate a set of moral values implicated by HireVue technologies, including justice, fairness, and accountability. HireVue gestures toward an interest in justice and fairness but provides little evidence of a commitment to accountability. In analyzing HireVue's claims, I found that their reliance on algorithmic psychometrics projects a strong faith in the power of artificial intelligence to quantify human experience. I criticized this approach as an assertion of a fixed rationality on an inconstant and dynamic subject. I further found that HireVue constructs aspects of their candidate assessments on the concept of "fit" within a company. I offered a critique of this approach as having the potential to reproduce prior patterns of hiring discrimination. Finally, I critiqued HireVue's stated commitments and technical approaches to fairness in hiring. I found that their technical approaches rest on methods that have been specifically contested as harmful and inaccurate for people of color, women, and people with non-binary and trans gender identities. I also found that HireVue's claimed commitment to anti-discrimination rests on rhetorical claims that are narrowly focused on a "bad actor" model of discrimination and are otherwise not supported with evidence, at least not any evidence that the public can examine.

CHAPTER 6 — REPUTATION AND POLITICAL JUSTICE

6.1 CHAPTER INTRODUCTION

My goal for this chapter is to conduct an analysis of algorithmic reputation—digital tools for decision-making—within a moral framework. I offer arguments about the moral status of algorithmic reputation situated in the liberal political theory of John Rawls. First, I argue that reputation is an appropriate frame for understanding certain technical practices and artifacts. Next I offer three arguments from Rawlsian theory that indicate how we can fairly assess persons as rational and reasonable members of a cooperative society. Ultimately, I close with a proposal for addressing problems of both justice and fairness for algorithmic reputation.

In egalitarian political philosophy, scholars generally seek to design society so that it recognizes both the fundamental liberty and the basic equality of all persons in respect to one another and as understood by the institutions that regulate their activities. In the egalitarian theory offered by John Rawls, which he labels "justice as fairness," a just society is conceived as a fair system of cooperation for mutual advantage. It is a system of rational self-interest coupled with a set of reasonable limitations and requirements aimed toward collective flourishing. Here, *fairness* is understood as meeting conditions that persons would generally agree to while *cooperation for mutual advantage* is understood to mean that people recognize that their good is intertwined with that of others. Fairness also means that there is a reciprocal relationship between persons and their societies that links together what one contributes and what one expects in return.[323]

[323] I offer a more detailed account of Rawlsian liberal theory below and in Chapter 3.

6.1.1 Fundamental Concepts

The political philosophy of John Rawls, to which he dedicated his entire career, is intended to provide a moral justification for democratic society.[324] Central to this theory is that the purpose of society and its governing institutions is to promote mutually advantageous cooperation under terms and conditions to which all can agree. Rawls's work spans decades and includes a long list of fundamental concepts and principles. I draw on several of these to conduct a normative inquiry into the societal function and effect of algorithmic profiling and automated decision-making, which I have labeled "algorithmic reputation." I discuss three of these concepts before conducting my inquiry: primary goods, the basic structure of society, and legitimacy.

First, distributive justice theories take as a given that there are certain material and non-material goods that are under society's control and also that there are reasons based in justice for allocating them in a particular way. These goods are considered *primary* goods. Second, the *basic structure* of society is the set of institutions that, in one way or another, regulate the lives of persons outside of (most) intimate human relations. Third, *legitimacy* is the basis from which a society is defensible to those whose lives are constrained by its rules and structures and through which a society maintains stability over time. These are concepts that are especially well-developed in the work of John Rawls. Primary goods and the basic structure are articulated in several volumes within his theory of justice called "justice as fairness." The subject of legitimacy is most fully developed in Rawls's theory on the requirements of democracy called "political liberalism," which he developed late in his career. What I argue here is that reputation lurks throughout Rawlsian theory and provides multiple entry points from which to analyze automated profiling and decision-making when understood as the digital expression of reputation.

---

[324] Freeman, *Rawls* (n 55).

6.1.2 Three Domains of Reputation

In the argument I offer in more detail below, I use the concepts of primary goods and the basic structure to construct a legitimacy framework for reputation. Entwined in this analysis is a tension in the construction of reputations between the responsibilities of assessors and those of the assessed. I describe this tension by considering reputational assessment across three domains of public and non-public live.

6.1.2.1 Primary Domain

In the primary domain of reputation, we owe all persons, first, basic moral respect; we are bound to recognize them as persons capable of rational self-interest and a reasonable commitment to justice *prior* to any determination as to their moral worth. This is reflected by the Kantian duty to recognize all of humanity not merely as means to our ends but as ends in themselves.[325] For both Kant and Rawls, persons (once defined) are accorded fundamental respect by virtue of their humanity. Ultimately, the *degree* of respect we accord others varies with what we know about them, but I interpret Rawlsian ethics as an expression of Kantian obligations applied to social cooperation that requires that we provide an opportunity for persons to demonstrate their moral worth prior to judging them. As discussed below, this is the basis of distributive justice—setting the initial conditions of cooperation and the social foundations from which persons form a sense of justice. It is only after these conditions are in place that subsequent evaluations of moral worth can be reasonably made. If society members fail to accord Kantian "ends" recognition to each

[325] Kant, 'Groundwork for the Metaphysics of Morals' (n 13). Kantian obligations motivate Rawls's philosophical stance, which I adopt here.

other at the outset of cooperation, it is unlikely that persons can make use of their talents and abilities to meet any subsequent expectations.[326]

6.1.2.2 Public Domain

In other reputation domains, the known or experienced public behaviors of subjects are the focus of assessment. Here, subjects make choices in light of the expectations indicated by the public rules that govern society and are judged based on their response to such rules. In a third domain, *non*-public rules of voluntary associations with others and individual affections operate to mediate various relationships and non-public systems of judgment (see Figure 6.1).

In Rawlsian theory, the basic structure of society is the set of institutions and doctrines that guide public life. These institutions, in turn, provide a framework of rights, expectations, and obligations. Both assessors and subjects are subject to public rules that indicate behaviors to be encouraged or discouraged, either by way of the hard constraints of law and order, or the design of public institutions that incentivize this rather than that endeavor or practice.

6.1.2.3 Non-public Domains

Not all domains of life are public or otherwise managed by public rules. Human relations include zones of intimacy between and among persons in which rules of conduct and expectations are understood to be flexible, even capricious. An example is the choice of romantic partners. Persons should not have to justify who they desire and why they desire them. Culture and other social

[326] Rawls, like Kant, equates personhood with rationality. Critics have argued that this permits the dehumanization of persons who are deemed to be lacking a rational will. For example, Nussbaum objects to Rawls's conception of the person because it risks erasing disabled and otherwise cognitively diminished agents from the realm of personhood. Nussbaum, *Frontiers of Justice* (n 134) . Young similarly faults Rawls by warning that Rawls's definition promotes a "normalization" of a set of persons that leaves others in position to be despised and abused Young, *Justice and the Politics of Difference* (n 126). I am sympathetic to these concerns but set them aside here and accept a potentially flawed account of personhood for the purposes of my argument. At least for those who qualify as persons under this account, the reciprocity demanded by a Rawlsian society indicates that we owe respect *first* and judgment second in order for members to proceed to a starting point of cooperation.

forces lend shape to the limitations of this (not all desires may be acted *upon*), as do some public rules. But within this scope are decisions left completely to the negotiations of individuals, either alone or within associations. Here, we might say that it is acceptable to treat others as means to our ends (fulfilling our desires or particular worldviews), so long as they accept this treatment and it does not transgress the boundaries of public rules, it does not cause them harm, and sociocultural mores permit it.

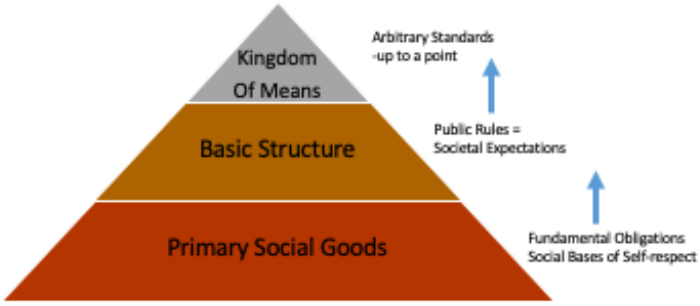

***Figure 6.1**. Obligations indicated by domains of reputational assessment*

6.2 Reputation as a Primary Social Good

In this section I expand upon the domains of reputation discussed above by placing them more firmly with Rawlsian constructs. First, I contend that there is some quotient of reputation that the institutions and individuals that constitute a just society are required to confer to individuals. I argue this by conceiving of reputation as a "primary social good" within a Rawlsian conception of society and his depiction of the necessity for "self-respect."

There are goods in the world that persons desire. Some of these goods are desirable because they are necessities, such as breathable air, fresh water, food, and shelter. In egalitarian theories of distributive justice, the list of goods is further refined to include those things that are necessary for persons to be "citizens," by which is meant that which enables persons to participate in public life. In Rawlsian theory, *justice as fairness*, primary goods play an important but possibly

overstated role. For Rawls, primary goods contribute toward a *flourishing* life.[327] However, it is not the goods themselves that denote flourishing. The assumption that drives egalitarian distributive justice is that there are things, tangible and intangible, that are not only valuable but are constituent to the construction of a complete moral person.[328] We refer to these as "goods," as in *what is good to want or to have*. In the political theory of John Rawls, something that "every rational man is presumed to want" is a primary good.[329] Beyond mere desires, the underlying point of distributive justice is to argue that possession of certain primary goods to a certain degree is the prerequisite for full participation in public life.[330]

Primary goods are closely linked in Rawlsian theory with his concept of the person as possessing the "two moral powers" necessary for full participation in society. Recall from Chapter 3 that of the two moral powers, one is a person's *rationality* through which they develop plans for their lives and identify the means for achieving those plans, which includes identifying necessary primary goods. It is "rational to want these goods whatever else is wanted, because they are in general necessary for the framing and the execution of a rational plan of life."[331] The initial list of primary goods offered by Rawls includes "rights and liberties, opportunities and power, income

---

[327] "Flourishing" is the inexact English translation for the Greek concept of "eudaimonia," or a good life by some definition. For example, flourishing may mean realizing one's essential purpose, or being able to pursue one's freely chosen life goals. See Adam D Moore, *Privacy Rights: Moral and Legal Foundations* (n 196) 39–42. Moore draws on accounts of flourishing from others including Philippa Foot, *Natural Goodness* (Clarendon; Oxford University Press 2001) and Nozick (n 76).

[328] *cf.* Anderson, (n 85), discussed in more detail in a subsequent footnote.

[329] John Rawls, *A Theory of Justice, Original Edition* (n 29) 62. Note that that account of primary goods offered by Rawls is but one conception. Other notable conceptions include those offered by Robert Nozick in *Anarchy, State, and Utopia* (n 76), and Michael Walzer's *Spheres of Justice* (1983).

[330] Rawls, like many egalitarians, argues that there is some minimal amount of some goods that are necessary for persons to exercise their two moral powers. However, many critics of egalitarianism overstate its emphasis on the distribution of primary goods and see a requirement for welfare-state economics the potential for the *re*distribution of goods. For example, Kekes sees Rawls's theory as a proposal for a welfare-state of guaranteed minimums for all persons, and that we would be required to harm some groups to produce equality for others. I believe this is a serious misreading. Rawls does not commit to any particular distributive regime. He emphasizes distributions only to the extent that there are sufficient conditions for justice. See: John Kekes, 'A Question for Egalitarians' (1997) 107 Ethics 13. Contrast with John Rawls, *A Theory of Justice, Revised Edition* (n 18) xv–xvi.

[331] Rawls, *A Theory of Justice, Revised Edition* (n 18) 380.

and wealth [and] a sense of one's own worth."[332] Samuel Freeman, surveying Rawls's portrayals across his works, adds to these "all purpose means … positions of office … the bases of self-respect" and those means that "are necessary to realizing citizens' fundamental interests and exercising their moral powers and pursuing their rational life-plans."[333] Freeman's depiction underscores the point that the distribution of primary goods is not the end goal of Rawlsian theory. By associating primary goods with a person's rational plans, they are presented as the foundations necessary for persons before they can be reasonably expected to act for themselves and to contribute fully to a cooperative, mutually advantageous society through which they flourish.

In the thought experiment offered by Rawls, the *original position*, and from behind a *veil of ignorance* in which persons know little about themselves and their specific interests, it is theorized that persons would desire a minimal distribution of primary goods, such as basic liberties and opportunities, because they are rational to want but also because they meet the requirements of participation in a society based in cooperation among persons who consider each other as equals. Understood in this way, goods that describe social relations are primary goods because they are *prerequisites* to both a person's own ends and their enablement as citizens.[334] Primary goods are the basis from which a person joins society as a free and equal member.

I contend that reputation is a primary good. Following Rawlsian theory, I argue that there is

---

[332] Rawls, *A Theory of Justice, Original Edition* (n 29) 92. Critics of Rawls's conception of primary goods have argued that conflating material goods, such as wealth, with immaterial goods, such as honor, confuses static and relational goods, potentially masking relations of power that shape their possession. *Cf.* Young (n 126).

[333] Freeman, *Rawls* (n 55) 478.

[334] A complementary view is offered by Anderson (n 85) in her account of egalitarian ethics. Anderson argues that primary goods are those things that are instrumental for ensuring relations of equality among members of a society and seeks to disentangle primary goods from facile notions of desert and obligation, departing from egalitarians that, on her view, over-emphasize the distribution of possession of primary goods as the point of justice. In her portrayal, justice is not specifically served by ensuring just distributions of any particular good in any particular way. Rather, Anderson argues that distributions provide a procedural path for liberal justice. The right of one-person-one-vote in a democratic society is a primary good not because voting is good in and of itself, but because it is fits into a scheme of requirements for a just society in a way that the possession of other goods, such as vacation homes and tennis lessons, does not.

some minimum quotient of reputational status that is required for the pursuit of rational plans and full participation in a society conceived as a fair system of cooperation for mutual advantage. This degree of reputation is what Rawls describes as the good of "self-respect." This primary good and its "social bases" motivate his privileging of liberal equality among the various principles of justice proposed in justice as fairness. The recognition of moral agency for persons is central to Rawlsian theory[335] and is the starting point for all further principles.

The minimum quotient of reputation is the portion we owe each other in recognition of a persons' moral worth as a human being; it is not only that portion of a person's reputation that is earned. The obligation to recognize the humanity of others is found in numerous philosophical traditions. In Kantian ethics, it is reflected in the formulations of the categorical imperative,[336] which are based in the presumption that all persons are free and equal and therefore presumed to be worthy of deep moral respect prior to subsequent assessment.[337] It is only from this presumption that other principles can be constructed. As embodiments of a rational will, we are compelled to confer and demand this degree of moral respect to others in light of their, and our own, rational wills. The principles that flow from Kant's categorical imperatives, and that lurk in Rawlsian principles, do not hinge on prior notions of desert or evaluation of another's worth.[338] The principles convey that each person is worthy of recognition for their humanity prior to any other considerations. Adapting Kant to the domain of reputation, we set aside that portion of our view

[335] As discussed in Chapter 3 and in n 327 above, the limits to personhood described by Rawls have raised objections, particularly from those concerned about the status of those who do not meet the definition of "rational." *Cf.* Nussbaum (n 134).

[336] Kant (n 13).

[337] See n 327.

[338] This contrasts with the view of some critics, such as Sterba, who argue that desert should be foundational to any system of justice—those who do not contribute to society deserve nothing from it. By centering self-respect as a primary good, Rawls argues instead that there are some goods whose absence so diminishes the person that we cannot expect them to exercise either rational eudaimonism nor a commitment to a cooperative society for mutual advantage. From this reading of Rawls, I argue that the good of self-respect is *prior* to desert. See James Sterba, 'Justice as Desert' (1974) 3 Social Justice Theory and Practice 17.

of others that concerns the most basic questions of their humanity in our initial evaluation. Each person, regardless of our feelings about them or their behaviors, is granted some minimal degree of reputational recognition.[339] This minimal degree of respect is an essential condition for self-respect and reciprocal respect for others.

Self-respect strongly affects a person's relation to others and is what motivates our rational pursuits. The goals we set for ourselves become worthwhile through their recognition as being so: "our self-respect normally depends upon the respect of others. Unless we feel that our endeavors are honored by them, it is difficult if not impossible for us maintain the conviction that our ends are worth advancing."[340] At the same time, feeling that our own endeavors are worthy situates us to be more inclined to appreciate the contributions of others, providing them with the motivation borne of self-respect to continue pursuing their rational ends. In a conception of society as a system of cooperation for mutual benefit, primary goods that encourage others to pursue their rational ends while reasonably adjusting those ends to satisfy the demands of cooperation with others, are goods that promote general, not just individual, flourishing.

It may be objected that self-respect is not necessarily or singularly conferred by others but is generated from within the person themselves. Indeed, Kant appears to agree where he argues that we are obligated to love ourselves and develop our own talents or risk a contradiction as bearers

[339] This raises difficult questions about how far we must go to defend a person's fundamental moral worth. Neither Kant nor Rawls offer a clear roadmap for how we should consider an "evil" person, though both suggest that such persons exist. My reading of Kant is that goodness is equated with rationality and that therefore only someone lacking rationality can be evil and everyone else, even a very bad person, deserves moral respect. This is tricky terrain given what we know of human history and depravity. However, I suspect Kant (and Rawls) hew close to Christian ethical tradition that requires we hold some fundamental regard for persons in recognition of our being reflections of God's own image and/or part of a divine plan. For example, Fleming refers to the primary aspect of reputation as "connatural" because it attaches to the person by virtue of their existence and prior to human judgment. (See Julia Fleming, 'The Right to Reputation and the Preferential Option for the Poor' (2004) 24 Journal of the Society of Christian Ethics 73.) Kant does not appeal to divine command but otherwise seems to hold persons (however defined) in a special category of consideration, possibly in adaptation of the Christian theology he sought to replace, brick by brick, with rationality.

[340] Rawls, *A Theory of Justice, Original Edition* (n 29) 178.

of rational wills who fail to cultivate those wills toward rational ends.[341] While it may be true that each person who possesses the two moral powers is equipped to produce self-respect, Rawls argues that whatever our position or condition, persons must develop a moral psychology through their interaction with others. Rawls demonstrates that persons do not develop self-respect on their own. Children develop their moral framework from the models of parents and other authorities. As adults, persons refine their moral psychology through associations with others and by coming to understand themselves through experiencing the requirements of cooperation and friendship. Ultimately, through these social processes, Rawls argues that we develop a "sense of justice" that is the expression of the two moral powers.[342] Self-respect, while based in the person and their potentiality, is socially co-constructed; subject and assessor jointly participate in reputational processes as members of social structures and this is what leads to a subject's reputation. As socially constructed, a reputation is vulnerable to social *de*struction where social structures and choices of individual assessors result in the denial of the subject's humanity. Without being given the chance to develop a sense of justice through social ties, one is denied the opportunity to locate the path toward moral worth.

An exemplar of failure in the allocation of the primary good of self-respect is the use of discriminatory standards in the treatment of others due to, for example, their race, ethnicity, gender, sexual preference, or abledness. Choosing such a standard is not only demeaning, it is a denial of the common humanity we share with others. Recent events in the United States have demonstrated how the good of self-respect may be systematically denied to African Americans and other non-white persons by police departments. This disrespect, which has many features

[341] Kant (n 13) 172–173.

[342] John Rawls, *A Theory of Justice, Revised Edition* (n 18) §70-72. (The Morality of Authority)

including disproportionate arrest and violence experienced by non-whites,[343] produces a dysfunction in the relationship between affected classes of persons and society as a whole. Similarly, the good of self-respect may be denied to disabled persons by airline regulations that generally permit disparate and demeaning treatment based on disabled status.[344] Denial of this important good is an obstacle to the pursuit of rational ends, and leaves little incentive for affected persons' endorsement of a society where it both limits their flourishing and actively demeans them.

Whether a person is entitled to any particular primary good is not in question when considering their baseline distribution. While we might attend to questions of desert when attending to the unequal distribution of primary goods for the good of all,[345] there is a baseline of distributions that is required in justice as fairness. This extends to immaterial goods such as self-respect as well as the basic liberties that are similarly necessary to the pursuit of rational plans for life. When a society denies the primary good of self-respect, whatever the reason, it denies something essential to the development and exercise of the two moral powers, which must be secured before persons can be seen as full participants in the scheme of mutually advantageous cooperation.[346] The first moral power is required for the realization and pursuit of a rational plan for life. Reciprocity indicates that constraints of one's ability to form rational plans are reflected in similar limitations in the ability to promote the good of others, diminishing the ability to exercise the second moral power of reasonableness. By this measure, self-respect is the core of the rational

---

[343] *cf.* Bryan Warde, (n 11).

[344] *cf.* Louise Hickman, 'The Avery Review | "On the Basis of Safety": The Forced Intimacies of Accessible Air Travel' (2020) 6 *The Avery Review* <http://www.averyreview.com/issues/48/on-the-basis-of-safety> accessed 29 June 2020.

[345] Rawls, *A Theory of Justice, Original Edition* (n 29) §13.

[346] I note that this is one expression of the "two moral powers" test in Rawlsian theory that appears frequently as a basis of justification for aspects of the theory. It is not without criticism. Brennan argues, for example, that this test is ambiguous and tells us too little about what is actually required for the realization of a flourishing life. See Jason Brennan, 'Against the Moral Powers Test of Basic Liberty' (2020) 28 European Journal of Philosophy 492.

and reasonable person from which they act in the world. In a society that fails to accord self-respect to some members while doing so for others, we cannot be confident in the standards by which we judge success or failure. Self-respect is both prior to desert and the basis upon which we begin to assess deservingness.

6.3 REPUTATION AND THE BASIC STRUCTURE

Reputation lurks within the very structure of society, particular within the institutions that guide and regulate human activity. In justice as fairness, Rawls employs the conception of the "basic structure" of society, comprised of the various constructs that regulate the relations and interactions of public life, including the distribution of rights, liberties, and obligations.[347] As a system of doctrines and institutions, the basic structure can be understood as a framework for decision-making about opportunities and limits. It is "the primary subject of justice" that defines each person's rights and duties, and influences their life prospects and expectations.[348] Freeman, surveying the many descriptions of the basic structure offered by Rawls, summarizes the collection as:

> … the design of the social and political institutions that structure daily life and individuals' decisions and actions, and which distribute fundamental rights and duties and determine the division of advantages of social cooperation. … The social institutions that make up the basic structure are the political constitution; the legal system of trials, property, and contracts; the system of markets and the regulation of economic relations; and the family.[349]

The basic structure includes all of the institutions and doctrines that construct society and regulate its conduct. In domain of government, the basic structure orders public life through the laws,

---

[347] Rawls, *A Theory of Justice, Original Edition* (n 29) §2. Among the descriptions provided by Rawls, the basic structure is "the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation." See also my discussion of the basic structure in Chapter 3.

[348] ibid 7.

[349] Freeman, *Rawls* (n 55) 464.

regulations, and systems of adjudication prescribed by government doctrine and mediated by government actors. Applied to contemporary society, this would seem to include the entire set of laws and regulations as well the various bureaucracies, political office holders, and civil servants. Many non-governmental institutions that "govern" people's lives, are also part of the basic structure. While these are only vaguely described by Rawls (*e.g*. "the economic system"), a consideration of lived society from the experience of daily life suggests that there are many institutions that govern in addition to those that are clearly governmental. The distribution of food, insurance markets, credit and lending, and non-public forms of transportation are examples of non-governmental institutions and constructs that regulate our lives by making some options available while limiting access to others.[350]

The basic structure constructs and enforces a significant share of a person's reputation, particularly as it is experienced in public life. It is most importantly a system of public rules that sets societal expectations for individual action. This point is developed by Rawls in the effects of the public rules set by the basic structure:

> What a person does depends upon what the public rules say he will be entitled to, and what a person is entitled to depends on what he does. The distribution which results is arrived at by honoring the claims determined by what persons undertake to do in the light of these legitimate expectations.[351]

Here, Rawls indicates that the rules guide *expectation, entitlement,* and *action* in society. These are the foundations from which reputational assessments in the domain of the basic structure are built. The basic structure is not merely rules but a social form that "shapes the wants and aspirations that its citizens come to have. It determines in part the sort of persons they want to be as well as

[350] While it is unclear exactly which institutions should be included in Rawls's account of the basic structure, others have argued that it should any holistic account of justice must include a larger scope of societal actors whose actions can be extremely consequential for others. *Cf.* Iris Marion Young, 'Taking the Basic Structure Seriously' (2006) 4 Perspectives on Politics.

[351] Rawls, *A Theory of Justice, Revised Edition* (n 18) 74.

the sort of persons they are."[352] The basic structure, through its various doctrines, encourages and discourages; it makes certain aspirations more attractive and more achievable than others, setting down the conditions and standards by which persons are judged as contributing and flourishing members of society.

This is most explicitly expressed by the effect of laws. A law that levies heavy fines and other sanctions for driving while intoxicated by liquor or drugs is intended not only to punish but to discourage certain behaviors, providing the grounds for negative assessments.[353] Government funded education programs encouraging designated drivers and limits on the consumption of intoxicants signal preferred behaviors and positive tags in the same domain. Non-governmental institutions that are part of the basic structure similarly indicate preferred and discouraged actions to direct behavior and provide bases of assessment. Economic institutions, such as banks and other aspects of the economic system, employ various means to set expectations and shape action. To encourage people to pay their credit card bills on time, companies threaten account cancelation, penalty fees, credit report entries, and other negative consequences. They may also offer rewards to "good customers" who pay on time or in full, such as special deals and discounts.

A characteristic of governing institutions, public and private, is their ability to issue rewards and sanctions. In the case of law, the sanctions are carried out by the state. In other domains of public life, institutions exercise their power to grant or withhold desirable goods, and to make life difficult for those who do not conform to their rules.

---

[352] ibid 229.

[353] This operates in both the public realm of treatment by courts and other institutions and private realms in which friends, family, colleagues, etc. who learn someone has violated the law may shift from or revise reputational assessments about the person.

6.3.1 The Basic Structure and Reputational Justice

I have argued that reputation lurks in the basic structure of society as described by Rawls. Unlike the depiction of our obligations in the assessment of others at the level of primary social goods, subjects of reputational assessment are understood to make choices that influence their public reputations. Here, the person is presumed capable of recognizing their rights and obligations and forming their aspirations within the bounds of the basic structure. Under these conditions and in exercise of the two moral powers of rationality and reasonableness, a person's reputation can be said to be a measure of their virtue. Yet, we might well ask, in virtue of what?

In the ideal society conceived by Rawls, the society is a reflection of pure procedural justice. The basic structure is comprised of institutions designed in an initial situation that is fair (the original position from behind the veil of ignorance) and produces the right public rules for the right reasons. It is therefore predicted to produce just results. It regulates, among other things, the allocation of a sufficient share of primary goods, including the good of self-respect, for persons to exercise their two moral powers and fully participate as citizens. The basic structure reflects the rational eudaimonism of individuals and the reasonable requirements of cooperation for mutual advantage. In such a system, the normative force of reputation expressed by the basic structure—the action guidingness of assessment—is therefore just. Persons are encouraged toward actions and pursuits all fellow citizens endorse (or at least tolerate) and are fairly recognized with positive reputational profiles to the extent they conform to these public standards. They are similarly discouraged from actions and pursuits that violate the requirements of justice.

Rawls offers his model of an ideal society as a means to evaluate our own. It is not surprising then that in actual societies we cannot be confident that the basic structure is an expression of pure procedural justice. In the United States, where a common depiction is that society was constructed with intentionality by high-minded individuals, we might claim that U.S. society hews close to the

Rawlsian ideal. However, a moment's reflection and observation about our world indicates that this is not the case. The origin story of the United States is complicated by the conquering and disenfranchisement of its original inhabitants, centuries of lawful human slavery, denial of gender equality, and other moral failures in its history that give shape to its present. Even in the fullness of time and various attempts to correct and ameliorate prior moral failings, most would agree that we do not live in a perfectly just society—and it is unlikely that one exists anywhere in the world.[354]

This has implications for the reputation standards that flow from the public rules of real societies. Recall that in Chapter 2, I described how reputation is a process in which there are multiple entry points for particular worldviews and priorities. In a basic structure in whose justice we are confident, the worldviews reflected in reputational processes are also just. However, in actual societies, the procedures that construct public life are clouded by historical legacies and contemporary interests that introduce worldviews and priorities that are not universally endorsed. It seems likely that the basic structures we know, at least some of the time, fail to encourage behaviors that exemplify a fair system of cooperation for mutual advantage while, at other times, are actively encouraging behaviors that impede such a system. Non-ideal conditions give rise to non-ideal reputational standards that, at least some of the time, would not meet with our reasoned moral reflections. Given these conditions, we have no choice but to place limits on our faith in their overall fairness. For example, if the basic structure permits, or does not sufficiently incentivize otherwise, a highly profitable retail chain to pay productive workers less than a living wage, particularly in locations where the firm's market domination has eliminated other economic

[354] A telling example is that, despite decades of law and dialog attempting to address the legacies of slavery and Jim Crow, claims of persistent racial injustice and inequality in the United States are routinely made and supported by grim evidence. *Cf.* Evan Hill and others, 'How George Floyd Was Killed in Police Custody' *The New York Times* (31 May 2020) <https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html> accessed 30 June 2020.

opportunities, it becomes unclear if we can fairly assess persons as "deadbeats" if they seek public assistance while employed by the firm.[355] We may engage in other deliberations about the moral obligations of employers and the desert of workers but we can also conclude that the conditions described muddy the easy assignment of a denigrating reputational assessment. Rawlsian justice as fairness suggests that before we can fairly judge persons operating under the rights, obligations, and expectations of an unjust basic structure, we may be required to deliberate on the fairness of the conditions provided by that structure and potentially intervene to improve conditions as a matter of justice.[356]

The basic structure provides a reputational construct by explaining "the good of activities"[357] and indicating the character of the citizen as rational and reasonable. A basic structure that is just, and serves as an expression of pure procedural justice is one that unambiguously assesses persons as not only in conformity with societal expectations but with moral expectations as well. A basic structure that is not just, or not entirely so, provides only ambiguous guidance in its public rules for individual conduct and lightens the moral weight of ensuing reputational assessments.

6.3.2 Recourse and the Basic Structure

For the basic structure as a system of rules and expectations to function in a liberal conception of society, the system must demonstrate respect for the two moral powers of rationality and reasonableness. Rational persons are treated as reasonable when provided the means to understand their assessments and the opportunity to form alternative plans based on that understanding.

---

[355] Clare O'Connor, 'Report: Walmart Workers Cost Taxpayers $6.2 Billion In Public Assistance' [2014] *Forbes* <https://www.forbes.com/sites/clareoconnor/2014/04/15/report-walmart-workers-cost-taxpayers-6-2-billion-in-public-assistance/> accessed 3 July 2020.

[356] This is most explicitly demonstrated by the principle of fair equality of opportunity and the difference principle in justice as fairness. Rawls, *A Theory of Justice, Revised Edition* (n 18) §11–17.

[357] ibid 350.

Furthermore, exercise of the two moral powers requires that assessments carried out through public rules are deliberative, affording opportunities for contestation and adjudication. For example, the legal system in such a society is unlikely to be accepted if it simply designates persons as criminals without due process. Instead, it includes processes through which a person is afforded the opportunity to maintain their innocence and contest evidence of their guilt. The provision of due process is essential to the sense of legitimacy in the coercive power of this aspect of the basic structure.

As defined by computer scientist Suresh Venkatasubramanian and philosopher Mark Alfano, the ability to understand and contest a decision can be described as "recourse."[358] Recourse is an expression of personal agency and, as Venkatasubramanian and Alfano argue, it is in itself a "modally robust good;" it affects not only a particular decision but potentially a network of other decisions and plans that affect one's life.[359] For example, being denied a loan is not merely a denial of the good of a loan but also a denial of access to the goods the loan was intended to secure. Therefore recourse is modally implicated for each of the goods. Recourse provides the ability to foresee or understand decisions. We can contrast it with subjugation to the unaccountable will of others while being denied a path to a more favorable outcome. If we know what is required for a positive decision, we may be able to do what is necessary to meet the requirements or challenge them as unreasonable.

Taken together, the basic structure has two general requirements in respect to reputational fairness. First, the public rules of the basic structure must be just. Persons cannot be reasonably obligated by expectations. Second, public rules of the basic structure must be accompanied by

[358] Suresh Venkatasubramanian and Mark Alfano, 'The Philosophical Basis of Algorithmic Recourse', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2020) <https://doi.org/10.1145/3351095.3372876>.

[359] ibid 285.

opportunities for recourse. Where the basic structure does not offer either a just construction or recourse for its expectations, ensuing reputational assessments are not guaranteed to be fair and become legitimate targets for direct moral evaluation.

6.4 A KINGDOM OF MEANS?

Immanuel Kant depicted a moral framework in which there is a system of moral laws that are universally applicable to all rational beings, including ourselves. The first formulation of the categorical imperative instructs us to choose action in consideration of all persons' rational wills. This collection of wills—of all rational beings—is characterized as the "kingdom of ends." Moral law is to recognize the fundamental moral worth of all subjects to this kingdom of ends and must demonstrate respect for persons as "ends in themselves." Moral laws must be simultaneously acceptable to all members of this kingdom, "a union of rational beings under common objective laws,"[360] as it is acceptable to ourselves. Rawlsian justice follows a similar line in the construction of an ideal society whose public institutions and doctrines are endorsed by all members because they follow from a pure procedural justice that reflects the good will of all persons, described by Rawls as their "sense of justice."[361] The sense of justice gives rise to a "political conception of justice" in which persons share in the endorsement of a set of public rules and doctrines, whatever else they believe as their good.[362]

Not all aspects of the society conceived by Rawls are directly regulated by the basic structure and its public rules. Rawls departs from Kant in recognizing that a democratic society, characterized by liberty and autonomy, must accommodate the "reasonable pluralism" of its

[360] Kant (n 13) 181–183. For an interesting interpretation of the "kingdom of ends," including its application to Rawlsian theory, see Korsgaard (n 38) 99–110.
[361] The sense of justice is most fully explained in Chapter VII: John Rawls, *A Theory of Justice, Revised Edition* (n 18).
[362] John Rawls, *Political Liberalism* (n 14) 12.

members and their varying conceptions of goodness.[363] These primarily include those conceptions that are reasonable but also include some degree of those that are *un*reasonable, for which a degree of tolerance is also expected.[364] As such, society may not be entirely a kingdom of ends, as Kant suggested. We may be generally required to demonstrate deep moral respect for the worth of persons, but there are domains in life in which this is not the most immediate source of action. Is it possible then that there is also a *kingdom of means* in this conception? Recall that, as rational and autonomous, persons are conceived as able to recognize the requirements for their own flourishing through the formulation and pursuit of rational plans. The conception of one's good is bound up in their "comprehensive doctrines," which express their values, including the bases of their relationships and associations.[365] These are formulated in moral, philosophical, and religious commitments, which persons hold themselves and, at the same time, likely share with others. Elemental to a comprehensive doctrine are the person's affections and judgments in regard to others, which shape their pursuits and plans. There are limits to the extent to which we can pursue rational self-interest, indicated by exercise of the second moral power of "reasonableness." This implies, that within some bounds of one's comprehensive doctrine and relationships we are not *always* expected to be reasonable and may pursue self-interest for the sake of oneself. We are free, up to a point, to be *un*reasonable.

One's comprehensive doctrine features in what motivates rational plans and the sense of what constitutes a flourishing life. While Rawls does not attend much to persons' *irrational* desires and interests, such as our affections for others, love of beauty and so on, he allows that these too play

[363] ibid 441.
[364] ibid 441f.
[365] ibid 13.

a part in an individual's conception of the good.[366] We enact those aspects of our rational life plans in associations with others that are non-public, meaning those domains of life in which we are free to construct or adopt rules and standards and be voluntarily held to them without the force of public rules that apply to all persons.[367] Within a religious community, all may agree that they are commanded by a sacred text to forego leavened bread at a specific time of the year. This is not a public rule, but one observed by the members of the association in keeping with a shared commitment to a doctrine. In such communities, as in other associations, unique, non-public standards of reputational assessment attend the non-public rules. A softball team includes expectations and norms that guide the behavior of team members, both as players of an organized sport and members of a team culture. Fulfilling the expectations of such an association are generally of concern only to those within it.

Within one's comprehensive doctrine, affections and many relationships operate within a kingdom of means and are freed, to some extent, from the demands of society as a whole. Here there is a degree of freedom to employ reputational standards that are unregulated by the public rules of the basic structure. We are free to decide whether or not to enter into friendship with someone based entirely on their physical appearance, political views, or choice of cologne, whereas such standards would likely be rejected as standards for admission as a student to a public university. We are thereby free to regulate many of the relationships in our lives as we please. However, a kingdom of means, if it exists, is sharply limited in its scope. While there are life domains and associations where rules and standards are open to latitude and primarily subject to local standards and constraints, this occurs within the broader constraints imposed by the public

[366] This is implied in Rawls's depiction of comprehensive doctrines as underlying our close relationships and also in his allowance that some degree of unreasonable conceptions of goodness are likely to be tolerated within the plurality of conceptions held in a free society.

[367] Rawls, *Political Liberalism* (n 14) 140.

rules of the basic structure. Rawls offers an illustrative, if tart, example; "while churches can excommunicate heretics, they cannot burn them. …"[368] While churches are associations that govern many aspects of their members' lives, a church is limited to the extent it can limit the general liberty of all persons. The basic structure provides the background conditions in which a church conducts itself, such as by including provisions for the free practice of religion, and the holding of church property, yet it also limits its reach. Where a church designation like "heretic" and its consequences conflicts with the priority of liberty for all, it loses force and is subservient to the public rules that would prevent heretic-burning. While there are many aspects of private life that are not included in the basic structure, there is some threshold upon which private acts become a matter of public concern and thereby, subjected to the domain of the basic structure.

What this example reveals is that the relatively weaker standard for rulemaking and assessment in associations and through affections is not total. The kingdom of means turns out not to be much of a kingdom at all. It is more of a domain within a larger kingdom. The political conception described by Rawls provides a sense of "background justice" that places limits on this domain of means. Similarly, associations, although constructed by shared comprehensive doctrines, exist against the "just background institutions within which associations and groups exist, and by which the conduct of their members is restricted."[369]

6.5 THE REQUIREMENTS OF LEGITIMACY

6.5.1 Coercion

Society is a system of authority enacted through its institutions and enforced against individual wills. While Rawls extols the priority of liberty in justice as fairness, the commitment to liberty is

[368] John Rawls, *Justice as Fairness: A Restatement* (n 33) 11.
[369] ibid.

bound up with our obligations to others in recognition of everyone's fundamental equality; we owe due consideration to the liberties of others.[370] Such considerations are imposed as limits upon our liberties. While social contract theories imply that members voluntarily submit to the authority that governs them in an exchange of liberties for benefits, we know that in practice this conception of society is purely formal; voluntary submission is merely hypothetical. As members of society "we enter only by birth and exit only by death."[371] We accept the coercive and inescapable power of governing structures as a feature of the social contract, and also because for most people there really is no choice. It may be possible to choose among societies and just pick up and leave one for another and thereby evade coercion, but in reality, most people lack the desire or the means to do this.[372] Even for those who would choose to shop for a more favorable society, coercion is present in any place governed by the rule of law. Coercion is a fact of the society we know, and that which John Rawls envisioned. For this coercion to be considered just, it must be viewed as legitimate.

Before proceeding, it is necessary to be clear about what I mean by "coercion." First, I mean that coercion is a force of constraining influence on behavior or attitude. I do not mean to characterize coercion as necessarily pejorative. We engage in and are subject to systems of coercion whose purposes may be agreeable or disagreeable, morally acceptable or unacceptable. However, coercion is a normative force; it is more than mere *suggestion.* Rules and norms that are coercive make non-compliance unappealing or impossible. For a rule or norm to be coercive, it is

---

[370] Rawls, *A Theory of Justice, Original Edition* (n 29) 60.

[371] Rawls, *Political Liberalism* (n 14) 135-136. It is interesting that Rawls does not attend to the status of persons who join a society after birth, as immigrants do. This is expressed in Rawls's requirements that his ideal society be closed and self-sufficient for the thought experiment to work, which is a target of criticism. See John Rawls, 'Kantian Constructivism in Moral Theory' (n 59).

[372] While the ability to leave a society one does not endorse is largely fixed by individual circumstance, the *desire* to exit is likely a reflection of the level of acceptability of the society and the availability of a better option.

accompanied by a foreseeable consequence. Coercion differs from *manipulation* in that agents are typically aware of how they are being directed or instructed to act.[373] In this way, coercion implies choice. Choosing to defy a rule or norm places one at risk of penalty. Choosing to conform invites reward, even if the reward is the merely the absence of penalty. When coercion is effective, it serves to reduce the scope of rational choice.[374]

The coercive effects of a social structure are most visible in a system of laws. In the intoxicated-driving example above, coercion is affected by the penalties provided by law. One *could choose* to drive a car just after drinking a six-pack of beer but would be at risk of serious consequences. The law, by providing heavy fines, loss of driving privileges, possible jail time, etc., is intended to make driving while intoxicated not only undesirable but to limit one's rational choice toward a particular end—driving only when sober.

6.5.2 Coercion and the Problem of Stability

In his later works on the political work of sustaining a just society, Rawls is concerned with what would be required to ensure fairness over time, and in particular the requirements of stability.[375] Stability is a problem for the society idealized by Rawls because many of its features are voluntarily agreed to only formally. Even in an ideal society that is endorsed by all, "political power is always coercive power backed by the government's use of sanctions."[376] As a result, citizens of this society must maintain their endorsement of society's basic structure, including its public rules, to maintain its enduring legitimacy. As discussed above, the basic structure is not the only domains of life in which persons are subject to rules and standards. Associations also impose

[373] I adapt the "forthrightness" of coercion from Daniel Susser, Beate Roessler and Helen Nissenbaum, 'Technology, Autonomy, and Manipulation' (2019) 8 Internet Policy Review 4.
[374] ibid.
[375] Rawls, *Political Liberalism* (n 14) §4.
[376] ibid 136.

rules and standards. A key distinction between these associations and domains directly regulated by the basic structure is the degree of voluntariness of these associations. While the rules and standards of voluntary associations are potentially coercive upon their members, one is theoretically free to conform or reject the demands of the association by exiting it.[377]

The public rules of the basic structure are largely involuntary and generally coercive. It is for this reason that a political conception of justice cannot be guided by any particular comprehensive doctrine; no single comprehensive doctrine or conception of the good can dominate society if all members are to endorse it as just. A fair but pluralistic society requires an "overlapping consensus" among comprehensive doctrines.[378] Holders of diverse comprehensive doctrines need not agree on all matters so long as there is a sufficient amount of overlap in which each conception of the good includes principles that support an agreement about political justice. Without this consensus, a society is unlikely to be stable, or stable "for the right reasons,"[379] meaning reasons that accord with the diverse conceptions of the good held by citizens—and so endorsed by them rather than imposed. A system that is stable only because of its coercive power imposed upon citizens is a system of domination by some over others rather than one of general endorsement; such a society, in which liberal equality does not hold for all members, is unlikely to remain stable over time.

6.5.3 Public Reason

The overlapping consensus of comprehensive doctrines is just one element of the a political conception of justice that stands apart from our individual conceptions of the good and from which

[377] I draw the conclusion that voluntariness is what sets non-public from public associations apart. This distinction does not take into account the actual choices persons have in their lives, which may be sharply limited by social bonds, including bonds that are dysfunctional.

[378] Rawls limits this only to comprehensive doctrines that are "reasonable," by which he means they endorse, whatever else they endorse, a liberal conception of justice such as justice as fairness. *Cf.* Rawls, *Political Liberalism* (n 14) 149; Freeman, *Rawls* (n 55) 366–367.

[379] Freeman, *Rawls* (n 55) 368.

we can adjudicate conflicting claims about political matters of importance.[380] To conduct these adjudications on the most essential matters, Rawls argues that we must appeal to the idea of *public reason.* Recall that in Chapter 3, I described public reason as a rigorous process of deliberation on matters of importance that should produce decisions that all can accept. Public reason provides the guidelines of inquiry, the principles for reasoning, and the rules of evidence that are used for questions about how the coercive power of society should be employed.[381] It is used to ensure that the institutions of the basic structure, including the rights and obligations that flow from it, are perceived as legitimate.

While the open, transparent, and rational deliberation of public reason seems like it could indeed provide the sense of legitimacy demanded in an ideal society, we must consider that, in practice, public reason is at best aspirational. It demands thorough and fully accountable deliberations and all agreements subjected to it are to be based in objective forms of evidence, such as scientific fact. We know from contemporary events that it is extremely challenging to identify a set of "objective facts" upon which all can agree. Scientific fact, which has enjoyed periods of great authority in many societies, is vulnerable to political contestation, even when pursued with rigor.[382] Furthermore, there are political questions that cannot be answered by science because they are matters of value rather than indisputable fact. There are also domains of decision-making in which openness and transparency are either not possible or undesirable, such as in matters of national security or matters in which transparency exposes persons to violations of important rights, such as reasonable privacy.

This complexity is perhaps why Rawls seeks to limit the scope of public reason only to

[380] Rawls, *Political Liberalism* (n 14) Ch. VI.

[381] ibid.

[382] I note, for example, politicized debates about evolution, climate science, and vaccines, as evidence of seemingly settled scientific facts failing to convince large numbers of people to share a unified truth.

"constitutional essentials," which only pertains to features of the basic structure that concern the most essential expressions of power. Indeed, as conceived by Rawls, the open, transparent, and evidence-based reasoning of public reason is limited to only those matters that affect the basic liberties of persons, by which he appears to mean those liberties potentially affected by public institutions and frameworks. Rawls appears to be concerned about the difficulty of reaching agreement even on these issues and is therefore conservative about its application elsewhere lest we lose focus on the most essential matters.

Here I offer a contrary view. Even the securing of basic liberties using the rigors of public reason is likely to be a task that challenges a society. By rejecting the application of public reason in a wider range of contexts, Rawls assumes this will improve the chances of public reason succeeding in the few domains he allows. There is no reason to make such an assumption. Public reason is challenging in *any* domain of use. For this reason, we would be better off viewing public reason as aspirational rather than formal; public reason can be the aim of many domains of deliberation while being an aim we cannot always attain. As such, this frees us to attempt its application wherever it can provide guidance and legitimacy to the exercise of institutional power. I suggest that, well beyond the limited scope of "constitutional essentials," public reason should be applied to any domain in which rights and obligations are made explicit, including those domains concerned with standards of assessing persons.

This view adapts "the broad view" of public reason proposed by Jonathan Quong as means of making decisions in a wide range of decision contexts.[383] As Quong reminds us, public reason is derived from Rawlsian principles of fairness and respect for persons.[384] We demonstrate that

[383] Jonathan Quong, (n 116) 233, 234.
[384] ibid 246.

respect by working to reach agreement on the use of coercive power through a process of open and transparent deliberation. In a society characterized by reasonable plurality, public reason provides a basis from which to arrive at decisions that everyone can reasonably accept. The broad view of public reason accepts that it can be a demanding process and universal acceptance may be difficult, even impossible to achieve in many contexts. And yet, in pursuit of the goals of fairness and respect, we should try to employ public reason when and where it can be successfully applied.

Recall that I have stated that reputation lurks in the basic structure of society. The public rules and institutions of the basic structure articulate a person's rights, obligations, and expectations, and thereby outline the standards by which they are to be assessed in a public life. Because the basic structure is the foundation of the political conception, which is coercive in its effects, its reputational standards are also coercive. To maintain their legitimacy such standards can be subject to public deliberation and may require the rigorous standard of public reason. We see aspects of this ideal expressed in the United States, where the Fair Credit Reporting Act places accountability requirements upon the financial industry and their methods of assessing a person's creditworthiness.[385] While the specific provisions of the FCRA have been criticized,[386] a public rule with some resemblance would likely be required in the society idealized by Rawls because of the coercive effects of credit ratings upon the lives and choices of persons. Regulations that provide openness and accountability to members of society contribute a measure of legitimacy to the institutions that govern our lives.

[385] Fair Credit Reporting Act 1970 (United States Code) 1361.
[386] *Cf.* Pasquale (n 167).

6.6 Algorithmic Reputation and Information Justice

I have argued in this dissertation that computationally mediated profiling and decisions systems can be conceptualized as systems of reputation, which I have labeled "algorithmic reputation." Here, I argue that we can apply certain requirements of the ideal society conceived by John Rawls to the domain of algorithmic reputation.

What is the point of using reputation as a means of conceptualizing profiling and decision-making performed by code? For the most part, it is about the power of language. Algorithmic profiling and automated decision systems are labels that denote engineering efforts. Reputation is a term to describe a process of human relations and reflects the standards and expectations constructed by society and its institutions. We are more likely to view reputational processes, particularly as I have described them, in normative terms. Similarly, when offering a position about the actors and frameworks that participate in the application of coercive power, I suspect we accomplish more by focusing our gaze on the human beings involved to a larger extent than focusing on their technical inventions.

I present three arguments to show why systems of algorithmic reputation are of moral concern. First, I argue that systems of algorithmic reputation diminish the good of self-respect for subjects by failing to recognize their humanity and by treating them as means rather than ends. Next, I argue that algorithmic reputation practices by institutions of the basic structure reflect whatever injustice exists in that structure. Additionally, I argue that the basic structure implies a set of conditions on the application of its expectations. Finally, I argue that algorithmic reputation may fail to meet the legitimacy requirements of a fair and just society as envisioned by Rawls and ought to be subject to the rigors of public reason.

6.6.1 Algorithmic Reputation and the Good of Self-Respect

I have argued, following Rawls, that self-respect is an essential feature of human flourishing. It is foundational to the rational autonomy of persons and the conditions under which other primary goods may be enjoyed. Self-respect is not only important to the persons but also for society generally in that it creates the conditions from which a person is motivated to participate in society for mutual advantage. I have also argued that we are obligated to accord a foundational amount of respect in making assessments of others that is prior to questions about what a subject deserves. It is assumed that subjects deserve, *a priori*, to be recognized for their humanity.[387]

Systems of algorithmic reputation that are instantiated as automated profiling and decision-making systems may be capable of accurately assessing persons based on their actions and may be found to make decisions about them that we can all endorse. However, there are reasons to be suspicious of such systems due to their functional constraints and the social contexts of their implementation that can deny persons' their humanity. There are two sources of concern. First, algorithmic profiling systems do not experience human beings as complete or living organisms but as collections of signals. They cannot confer the respect of recognizing a subject's humanity because they do not recognize humanity.[388] The various signals acquired by algorithmic systems are discrete and processed independently by multiple systems employed separately based on the

[387] As discussed in n 327, there are difficulties with this assertion. Some would argue that a person can be so vile in their actions as to lose even this trace of respect. And yet, I interpret Kant and Rawls to hold that, so long as a person is "rational" (and for Rawls, capable of the two moral powers), they cannot be entirely denied moral worth. It is the absence of rationality that gives rise to evil. This view is controversial both to those who hold that there is pure evil in the world that demands total retribution and also to those who question an assertion of moral worth as requiring a baseline of capacities that not all persons possess. In either case, it leaves us with an ambiguous class of persons who can be said not to be rational (but by whose judgment?) and therefore appear to fall outside of moral worthiness. Even here, we must have *some* obligations. In the matter of reputation, the question of rationality is circular—it is also an assessment. As such, I cannot confront this question here, however; I share some of Nussbaum's concerns about Rawls's conception of the person as prior their moral consideration, which I discuss in Chapter 3.

[388] This may raise the specter of a metaethical discourse on human essence or ontology, such as that raised in the work of Emanuel Levinas and others. I do not engage with this topic here but note it for further exploration.

requirements of design and implementation. By example, the HireVue employment screening system described in Chapter 5 processes video, audio, and interaction data, acting on each flow using various processing algorithms and constructs a score from various aspects, such as "facial action units" and "audio features."[389]

Based on information provided in patents held by HireVue, audio data is analyzed by diagramming and comparing audio waveforms against a model based on other audio waveforms.[390] Waveforms that resemble models of tone tagged as "cheerful" or "anxious" are assigned a similarity score based on their correspondence. Natural language processing algorithms (NLP) compare collections of audio waveforms against libraries that have been previously tagged with language units, looking for pattern matches in a similar fashion. Meanwhile, other algorithms process other signals acquired on the platform. The AI does not distinguish human beings as bodies with minds, only the signals they have been trained to recognize and process. Algorithmic profiling, as exemplified by the technical description of HireVue, is limited in its analysis to discrete aspects of human expression. With relatively little modification and by reference to different data models, the same algorithms could be used to identify salient features of bird song or engine noise. As a result, when the HireVue issues a score that is influential in hiring decisions, it does so without first recognizing the basic humanity of its subject.

Writing about the dehumanizing effects of surveillance infrastructures, sociologists Kevin Haggerty and Richard Ericson similarly observe that the reduction of humans into discrete informational units has destructive effects on selfhood and social structure. By converting features of human activity into commodifiable units, data processing technologies engage in "abstracting

[389] Harwell(n 229).
[390] *Cf.* Loren Larsen and Benjamin Taylor, 'United States Patent: 8751231' (n 262).

human bodies from their territorial settings and separating them into a series of discrete flows. These flows are then reassembled into distinct 'data doubles' which can be scrutinized and targeted for intervention."[391] HireVue, by operating on subjects as "discrete flows" separates persons from the details of their humanity and acts only upon those details, thereby denying them the respect of recognizing them as humans rather than objects. While it is conceivable that a collection of sensing systems and machine learning algorithms might work together to approximate human sensing, processing, and response in a human-like way, passing the so-called "Turing Test," there is no evidence that such systems currently exist or will exist anytime soon.

The second way algorithmic profiling fails to recognize subjects' humanity is by employing classification logics that misinterpret human identities, leading to failures of respectful identification. Failures in classification frequently fail for identities attached to classes or persons that face discriminatory exclusion and violence in society. As discussed in Chapter 5, and articulated by gender and technology scholar Os Keyes, prominent machine learning research operationalizes gender as both binary and immutable, thereby erasing non-binary and trans identities in their classifications.[392] When this research is instantiated into AI artifacts that identify and classify their subjects, this erasure expresses a failure of respect for the person and potentially opens them to harm when gender identification systems are used in situations that carry existential risk.[393]

Many algorithmic recognition systems fail to identify faces with darker skin tones[394] and

[391] Kevin D Haggerty and Richard V Ericson, 'The Surveillant Assemblage' (2000) 51 The British Journal of Sociology 605, 606.

[392] Keyes(n 317).

[393] Anna Lauren Hoffmann, 'Data Violence and How Bad Engineering Choices Can Damage Society' (*Medium*, 30 April 2018) <https://medium.com/s/story/data-violence-and-how-bad-engineering-choices-can-damage-society-39e44150e1d4> accessed 24 January 2019.

[394] Buolamwini and Gebru (n 313).

interpret the speech of African Americans less accurately than for whites.[395] Misidentification and misunderstanding is disrespectful regardless of the actor but is worsened in the case of automation which carries with it a perception of mathematical infallibility and operates at greater speed and scale than human decision-making—both of which impede accountability and the detection of error. In a recent case, police officers accepted, apparently without double-checking, the conclusion of a facial recognition system that tagged Robert Williams, an African American man, as a suspect in a shoplifting case. A close look at the reference image should have made it obvious that it was not a picture of Mr. Williams, but police officers arrested him anyway based on the conclusion of the algorithmic system and subjected him to hours of detention and a court appearance even after acknowledging the system's failure.[396] The algorithmic system, and its human users, failed to demonstrate the respect of identification to Mr. Williams and subjected him to needless humiliation and risk. Similar errors of identification can be committed by human investigators and often are. Whether by algorithmic system or by human beings, misidentification diminishes the self-respect of the subject. In the case of law enforcement, where there is an ever-present threat of violence, misidentification has potentially fatal consequences, deepening the disrespect.

As argued by Rawls, self-respect has social bases; it emerges through the development of a moral psychology shaped by our interactions with others. Self-respect is the basis of rationality and a requirement of liberal equality. Our notions of desert are based in the assumption that a person is rational and therefore capable of exercising the two moral powers for their own good and as functioning members of society. Algorithmic systems that deny a person's humanity diminish

[395] Koenecke and others, (n 314) .

[396] Kashmir Hill, 'Wrongfully Accused by an Algorithm' *The New York Times* (24 June 2020) <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html> accessed 3 July 2020.

their ability to exercise the two moral powers. When this occurs, our ability to legitimately assess or hold to account what a person does is hampered by a failure to produce the required conditions from which a person can reasonably form and pursue rational plans for their life.

6.6.2 Algorithmic Reputation and the Basic Structure

The basic structure of society contains the institutions and doctrines of society including its public rules. From the public rules, persons understand their rights, obligations, and expectations in those domains of life controlled by the basic structure. Systems of algorithmic reputation can be found in the basic structure of many societies. Governmental institutions, including law enforcement agencies, public schools and universities, and various bureaucracies are among the contexts in which algorithmic profiling and decisions systems—systems of algorithmic reputation—may be used. The employment screening platform HireVue, for example, is used by government agencies for selecting employees[397] and its designers conceive of its use in college admissions.[398] While the exact content of the basic structure is not fully specified by Rawls, we can be most confident that it includes government agencies, public schools and universities, and civil and criminal justice systems, among other institutions. By considering the scope and limitations of reputational processes instantiated by these institutions, we can be most confident we are considering reputation from the perspective of the basic structure. Where such institutions employ systems of algorithmic reputation, they can be evaluated as reflections of the public rules that indicate the rights and obligations operating under the requirements of justice as fairness.

---

[397] Loren Larsen, 'HireVue Poised to Bring US Government Agencies' Recruiting Up To Speed' (*HireVue*, 16 May 2019) <https://www.hirevue.com/blog/hirevue-poised-to-bring-us-government-agencies-recruiting-up-to-speed> accessed 3 July 2020.

[398] Larsen and Taylor, 'United States Patent: 9305286' (n 262).

6.6.2.1 Recourse

The public rules of the basic structure cannot be said to be just unless they include opportunities for recourse. Systems of algorithmic reputation frequently deny the good of recourse by obscuring the reasoning behind an assessment or decision so that it is inaccessible or indecipherable to the subject. When we understand why decision *X* occurs rather than *Y*, we gain two important avenues of recourse: First, when we understand the reasoning behind a decision, we can plan future action in light of that reasoning. Understanding a decision enables planning for future decision states. Other than by guessing, a subject who desires future decisions to favor them can only adjust their actions or otherwise produce signals to improve their chances if they understand the standards that obtain.[399] For example, understanding that you were not offered a job because you lack a particular certification could provide an incentive to do the work of attaining that certification. Similarly, knowing in advance that the certification is preferred for a desired position is useful information for planning. Lacking information about the reasons behind decisions provides no recourse for taking action to remedy or plan for a given situation.

Second, we can hold the decision-maker accountable for their decision. A decision-maker who uses unfair or deceptive standards is difficult to hold to account if the standards are hidden or obscure. While decisions made by human decision-makers can be obscure and coercive, algorithmic systems are generally less accessible for accountability measures than other reputation systems. An employer who rejects every job applicant with a Persian-sounding name can be questioned about their potential anti-Persian bias. They may lie or dissemble but the opportunity to question, to gather evidence is more accessible than with a computer program that is functionally or lawfully blocked from questioning.

---

[399] Venkatasubramanian and Alfano (n 358).

The job candidate screening platform provided by HireVue fails to provide adequate recourse. Recall from Chapter 5 that among the questions raised about the HireVue system is that it does not provide feedback to people subjected to it. Job candidates may be provided with their profile scores if the employer using the system chooses to reveal them, but this is not a guaranteed outcome. The system is inscrutable even under such conditions. The scores provided by the system do not indicate exactly which gestures, facial expressions, word choices, or tones of voice are particularly favored or disfavored thereby preventing job candidates from adjusting their presentation to suit the requirements of the system. The validity of the results is also in question. While the producers of the HireVue system claim that the system has been rigorously evaluated for accuracy and fairness, we have only their claims to rely on. They have not offered evidence that can be assessed by disinterested investigators. The system fails to provide recourse in the form of either understanding or accountability.

Recourse is not guaranteed in non-algorithmic contexts, yet the prevailing practices and affordances of algorithmic reputation systems compound preexisting conditions. First, algorithmic systems are difficult to understand. Some systems make decisions that even their authors cannot explain.[400] This inscrutability renders algorithmic decisions especially unaccountable. While humans can also be inscrutable, we have the option of asking them questions. Even if human decision-makers resist questioning or can otherwise hold themselves unaccountable, this does not justify accepting the inscrutability of algorithmic systems.

6.6.2.2 Algocracy

Unaccountability is morally suspect whether from humans or from algorithmic systems.

[400] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31 Harvard Journal of Law & Technology 841.

Indeed, as argued by philosopher John Danaher, opaqueness in decision-making is the hallmark of "epistocracy," in which elites make decisions for others without engaging in an open process that confers the process as legitimate in the eyes of those affected.[401] Danaher suggests that "algocracy," or processes of opaque yet influential decision-making by algorithmic systems, is similarly an expression of elite decision-making with legitimacy problems.[402] The denial of rational autonomy for decision subjects obtains under both epistocracy and algocracy and may be worsened in the algorithmic context due to the additional opacity of such contexts. Second, algorithmic systems are typically shielded from analysis by intellectual property and market-protection concerns. The producers of algorithmic reputation systems have a financial incentive to shield their methods from public scrutiny to prevent their inventions from enriching others at their expense. Non-algorithmic decision-makers are less likely to have such concerns about explaining their reasoning and have fewer options to shield themselves in any case. These two factors place additional obstacles in the way of recourse for the subjects of algorithmic reputation systems over non-algorithmic systems. We are left to trust the word of the providers of such systems that their decision systems are both accurate and fair. This is hardly sufficient to meet even a minimal standard of accountability and therefore denies the good of recourse.

For reputation to function as an element of the expectations of the basic structure of society, recourse is a legitimate expectation. The exercise of the two moral powers is severely limited in the absence of recourse. First, without recourse, we cannot exercise rationality in the making of plans. The uncertainty of what is expected and what is discouraged renders all planning speculative at best. Second, without accountability, we are denied the reciprocity required by reasonableness.

---

[401] John Danaher, 'The Threat of Algocracy: Reality, Resistance and Accommodation' (2016) 29 Philosophy & Technology 245.
[402] ibid 252.

We expect others to be held to standards as they hold us to theirs. In the absence of accountability, persons are unlikely to find it reasonable to shape their rational plans such that they contribute to a cooperative, mutually advantageous society.

6.6.3 Algorithmic Reputation and Legitimacy

In the second stage of the society envisioned by Rawls, we move beyond the original position to the requirements of stability for a society conceived as fair system of cooperation for mutual advantage to endure from one generation to the next. Here, fairness is expressed as the sense that the basic structure of society and the decision-making frameworks included within it are perceived as legitimate by those who are subjected to it. Elemental to legitimacy are limitations to the control that individual comprehensive doctrines have in the political conception of society.

6.6.3.1 Comprehensive Doctrines

At least some set of systems of algorithmic reputation appear not to be legitimized in the overlapping consensus of reasonable comprehensive doctrines. These systems arguably fit within the scheme of some comprehensive doctrines but not others. For example, some comprehensive doctrines hold that *efficiency,* particularly in business matters, is more valuable than most other goods, such as many commonly held beliefs about privacy. An example of adherence to this view is that of Richard Posner who argues that firms have a maximal right to uncover information about others with whom they do business.[403] In contrast to arguments that promote personal autonomy in decisions about the release of personal information, Posner argues that the only justification for personal privacy is the "concealment of personal facts,"[404] which, on Posner's view, is morally indefensible. Concealment potentially limits access to discrediting information that could aid

[403] Posner (n 176).
[404] ibid 11.

transactants in making the most efficient and productive moves.[405] This applies to the details that contribute to a person's reputation, to which Posner has demonstrated particular concern. The role of reputation is to put "resources to their most valuable employments,"[406] and thereby lower the cost of searching out information necessary for transactions, both business and personal. For Posner, the most important features of privacy are protected by the natural limit of market efficiency; once the cost of inquiry into a person grows too high to justify its expense, the search will cease. What remains beyond this cost threshold is a domain of privacy presumably sufficient for the dignity of the person. While Posner is skeptical about the privacy claims of a firm's targets, he is bullish on the privacy requirements of the firm. Intellectual property rights and trade secrets are an essential ingredient for the firm's efficiency and prosperity. So, the firm should not be required to disclose its methods but should be granted latitude in "appropriating the social benefit of superior knowledge" by discovering what others would conceal.[407]

Based on this view of the firm's right to information to the limits of market efficiency, Posner would likely endorse the use of the HireVue platform for selecting employees. So long as the product maximizes the firm's efficiency by selecting "successful" employees at a reasonable cost, the functional extent of its reach is acceptable. Posner's market efficiency boundary for privacy suggests that he would find that a system capable of revealing a person's psychological states respects the legitimate domain of privacy, where the legitimate domain is understood as whatever is too expensive to enter. Posner would likely also hold that employers and their agents (*i.e.* HireVue) cannot be required to disclose their methods because this might dilute the benefits of superior knowledge from which to defend their competitive advantage. An interesting feature of

[405] ibid.
[406] ibid 31.
[407] ibid 10.

Posner's economics argument is that the domain of privacy required for human dignity is whatever is too costly to enter. If it works as claimed, a technology like that of HireVue appears to dramatically lower the cost to discover the most private information a person can possess, which is knowledge of their own state of mind. Following Posner, the claims of what information technologies can reveal has us headed for a privacy remainder of approximately zero.

Posner's view is not only a position. It reflects a comprehensive doctrine in which the efficiency, particularly economic efficiency, is the preeminent value through which the world should be viewed. On this view, individual privacy has little value because it creates inefficiencies that limit the maximal productivity of firms. While this comprehensive doctrine may attract other adherents, we can ask if it is *reasonable* from a Rawlsian perspective; we might ask if it contains enough content that overlaps with the comprehensive doctrines of others. Posner's doctrine appears to privilege the objectives and priorities of one set of citizens while setting aside those of others to a degree they would not accept.

6.6.3.2 Algorithmic Reputation and Public Reason

While the overlapping consensus of reasonable comprehensive doctrines provides legitimacy for the principles of a well-ordered society, the details of enacting those principles requires deliberation. In the vision of a democratic society committed to egalitarian ideals offered by Rawls in his later works, important decisions that affect the coercive governance of people are generally expected to be subject to open and transparent deliberative processes.

Where systems of algorithmic reputation distribute access to goods that are controlled by the basic structure (*e.g.* a government job), the openness and transparency requirements of the basic structure make them legitimate targets for accountability. For example, an admissions officer at a public university would face scrutiny if she was to use skull measurements in selecting students.

There may be admissions officers who sincerely believe that skull size corresponds with educational success, but they know (or we hope they know) that they would be held accountable for making admissions decisions based on skull-measuring; they would be required to defend this standard. This is not to say that such choices are never taken. It is not even to say that all such choices are indefensible; it is to say that we would be right in holding the decision-maker accountable for the choice.

There is no reason why a system of algorithmic reputation should not be subject to a similar degree of scrutiny, particularly in matters regulated by the basic structure. The fact that a decision is made using a collection of formal programming steps or is the product of a deep learning algorithm does not remove the accountability we should expect. However, this is challenging to accomplish. Algorithmic systems, be they formal systems of stepwise instructions or "learning" systems such as neural networks, are understood by only a small minority of human beings. Even within this narrow segment, some algorithmic systems produce decisions that even software engineers cannot explain. Added to this is the secrecy that businesses employ to protect their code as intellectual property. Technology firms are reluctant to divulge the details of how their systems work for economic reasons. Taken together, these aspects of algorithmic systems present significant challenges for subjects or other arbiters to understand them. However, the barriers to this understanding should not be confused with objectivity or an otherwise special classification stemming from its basis in mathematics and so on. We have reason to be concerned with how such systems make decisions. For automated decisions systems to "learn," they reference the results of prior decisions and preexisting data; they learn from the world as it is and how they find it. An algorithm that leans from prior hiring decisions in a firm that historically hired persons of one gender, race, and/or ethnicity will be challenged to appreciate the qualities of people with other

identities.[408] Such a system should be no less accountable than humans making similar choices.[409] Public reason provides an expectation that citizens demonstrate the respect of explaining themselves fully when exercising their coercive power over one another.

From the case study of HireVue, we know that there are significant obstacles for HireVue to meet the ideal of public reason, including existing intellectual property laws and epistemic barriers. First, the inner workings of the HireVue platform are protected by trade secret laws and the company has a strong incentive to maintain a degree of secrecy to protect its business position. However, it is possible to maintain a trade secret while submitting technology to a third-party audit, as suggested by Ifeoma Ajunwa.[410] Yet, as discussed in Chapter 5, the HireVue company has not made its system available to inspection and provides no verifiable evidence about its accuracy and fairness. HireVue thereby fails to meet the requirements of openness and transparency. Second, HireVue relies on particular epistemic commitments about the power of technology to reveal and predict the psychological states of human subjects. The company projects confidence that an algorithmic analysis of audio, video, and interaction data reliably reveals who people *really* are. The science of psychometrics is not settled, nor is the science of similarly interpreting behavioral cues using software. The depth of doubt in this science is not trivial; adherents of these claims have been compared to practitioners of phrenology—a science that has not only been discredited but has been accused of being based in discriminatory ideologies. Whether the phrenology accusation is fair is beside the point; the volume of critique calls this epistemic commitment into question.

The HireVue system appears to import the comprehensive doctrine of its producers. This is

[408] Ajunwa, 'The Paradox of Automation as Anti-Bias Intervention' (n. 306).
[409] This is so whether we agree or disagree, upon reflection, with the choice.
[410] Ifeoma Ajunwa, 'Automated Employment Discrimination' (n 49).

likely the case in many systems of algorithmic reputation. If this is so, then following Rawls, we cannot accept a society ruled under the conceptions held by any particular comprehensive doctrine. In the "free standing" political conception of society proposed by Rawls, members share a commitment to common principles. This includes a commitment to justify coercive power in open and transparent deliberation. While operating from the protections of intellectual property and without providing acceptable evidence for their claims, systems of algorithmic reputation are not likely to survive public reason. We are therefore unlikely to endorse such systems in our reasoned moral reflections about the requirements of a society conceived as a fair system of cooperation for mutual advantage.

6.7 CHAPTER CONCLUSION

In this chapter I have argued that systems of algorithmic reputation, such as systems that profile people and either make or support decisions about their lives, can be subjected to a moral analysis using the political philosophy of John Rawls. I suggest that algorithmic profiling of consumers and citizens and the automated decisions that flow from such profiles, are properly understood as systems of reputation. As such, they are targets of normative inquiry rather than merely neutral technical projects.

The cornerstone of Rawlsian theory is the model conception of the "two moral powers" expected to be practiced by citizens. Rawlsian ethics is person-centered. Society is constructed by persons in light of their rational eudaimonism in which they recognize their own good in combination with their reasonableness in recognizing how their good is wrapped in the good of others. This is the basis of a society conceived as a fair system of cooperation for mutual advantage.

I find that systems of algorithmic reputation potentially diminish the good of self-respect by failing to recognize persons as human beings and also by misclassifying them in ways that can

cause them tremendous harm. When we fail to accord sufficient respect to persons, we deny their rational autonomy from which they act as full citizens who are aware of their own good and committed to the good of society. As such, we cannot be confident in fairly assessing persons who have been so denied. I further argue that the public rules of the basic structure, which set out the rights and obligations of citizens, create the expectations under which a person's public reputation is assessed. A basic structure that is just should produce reputational assessments that are fair. If we doubt the justice of the basic structure, we may also doubt its reputational standards. While there are realms of life where reputational standards are governed by non-public rules, such as our individual comprehensive doctrines and our affections, this realm is limited; it is bounded by the requirements of the basic structure. While we may be uncertain about the requirements for a system of algorithmic reputation that operates in the non-public realm, we can be confident that reputational standards in public settings, such as government offices and public universities, are subject to public rules and require accountability, including a reasonable degree of legibility and opportunities for recourse.

Finally, I argued that the exercise of coercive power must meet standards of legitimacy if it is to endure as a system of liberal equality. Legitimacy flows from the members' sense that society not only serves their interests but demonstrates respect for them through deliberations that are open and transparent. Employing a broad view of public reason, I suggest that systems of reputation that do not make themselves available for open and transparent deliberation while engaged in the work of the basic structure do not meet the requirements of legitimacy, placing the stability of society at risk. Systems of algorithmic reputation, like the HireVue employment screening system, are difficult to hold to account both because such systems are challenging to understand, and also because their producers choose not to make them available for a more complete understanding.

While it may make practical or business sense for firms to maintain secrecy about their systems, the coercive power of such systems exerts pressure on the stability of a society conceived as a fair system of cooperation for mutual advantage. Public reason demands a baseline of openness as a measure of the respect shown by persons one to another, including firms. To the extent that we expect openness and transparency on the subject of reputation systems, we can demand it to a similar degree from assessors.

BIBLIOGRAPHY

Ajunwa I, 'Automated Employment Discrimination' [2019] SSRN Electronic Journal <https://www.ssrn.com/abstract=3437631> accessed 13 March 2020

——, 'The Future of Work: Protecting Workers' Civil Rights in the Digital Age' (witness statement), House of Representatives Joint Hearing of Subcommittee on Civil Rights and Human Services, 116th Congress. Washington, D.C., United States, February 5, 2020 <https://edlabor.house.gov/imo/media/doc/AjunwaTestimony02052020.pdf> accessed 28 December 2020.

——, 'The Paradox of Automation as Anti-Bias Intervention' [2020] Cardoza Law Review 55.

Anderson C and Shirako A, 'Are Individuals' Reputations Related to Their History of Behavior?' (2008) 94 Journal of Personality and Social Psychology 320.

Anderson ES, 'What Is the Point of Equality?' (1999) 109 Ethics 287.

Aristotle, *Nicomachean Ethics* (JEC Welldon tr, Hackett Pub Co 1999).

'Bias, AI Ethics, and the HireVue Approach' (*HireVue*) <https://www.hirevue.com/why-hirevue/ethical-ai> accessed 19 March 2020.

Bicchieri C, *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge University Press 2006).

Bogen M and Rieke A, 'Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias' (Upturn 2018) <https://www.upturn.org/reports/2018/hiring-algorithms> accessed 30 September 2019.

Borning A and others, 'Informing Public Deliberation: Value Sensitive Design of Indicators for a Large-Scale Urban Simulation' in Hans Gellersen and others (eds), *ECSCW 2005* (Springer-Verlag 2005).

Borning A, Friedman B and Kahn PH, 'Designing for Human Values in an Urban Simulation System: Value Sensitive Design and Participatory Design' (2004).

Brennan J, 'Against the Moral Powers Test of Basic Liberty' (2020) 28 European Journal of Philosophy 492.

Brey P, 'Disclosive Computer Ethics' (2000) 30 SIGCAS Comput. Soc. 10.

——, 'The Technological Construction of Social Power' (2008) 22 Social Epistemology 71.

Buolamwini J and Gebru T, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', *Proceedings of Machine Learning Research* (2018).

Carey JW, *Communication as Culture: Essays on Media and Society* (Unwin Hyman 1989).

Cerrato J and Freyermuth J, 'Market Guide for Talent Acquisition Applications' (*Gartner*, 18 December 2018) <https://www.gartner.com/document/3896176?ref=solrAll&refval=243416472 > accessed 11 March 2020.

Cohen GA, *Self-Ownership, Freedom, and Equality* (Cambridge University Press ; Maison des sciences de l'homme 1995).

Crawford K and Schultz J, 'AI Sytems as State Actors' (2020) 119 COLUMBIA LAW REVIEW 33.

Creswell JW, *Research Design : Qualitative, Quantitative, and Mixed Methods Approaches* (3rd ed., Los Angeles : Sage 2009).

Danaher J, 'The Threat of Algocracy: Reality, Resistance and Accommodation' (2016) 29 Philosophy & Technology 245.

Dastin J, 'Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women' *Reuters* (10 October 2018) <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G> accessed 20 September 2019.

Dewey C, 'Creepy Startup Will Help Landlords, Employers and Online Dates Strip-Mine Intimate Data from Your Facebook Page' *The Washington Post* (9 June 2016) <https://www.washingtonpost.com/news/the-intersect/wp/2016/06/09/creepy-startup-will-help-landlords-employers-and-online-dates-strip-mine-intimate-data-from-your-facebook-page/> accessed 11 June 2016.

Douglas R, '2020 Identity Theft Statistics' [2020] *Consumer Affairs* <https://www.consumeraffairs.com/finance/identity-theft-statistics.html> accessed 5 March 2020.

Durkheim É, *Émile Durkheim on The Division of Labor in Society* (George Simpson tr, Macmillan 1933).

Fertik M and Thompson DC, *The Reputation Economy: How to Optimize Your Digital Footprint in a World Where Your Reputation Is Your Most Valuable Asset* (First edition, Crown Business 2015).

FinTech Silicon Valley, *Video Interview with Brian Ley, CEO/Founder Alpharank* (2017) <https://www.youtube.com/watch?v=JJe9TlSM_8M> accessed 26 July 2018.

Fleming J, 'The Right to Reputation and the Preferential Option for the Poor' (2004) 24 Journal of the Society of Christian Ethics 73.

Fourcade M and Healy K, 'Classification Situations: Life-Chances in the Neoliberal Era' (2013) 38 Accounting, Organizations and Society 559.

Freeman SR, *Rawls* (Reprinted, Routledge 2010).

Friedman B and Hendry D, *Value Sensitive Design: Shaping Technology with Moral Imagination* (MIT Press 2019).

Friedman B and Hendry DG, 'The Envisioning Cards: A Toolkit for Catalyzing Humanistic and Technical Imaginations' (ACM Press 2012) <http://dl.acm.org/citation.cfm?doid=2207676.2208562> accessed 24 January 2015.

Friedman B, Kahn PH and Borning A, 'Value Sensitive Design and Information Systems' in Ping Zhang and Dennis F Galletta (eds), *Human-computer interaction and management information systems: foundations* (ME Sharpe 2006).

Gillespie T, *Wired Shut: Copyright and the Shape of Digital Culture* (MIT Press 2007).

Goffman E, *The Presentation of Self in Everyday Life* (Repr, Penguin 1990).

Goodman W, 'Lie Detectors Don't Lie' *New York Times* (New York NY USA, 1965) SM12.

Greif A, 'Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders' (1989) 49 The Journal of Economic History 857.

——, 'Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition' (1993) 83 The American Economic Review 525.

Gunkel DJ, *The Machine Question: Critical Perspectives on AI, Robots, and Ethics* (The MIT Press 2012).

Haggerty KD and Ericson RV, 'The Surveillant Assemblage' (2000) 51 The British Journal of Sociology 605.

Harris CI, 'Whiteness as Property' (1993) 106 Harvard Law Review 1707.

Harwell D, 'A Face-Scanning Algorithm Increasingly Decides Whether You Deserve the Job' *Washington Post* (6 November 2019) <https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/> accessed 30 March 2020.

Hickman L, 'The Avery Review | "On the Basis of Safety": The Forced Intimacies of Accessible Air Travel' (2020) 6 *The Avery Review* <http://www.averyreview.com/issues/48/on-the-basis-of-safety> accessed 29 June 2020

Hill E and others, 'How George Floyd Was Killed in Police Custody' *The New York Times* (31 May 2020) <https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html> accessed 30 June 2020.

Hill K, 'Wrongfully Accused by an Algorithm' *The New York Times* (24 June 2020) <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html> accessed 3 July 2020.

'HireVue Rolls Out Data-Driven Candidate and Interviewer Recommendation Engine' [2014] *Professional Services Close-Up* <https://advance-lexis-com.offcampus.lib.washington.edu/document/?pdmfid=1516831&crid=7a5bcf19-8dfe-420a-97c3-0a744fb26478> accessed 16 March 2020.

Hoffmann A, 'Google Books as Infrastructure of in/Justice: Towards a Sociotechnical Account of Rawlsian Justice, Information, and Technology' (University of Wisconsin 2014).

Hoffmann AL, 'Data Violence and How Bad Engineering Choices Can Damage Society' (*Medium*, 30 April 2018) <https://medium.com/s/story/data-violence-and-how-bad-engineering-choices-can-damage-society-39e44150e1d4> accessed 24 January 2019.

——, 'Making Data Valuable: Political, Economic, and Conceptual Bases of Big Data' (2018) 31 Philosophy & Technology 209.

——, 'Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse' (2019) 22 Information, Communication & Society 900.

Howe DW, 'American Victorianism as a Culture' (1975) 27 American Quarterly 507.

Johnson JA, 'From Open Data to Information Justice' (2014) 16 Ethics and Information Technology 263.

Johnston C, 'Data Brokers Won't Even Tell the Government How It Uses, Sells Your Data' (*Ars Technica*, 21 December 2013) <http://arstechnica.com/business/2013/12/data-brokers-wont-even-tell-the-government-how-it-uses-sells-your-data/> accessed 21 July 2016.

Kant I, 'Groundwork for the Metaphysics of Morals', *The philosophy of Kant: Immanuel Kant's moral and political writings* (Random House 1949).

Katell M, 'Adverse Detection: The Promise and Peril of Body-Worn Cameras' in Bryce Clayton Newell, Tjerk Timan and Bert-Jaap Koops (eds), *Surveillance, Privacy, and Public Space* (Taylor & Francis 2018).

Kekes J, 'A Question for Egalitarians' (1997) 107 Ethics 13.

Keyes O, 'The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition' (2018) 2 Proceedings of the ACM on Human-Computer Interaction 1.

Kitchin R, 'Thinking Critically about and Researching Algorithms' (2017) 20 Information, Communication & Society 14.

Koenecke A and others, 'Racial Disparities in Automated Speech Recognition' [2020] Proceedings of the National Academy of Sciences.

Koops B-J, ‘Law, Technology, and Shifting Power Relations’ (2010) 25 Berkeley Technology Law Journal 973.

Korsgaard CM, *The Sources of Normativity* (Cambridge University Press 1996).

Kranzberg M, ‘Technology and History: “Kranzberg’s Laws”’ (1986) 27 Technology and Culture 544.

Lange A, ‘She Trusted Her Husband To Handle Her Money. It Cost Her More Than She Imagined.’ [2019] *BuzzFeed News* <https://www.buzzfeednews.com/article/arianelange/coerced-debt-financial-abuse-fix-credit-score> accessed 5 March 2020.

Larsen L, ‘HireVue Poised to Bring US Government Agencies’ Recruiting Up To Speed’ (*HireVue*, 16 May 2019) <https://www.hirevue.com/blog/hirevue-poised-to-bring-us-government-agencies-recruiting-up-to-speed> accessed 3 July 2020.

Larsen L and Taylor B, ‘United States Patent: 8751231 – Model-Driven Candidate Sorting Based on Audio Cues.’

——, ‘United States Patent: 8856000 – Model-Driven Candidate Sorting Based on Audio Cues.’

——, ‘United States Patent: 9009045 – Model-Driven Candidate Sorting.’

——, ‘United States Patent: 9305286 – Model-Driven Candidate Sorting.’

——, ‘United States Patent: 10438135 – Performance Model Adverse Impact Correction.,’

Latour B, ‘Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts’ in Bijker, Wiebe E and John Law (eds), *Shaping Technology/Building Society: Studies in Sociotechnical Change* (MIT Press 1992).

Locke J, *Second Treatise of Government: An Essay Concerning the True Original, Extent and End of Civil Government* (John Wiley & Sons 2014).

Lyotard J-F, *The Postmodern Condition: A Report on Knowledge* (Geoff Bennington and Brian Massumi trs, University of Minnesota Press 1984).

McKeever K and Ditcheva B, ‘The Circular Letter of Credit’ (2006) <http://library.law.columbia.edu/CircularLetterOfCredit/> accessed 20 December 2019.

Metz R, ‘There’s a New Obstacle to Landing a Job after College: Getting Approved by AI’ *CNN* (15 January 2020) <https://www.cnn.com/2020/01/15/tech/ai-job-interview/index.html> accessed 15 January 2020.

Miller J and others, ‘Value tensions in design’, *Proceedings of the 2007 International ACM Conference / Supporting Group Work (GROUP ’07)* (2007).

Mills CW, ‘Rawls on Race/Race in Rawls’ (2009) 47 The Southern Journal of Philosophy 161.

——, ‘Decolonizing Western Political Philosophy’ (2015) 37 New Political Science 1.

Mondragon DN, Aichholzer C and Leutner DK, ‘The Next Generation of Assessments’.

Moore AD, *Privacy Rights: Moral and Legal Foundations* (Pennsylvania State University Press 2010).

Moore AD, ‘Privacy, Security, and Government Surveillance: Wikileaks and the New Accountability’ (2011) 25 Public Affairs Quarterly 141.

Moore AD and Martin S, ‘Privacy, Transparency, and the Prisoner’s Dilemma’ [2018] SSRN Electronic Journal <https://www.ssrn.com/abstract=3212217> accessed 13 April 2020.

Morris RJ, ‘Anatomy of a Patent’ in Avery Goldstein (ed), *Patent law for scientists and engineers* (Taylor & Francis 2005).

Nadler A and McGuigan L, ‘An Impulse to Exploit: The Behavioral Turn in Data-Driven Marketing’ (2018) 35 Critical Studies in Media Communication 151.

Nissenbaum H, ‘Accountability in a Computerized Society’ (1996) 2 Science and Engineering Ethics 25.

Noble SU, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press 2018).

Nozick R, *Anarchy, State, and Utopia* (Basic Books 1974).

Nussbaum MC, *Frontiers of Justice: Disability, Nationality, Species Membership* (First Harvard University Press paperback edition, The Belknap Press of Harvard University Press 2007).

Ochigame R, ‘The Long History of Algorithmic Fairness’ [2020] *Phenomenal World* <https://phenomenalworld.org/analysis/long-history-algorithmic-fairness> accessed 5 March 2020.

O’Connor C, ‘Report: Walmart Workers Cost Taxpayers $6.2 Billion In Public Assistance’ [2014] *Forbes* <https://www.forbes.com/sites/clareoconnor/2014/04/15/report-walmart-workers-cost-taxpayers-6-2-billion-in-public-assistance/> accessed 3 July 2020.

Ossorio P and Duster T, ‘Race and Genetics: Controversies in Biomedical, Behavioral, and Forensic Sciences’ (2005) 60 American Psychologist 115.

Ostrom E, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press 1990).

Pasquale F, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015).

Paton HJ, *Groundwork of the Metaphysics of Morals* (Harper & Rowe 1964).

Patrick K, ‘Arnold Foundation to Roll Out Pretrial Risk Assessment Tool Nationwide’ (*InsideSources*, 4 September 2018) <https://www.insidesources.com/arnold-foundation-to-roll-out-pretrial-risk-assessment-tool-nationwide/> accessed 7 April 2019.

Pickard AJ, *Research Methods in Information* (Facet 2007).

‘PitchBook Profile - HireVue’ (*Pitchbook*, 3 February 2020) <https://my.pitchbook.com/profile/47555-65/company/profile#insights> accessed 17 March 2020.

Poon M, ‘Scorecards as Devices for Consumer Credit: The Case of Fair, Isaac & Company Incorporated: Scorecards as Devices for Consumer Credit: The Case of Fair, Isaac & Company Incorporated’ (2007) 55 The Sociological Review 284.

Posner RA, ‘Privacy, Secrecy, and Reputation 1978 James McCormick Mitchell Lecture, The’ (1978) 28 Buffalo Law Review 1.

‘Pre-Employment Testing Software | Online Gamified Assessments | HireVue’ <https://www.hirevue.com/products/assessments> accessed 30 March 2020.

Quong J, ‘The Scope of Public Reason’ (2004) 52 Political Studies 233.

Raghavan M and others, ‘Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices’ 469 <https://doi.org/10.1145/3351095.3372828>.

Rawls J, ‘Kantian Constructivism in Moral Theory’ (1980) 77 The Journal of Philosophy 515.

——, *A Theory of Justice, Revised Edition* (Rev ed, Belknap Press of Harvard University Press 1999).

——, *Justice as Fairness: A Restatement* (Erin Kelly ed, Harvard University Press 2001).

——, *A Theory of Justice, Original Edition* (Original ed, Belknap Press 2005).

——, *Political Liberalism* (Expanded ed, Columbia University Press 2005).

Rice L and Swesnik D, ‘Discriminatory Effects of Credit Scoring on Communities of Color Symposium: Credit Scoring and Credit Reporting’ (2013) 46 Suffolk University Law Review 935.

Rosenberg A, *Philosophy of Social Science* (4th ed, Westview Press 2012).

Sabel J-M, ‘How Artificial Intelligence Helps Humans Find the Best Talent [Infographic]’ (*HireVue*, 5 October 2017) <https://www.hirevue.com/blog/how-artificial-intelligence-helps-humans-find-the-best-talent> accessed 5 December 2018.

Sánchez-Monedero J, Dencik L and Edwards L, ‘What Does It Mean to “solve” the Problem of Discrimination in Hiring?: Social, Technical and Legal Perspectives from the UK on Automated Hiring Systems’, *Proceedings of the 2020 Conference on Fairness, Accountability, and*

Transparency (ACM 2020) <http://dl.acm.org/doi/10.1145/3351095.3372849> accessed 29 January 2020.

Segrave K, *Lie Detectors: A Social History* (McFarland 2004).

Shankar S and others, 'No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World' [2017] arXiv:1711.08536 [stat] <http://arxiv.org/abs/1711.08536> accessed 24 July 2018

Shirky C, *Here Comes Everybody: The Power of Organizing without Organizations* (Penguin Press 2008).

Singer N, 'The Scoreboards Where You Can't See Your Score' *The New York Times* (27 December 2014) <http://www.nytimes.com/2014/12/28/technology/the-scoreboards-where-you-cant-see-your-score.html> accessed 29 December 2014.

Solove D, *The Digital Person: Technology and Privacy in the Information Age* (New York University Press 2004)

——, *The Future of Reputation: Gossip, Rumor, and Privacy on the Internet* (Yale University Press 2007).

Star SL, 'The Ethnography of Infrastructure' (1999) 43 American Behavioral Scientist 377.

Stark L, 'Algorithmic Psychometrics and the Scalable Subject' (2018) 48 Social Studies of Science 204.

Sterba J, 'Justice as Desert' (1974) 3 Social Justice Theory and Practice 17.

Susser D, Roessler B and Nissenbaum H, 'Technology, Autonomy, and Manipulation' (2019) 8 Internet Policy Review.

Taylor B and Larsen L, 'United States Patent: 9652745 – Model-Driven Evaluator Bias Detection.'

Timcke S, 'The One-Dimensionality of Econometric Data: The Frankfurt School and the Critique of Quantification' 15.

van den Hoven J, 'Moral Methodology and Information Technology' in Kenneth Einar Himma and Herman T Tavani (eds), *The Handbook of Information and Computer Ethics* (John Wiley & Sons, Inc 2008).

Veale M and Binns R, 'Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data' (2017) 4 Big Data & Society 2053951717743530.

Venkatasubramanian S and Alfano M, 'The Philosophical Basis of Algorithmic Recourse', *Proceedings of the 2020 conference on fairness, accountability, and transparency* (Association for Computing Machinery 2020) <https://doi.org/10.1145/3351095.3372876>.

Victor J and others, ‘SHRM Survey Findings: Background Checking-The Use of Credit Background Checks in Hiring Decisions’ (Society for Human Resource Management, 19 July 2012) <https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Pages/creditbackgroundchecks.aspx>.

‘Video Interview Software | On-Demand Interview Technology’ (*HireVue*) <https://www.hirevue.com/products/video-interviewing> accessed 26 March 2020.

Volpone SD and others, ‘Exploring the Use of Credit Scores in Selection Processes: Beware of Adverse Impact’ (2015) 30 Journal of Business and Psychology 357.

Wachter S, Mittelstadt B and Russell C, ‘Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR’ (2018) 31 Harvard Journal of Law & Technology 841.

Warde B, ‘Black Male Disproportionality in the Criminal Justice Systems of the USA, Canada, and England: A Comparative Analysis of Incarceration’ (2013) 17 Journal of African American Studies 461.

Wei Y and others, ‘Credit Scoring with Social Network Data’ (2015) 35 Marketing Science 234.

Winner L, ‘Do Artifacts Have Politics?’ (1980) 109 Daedalus 121.

Yin RK, *Applications of Case Study Research* (2nd ed, Sage Publications 2003).

Yoo D and others, ‘A Value Sensitive Action-Reflection Model: Evolving a Co-Design Space with Stakeholder and Designer Prompts’ (ACM 2013).

Young IM, *Justice and the Politics of Difference* (Princeton University Press 1990).

——, ‘Taking the Basic Structure Seriously’ (2006) 4 Perspectives on Politics.

Zalta EN and others, ‘Stanford Encyclopedia of Philosophy’ 29.

Zuboff S, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (First edition, PublicAffairs 2018).