HUMAN IN THE LOOP: THE ROLE OF INDIVIDUAL AND INSTITUTIONAL BEHAVIOR ON PREDICTIVE ALGORITHMS

By

William Isaac

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Political Science – Doctor of Philosophy

2018

ABSTRACT

HUMAN IN THE LOOP: THE ROLE OF INDIVIDUAL AND INSTITUTIONAL BEHAVIOR ON PREDICTIVE ALGORITHMS

By

William Isaac

Over the past decade, algorithmic decision systems (ADS) – applications of statistical or computational techniques designed to assist human-decision making processes – have moved from an obscure domain of statistics and computer science into the mainstream. The rapid decline in the cost of computer processing and ubiquity of digital data storage has created a dramatic rise in the adoption of ADS using applied machine learning algorithms, transforming various sectors of society from digital advertising to political campaigns, risk modeling for the banking sector, healthcare and beyond. Many agencies and practitioners in the public sector turn to ADS as a means to stretch limited public resources amidst growing public demands for equity and accountability.

However, recent research from multiple fields has found that social and institutional biases, often reflected by input data used to generate predictions. The potential of perpetuated discrimination via input data is a particular concern in fields such as criminal justice where historical biases against minorities have the potential to exacerbate existing racial inequalities. In a series of three essays, this dissertation seeks to outline how institutional norms often shape algorithmic predictions, examine how ADSs alter the incentive structures for agents using the tools, and ultimately its impact on human decision-making.

To Bobbi and Charlotte, who gave me the strength to finish the race.

ACKNOWLEDGEMENTS

While I had plenty of warning beforehand, it is shocking how many people support you while writing a dissertation. This experience has shown me that I have been blessed a with truly amazing support system and some of the most thoughtful and caring people in the world. Starting with dissertation committee (Eric Juenke, Matt Grossmann, Josh Sapotichne, and Cory Smidt) who have mentored me and given the space to develop into the scholar I always wanted to be. Even during the low moments of my graduate school tenure they always believe in my potential and pushed me into positions to succeed as a person and academic. I truly wish every doctoral student could have a committee as supportive as mine.

I also want to express my deepest thanks to my colleagues at the Human Right Data Analysis Group (HRDAG) including, Patrick Ball, Kristian Lum, and Megan Price. The opportunity to work at HRDAG changed my life. They taught me that it is possible to be rigorous in your research while still speaking truth to power. Regardless of where my adventures in life take me next, I will always be proudest of the work I did during my tenure there. I sincerely hope this institution can remain a beacon for data geeks to pursue statistics in the justice and progress.

I would like to thank my family. The countless nights spent in front of a computer writing this document or traveling around the world to communicate its findings would not have been possible without my entire family (Reta Adams, Jesse Adams, Joyce White, Jesse Adams Jr., Elvis Isaac, Jennie Isaac and Geraldine Isaac) and countless others filling in the gaps and chipping in. Though they may not always understand what my research is about, they have always been extremely proud of me using my platform for change.

Last but certainly not least, I would like to thank my wife Bobbi. Before the success and achievements, you took a gamble on a kid cleaning cars inside of a dusty auto garage in Tuscaloosa, Alabama. Not many people would have believed my outlandish dream of becoming an academic or getting a PhD, but you stood by me the entire way and never wavered. For that I am eternally grateful. And Charlotte, you were perhaps the most important inspiration during this process. I

work everyday with the goal of creating a better world for you. I truly hope you can one day live in a world where people can judge you by the content of your character and not the color your skin. While the current times may give us cause to question this possibility, I firmly believe we will achieve this promise one day. As Dr. Martin Luther King once said, "the arc of the moral universe is long, but it bends toward justice."

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER 1 HOPE, HYPE, AND FEAR: THE PROMISE AND POTENTIAL PITFALLS OF THE BIG DATA ERA IN CRIMINAL JUSTICE 1
- 1.1 Background 2
- 1.2 Case Study: Predictive Policing 5
- 1.3 Discussion 13

CHAPTER 2 THE NUMBERS GAME: GOODHART'S LAW AND THE USE OF THE INDEX CRIMES AS A PERFORMANCE METRICS 15
- 2.1 Performance Metrics and Goodhart Effects 16
  - 2.1.1 Defining Goodhart Effects 16
    - 2.1.1.1 Goodhart's Law 16
    - 2.1.1.2 Campbell's Law 17
- 2.2 Goodhart Effects and Policing 19
  - 2.2.1 Methods & Data 25
    - 2.2.1.1 Data 28
  - 2.2.2 Results 29
    - 2.2.2.1 Feedback Effects 33
- 2.3 Conclusion 35

CHAPTER 3 BACK TO THE FUTURE: PREDICTIVE POLICING AND RESIDUAL DISCRIMINATION FROM RACE-NEUTRAL CLASSIFIERS . 37
- 3.1 This is Your Brain on Data 39
  - 3.1.1 Principles of Modern Technology and Human Behavior 41
  - 3.1.2 Technology-Behavior Interaction in Policing 47
- 3.2 Case Study: LAPD Deployment of Predpol 50
- 3.3 Conclusion 69

BIBLIOGRAPHY 72

LIST OF TABLES

Table 2.1: Segmented Regression Table of MAD Values on Predictive Policing Deployment 33

LIST OF FIGURES

Figure 1.1: Estimated number of drug users, based on 2011 National Survey on Drug Use and Health 8

Figure 1.2: Number of drug arrests made by Oakland police department, 2010 9

Figure 1.3: Number of days with targeted policing in areas flagged by PredPol analysis of Oakland police data 10

Figure 1.4: Number of days with targeted policing in areas flagged by PredPol analysis of Oakland police data 11

Figure 1.5: Predicted odds of crime in locations targeted by PredPol algorithm, relative to non-targeted locations. 12

Figure 2.1: Cumulative Digit Distribution for Reported Property Crimes, Foothill District . 28

Figure 2.2: Monthly Digit Distribution for Reported Property Crimes, Foothill District . . . 29

Figure 2.3: Cumulative Digit Distribution for Reported Property Crimes, Southwest District 30

Figure 2.4: Annual Digit Distribution for Reported Property Crimes, Southwest District . . 31

Figure 2.5: Cumulative Digit Distribution for Reported Property Crimes, North Hollywood District 31

Figure 2.6: Monthly Digit Distribution for Reported Property Crimes, North Hollywood District 32

Figure 2.7: Monthly Time Series of MAD Scores, North Hollywood District 34

Figure 3.1: Percentage of Arrests by Crime Type and Data Source, Foothill District 2010-2017 59

Figure 3.2: Percentage of Arrests by Crime Type and Data Source, Southwest District 2010-2017 60

Figure 3.3: Percentage of Arrests by Crime Type and Data Source, North Hollywood District 2010-2017 61

Figure 3.4: Demographic Composition of LAPD Districts 63

Figure 3.5: Arrests Rates in Control and Treatment Boxes for Predpol Deployed LAPD Divisions 64

Figure 3.6: Arrests Rates Non-Predicted Boxes in Predpol Deployed LAPD Divisions . . . 65

Figure 3.7: Pre-Post Monthly Arrest Rates by Race and Quadrant, Foothill District 67

Figure 3.8: Pre-Post Monthly Arrest Rates by Race and Quadrant, Southwest District 68

Figure 3.9: Pre-Post Monthly Arrest Rates by Race and Quadrant, North Hollywood District 69

CHAPTER 1

HOPE, HYPE, AND FEAR: THE PROMISE AND POTENTIAL PITFALLS OF THE BIG DATA ERA IN CRIMINAL JUSTICE

> *Crime and disorder are not natural phenomena. These events have to be observed, noticed, acted upon, collected, categorized, and recorded — while other events aren't.*
>
> — Elizabeth Joh*, Feeding the Machine: Policing, Crime Data, & Algorithms*

Over the past decade, algorithmic decision systems (ADS) – applications of statistical or computational techniques designed to assist human-decision making processes – have moved from an obscure domain of statistics and computer science into the mainstream (Munoz et al. 2016, O'Neil 2016). The rapid decline in the cost of computer processing and ubiquity of digital data storage has created a dramatic rise in the adoption of ADS using applied machine learning algorithms, transforming various sectors of society from digital advertising to political campaigns (Hersh 2015, Nickerson & Rogers 2014), risk modeling for the banking sector (Wei et al. 2016, pg. 234), healthcare (Powles & Hodson 2017, pg. 351) and beyond. Many agencies and practitioners in the public sector turn to ADS as a means to stretch limited public resources amidst growing public demands for equity and accountability. Advocates of these "intelligence-led" or "evidence-based" policy approaches assume these algorithmic tools will allow government agencies to use *objective* data to overcome historical inequalities to serve underrepresented groups (Podesta et al. 2014, Chowdhry et al. 2016, Miller 2018) better.

However, the assumption of objective data is flawed. All human behavior or social phenomenon that machine learning algorithms attempt to predict come from a data-generation process (DGP) which comprises trillions of complex interactions between the roughly 7 billion people that inhabit our planet. The DGP is often unseen to the analyst, but we make assumptions regarding this process by choice of statistical models or the inferences we derive from our analysis. Machine learning algorithms are often theorized and developed in cases of simulated data or data with outcome variables with little ambiguity in interpretation or method of collection. However, if we assume incorrectly about the DGP, the predictions and conclusions we generate will be highly inaccurate (Cederman & Weidmann 2017, pg. 475). Furthermore, because the *true* DGP is unseen, it is nearly impossible to determine whether a proposed measure captures the phenomenon or outcome of interest for decision-makers.

1.1 Background

Defining what is considered objective data is a particularly acute problem in criminal justice. Dating back to the turn of the 20th century, statisticians and criminologists have raised concerns over the operationalization and measurement of crime (Morrison 1897). At its core, crime is a social phenomenon that has had multiple definitions and interpretations across time. Since the early 1930s, the United States Department of Justice uniform crime reporting (UCR) data – considered the official assignment of national crime trends – are based on crimes known to the police, either reported by the public or witnessed by members of law enforcement (Beattie 1941, Mosher et al. 2010, Wormeli 2018). While this operationalization of crime is reliable in the statistical sense (i.e. it will consistently measure the same concept over time), multiple scholars have pointed out this approach systematically undercounts crimes committed by some groups. These "hidden” or unreported instances are often referred to as the ”dark figure of crime" (Mosher et al. 2010, pg. 45).

One of the most common reasons for the emergence of "dark figures” has been the policies and practices of individual police departments. Robison (1936), who was among a group of early scholars to originally make a linkage between systemic bias in the criminal justice system and measurement, found that arrest data for delinquency of juveniles in New York state (defined as truancy, theft, or malicious mischief) was not a function of a persons' race or socioeconomic status directly – as previously theorized – but rather the differential treatment of these individuals by the criminal justice system. Beattie (1941, pg. 21) also noted that in addition to demographic factors, police statistics were likely manipulated based on the local political conditions, often with a tendency to "report those facts which show a good administrative record on the part of the department.”

More recently, Levitt (1998) analyzed crime victimization and reporting data for 26 large American cities and found that the likelihood of a crime reported to the police increases as the size of the city’s police force increases. The study found a 10% increase in the number of sworn officers per capita corresponds to a 1.54% increase in the reporting rate of Household Larcenies. MacDonald (2002) assessed the likelihood of reporting a crime to law enforcement in the United Kingdom, and found that non-White (except Asian), unemployed, and low-income residents were less likely to report crimes. A longitudinal study by Baumer & Lauritsen (2010) of crime reporting from the US National Crime Victimization Study (NCVS) between 1973 and 2005 had similar findings. While the authors found increasing rates of reporting over time, non-White victims and male victims were much less likely to report crimes to the police. More surprising was that overall less than half of nonlethal violent incidents (40%) and property crimes (32%) were reported to the police.

In addition to the underlying demographics and method of measurement, policing strategy behavior can be a significant factor in the measurement and perception of crime in a particular area. Golub et al. (2006) examined the spatial shifts in MPV (Possession of Marijuana in Public) arrests in New York City after the NYPD shifted its enforcement strategy in the early 1990’s as part of Chief Bill Bratton’s embrace of "broken windows” or quality-of-life policing strategy (Golub et al. 2007, Harcourt 2009, Bornstein 2015). The authors found that as the NYPD shifted the focus of MPV enforcement from Transit locations and tourist areas near lower Manhattan to public housing projects in Brooklyn and Queens, this led to a significant shift in both the level and intensity of arrest patterns over time.

As a result of the change in strategy, MPV arrests spike in subsequent years, increasing from 1,851 in 1994 to 39,212 in 2003, with the highest proportion of arrests occurring in predominately Black and Latino neighborhoods. When comparing police units, the NYPD Housing police significantly increased their activity, going from zero recorded arrests in 1994 to 3,769 MPV arrests in 2003 and accounting for 10% of the total in 2003. Conversely, arrests from transit police declined from 499 in 1995 to only 57 in 2003. Aggressive police policies can also have a chilling effect on community behavior. For example, Desmond et al. (2016) used an interrupted time series approach on administrative data to find that 911 calls from minority neighborhoods had a net loss of 22,200 calls after the beating of Frank Jude by police officers in Milwaukee, Wisconsin.

Simply put, crimes recorded by police are not a complete census of all criminal offenses, nor do they constitute a representative random sample. Instead, police records are a compilation of complex interactions between criminality, policing strategy, and community-police relations. Moreover, while questions about how to measure and operationalize crime are often see as debates for academic criminologists, it has become more relevant as the use of criminal justice data has moved into the era of machine learning.

The next sections will discuss in detail how machine learning algorithms are often unaware, and in many cases, unable to adjust for institutional biases and norms embedded within policing data. As a result, the presence of bias in the initial (training) dataset leads to predictions that are subject to the same biases that already exist within the dataset. Further, these biased forecasts can often become amplified if practitioners begin to concentrate resources on an increasingly smaller subset of these forecasted targets. Thus, a failure to understand the limitations of data used in these predictive tools – and create more transparent and accountable mechanisms to mitigate these potential harms may perpetuate historical discrimination toward underrepresented groups and violate their civil and human rights.

1.2 Case Study: Predictive Policing

One of the most popular and fastest growing ADS in criminal justice is "predictive policing" tools, or applications designed to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical predictions (Perry 2013) In a survey of the nation's 50 largest police forces, Robinson & Koepke (2016) reported that more than 30 departments have either deployed or are actively exploring the deployment of a predictive policing system. Outside the United States, European cities such as Kent, London, and Berlin are considering the use of predictive policing (or precrime) tools to predict potential violent gang members (Baraniuk 2015). Winston (2018) found the technology firm Palantir sold their predictive policing technology to the Israeli government for targeting Palestinian dissidents, and a report by Human Rights Watch (2018) uncovered the use of predictive policing by the Chinese government on the Muslim Uyghur population of the Xinjiang region.

While proponents of predictive policing have viewed this trend as a significant step towards transparency and pragmatic, data-driven policymaking, the use of predictive policing and other ADS within police departments has also raised very serious concerns among activists and scholars (Ferguson 2014, Joh 2014, Robinson & Koepke 2016, ACLU 2016, Joh 2017) regarding this new intersection between statistical learning and public policy. Civil liberties advocates have argued the growth of predictive policing means that officers in the field are more likely to stop suspects who have yet to commit a crime under the guise of historical crime patterns that are not representative of all criminal behavior.

Is there any evidence of these potential harms? In their analysis, Lum & Isaac (2016) replicate Predpol's algorithm initially proposed in Mohler et al. (2011) to generate predictions with publicly available data on drug crimes in the city of Oakland from 2009 to 2011. Specifically, the algorithm at the center of this study is the Epidemic-Type Aftershock Sequence (ETAS) crime forecasting model developed by Predpol Inc., one of the largest vendors of predictive policing systems in the country and one of the only companies to publicly release details of their algorithm in a peer-reviewed journal (Mohler et al. 2015). The foundation of the ETAS model is a spatio-temporal branching or "self-exciting" Poisson process referred to as a Hawkes Process based on the seminal research by (Hawkes 1971) into using seismographic activity to predict earthquake aftershocks. More recently, the Hawkes process has used in a wide array of fields, from criminology (Mohler 2013, Mohler et al. 2015) to finance (Bacry et al. 2015), social media (Du et al. 2015), and counter-terrorism (Tench et al. 2016).

Equation 1.1 outlines Predpol's ETAS model as defined in Mohler et al. (2015). Predpol takes a defined geographic area such as a city or police district and divides it into discrete interlocking boxes or bins for isolating allocations of additional policing surveillance. The model then determines which bins are selected for targeting by generating a conditional intensity rate of crime for each bin n at time t by calculating $\lambda_n(t)$ as a function of the background rate μ_n and the triggering kernel $\Theta\omega \exp^{-\omega(t-t_n^i)}$. The background rate is a nonparametric histogram of the counts of recorded crimes in bin n over time $t - t_i$ and can be thought of as the fixed level of crime in a given area.

The triggering kernel parameter captures the model's "near-repeat" or "contagion" effects in crime data. In particular, the decaying exponential function gives a higher weight to bins with recent spikes in recorded crimes compared to bins with higher background rates and declining rates of recently recorded crimes. This parameter is very similar to the hotspot maps that have become very common within police departments across the country. As Mohler et al. (2015) point out, a critical difference between the ETAS model and hotspot maps such as Compstat which model near-repeat effects are the introduction of the background rate μ_n.

$$\lambda_n(t) = \mu_n + \sum_{t_n^i < t} \Theta\omega \exp^{-\omega(t-t_n^i)} \tag{1.1}$$

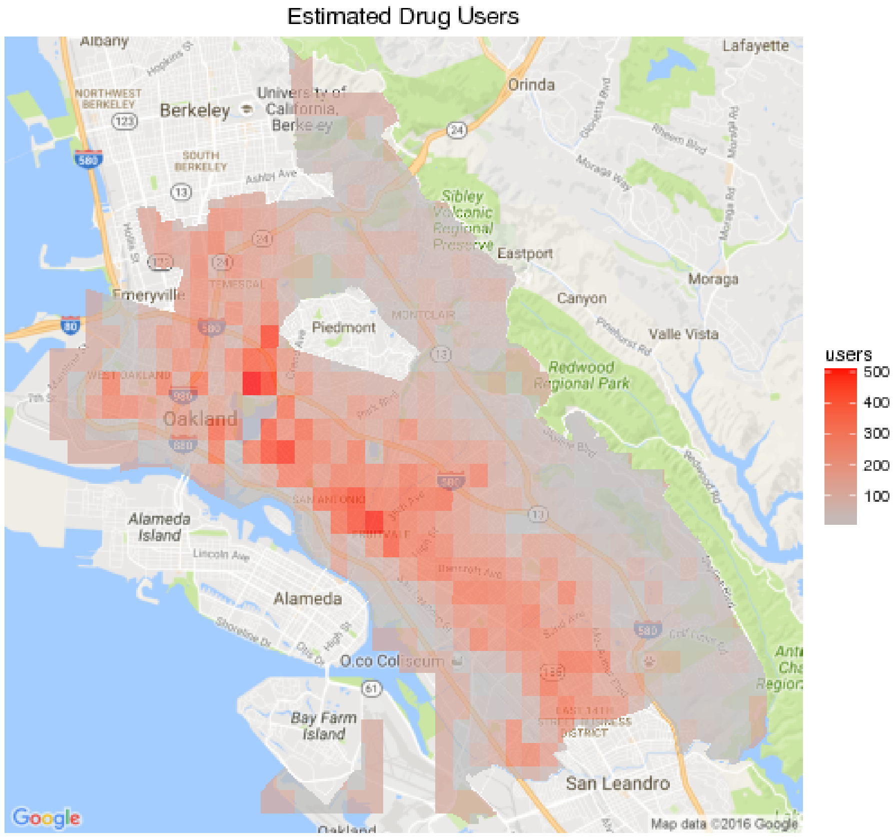
While the ETAS model is novel for its ability to capture both short and long-term fluctuations in criminal activity, it is also heavily dependent on the quality of the input data used. Its founders have touted Predpol as a parsimonious "race-neutral system" that uses only three data points in making predictions: past type of crime, place of crime and time of crime (OMalley 2013, Smith IV 2016). In their paper, Mohler et al. (2015) note that the use of criminal records alone is a feature rather than a bug, as using fixed environmental characteristics of a hotspot explicitly such as census data or locations of crime attractors could introduce bias. However, Lum & Isaac (2016) argue that the unrepresentative nature of recorded crime data could bias predictive policing forecasts in two crucial ways.

First, the presence of bias in the initial training data leads to predictions that are subject to the same biases that already exist in crime data. As predictions generated by ETAS are likely to over-represent areas that were already known to police, and as a result increase patrols of these same areas. In the process of this increased surveillance, officers are more likely to observe new criminal acts that confirm their prior beliefs regarding the distributions of criminal activity. To test this hypothesis, the authors sought to address a vital counter-factual: what is the universe of crimes in a given location? Determining this counter-factual scenario is often tricky because virtually all available crime data originates with police recorded databases. Thus, the authors generated local estimates of the number of illicit drug users within the city using public health responses from the 2011 national survey on drug use and health (NSDUH) on recent use of a series of illicit drugs with a synthetic population of the city of Oakland. These estimates provided a unique and novel way to compare an estimated counter-factual scenario with the observed activity of the Oakland Police department.

Figure 1.1 below, reprinted from the original study, plots the estimated number of drug users in Oakland, California divided by 150 x 150-meter bins, with the greater color saturation indicating a higher number of estimated drug users in a particular location. As we see from figure 1.1, the spatial distribution of drug users appears to spread across the city, with elevated levels along International Boulevard, which is a primary thoroughfare in the city of Oakland. In comparison to figure 1.2, also reprinted, which plotted the number of reported crimes across the city divided by bins, and demonstrates a pattern concentrated around the neighborhoods around West Oakland and Fruitvale, two neighborhoods with mostly non-white and low-income populations. Variations in the latter are driven primarily by differences in population density, as the estimated rate of drug use is relatively uniform across the city.

From these figures, it is clear that police databases and public health-derived estimates tell

Figure 1.1: Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

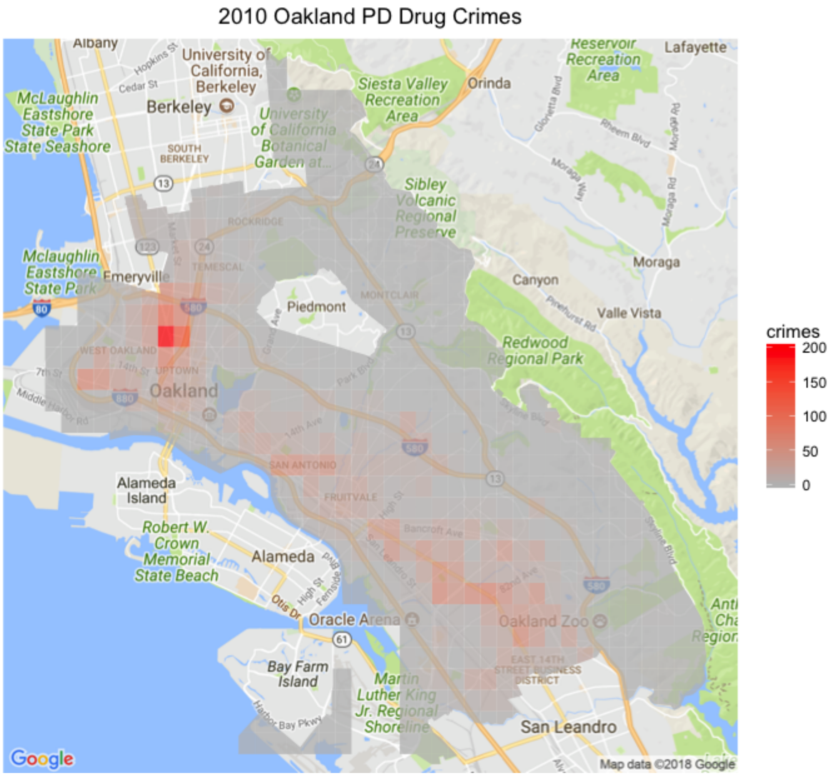


Source: Lum & Isaac (2016, pg. 17)

dramatically different stories about the pattern of illicit drug use in Oakland. The critical question is whether the predictions generated from the ETAS model align with the crime data or estimates from public health data. Figure 1.4 plots the number of days each bin would have been flagged by Predpol for targeted policing, with greater color saturation indicating a higher number of days targeted.

Given the few numbers of bins targeted by the ETAS algorithm, it is clear the model failed to capture the more considerable spatial variation found in the public health data and closely resembles the reported crime data. Further, this optimized targeting led to a higher concentration of targeted policing among minority and low-income neighborhoods, as Black populations in Oakland would

Figure 1.2: Number of drug arrests made by Oakland police department, 2010

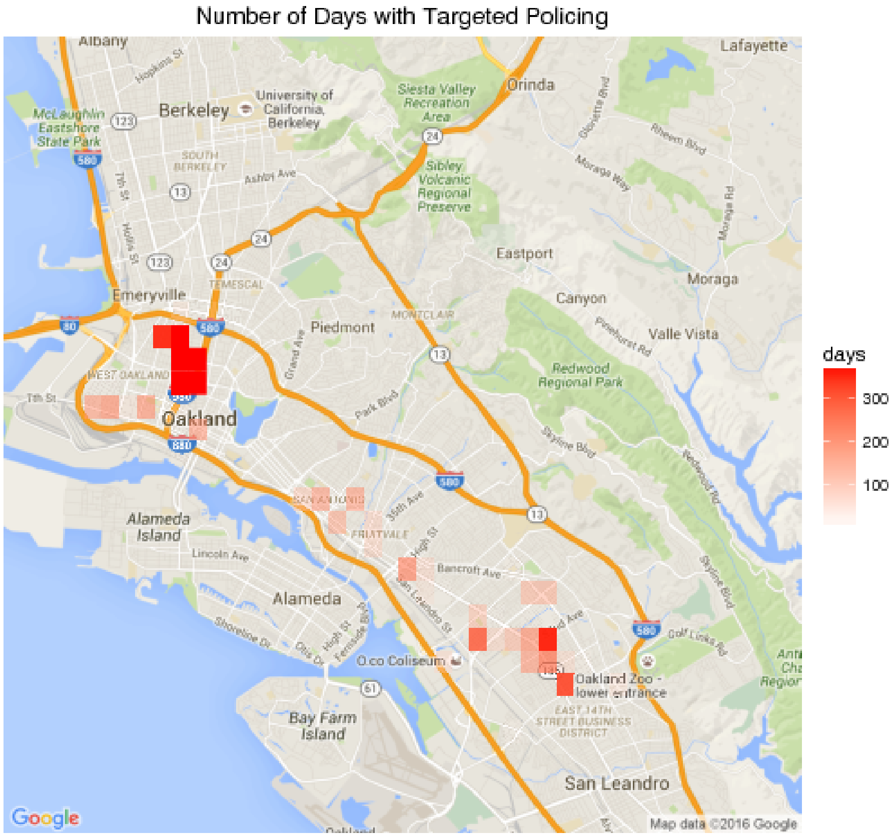


Source: Lum & Isaac (2016, pg. 17)

have received targeted police at approximately twice of the rate of the White population and other racial groups would have received targeted policing at 1.5 times of whites. The finding appears to run counter to the claims of predictive policing proponents who suggest that police recorded data is neutral and will generate unbiased allocations of policing resources.

In addition to mimicking the existing social norms in the data, the newly observed criminal acts that police document as a result of these targeted patrols then feeds into the predictive policing algorithm on subsequent days, generating increasingly biased predictions. The newly observed criminal acts that police document as a result of these targeted patrols then feeds into the predictive policing algorithm on subsequent days, generating increasingly biased predictions. This feedback loop or "ratchet effect" (Robinson & Koepke 2016) can lead to model over-fitting in that the locations most likely to experience further criminal activity are precisely the locations they had previously believed to be high in crime.

Figure 1.3: Number of days with targeted policing in areas flagged by PredPol analysis of Oakland police data

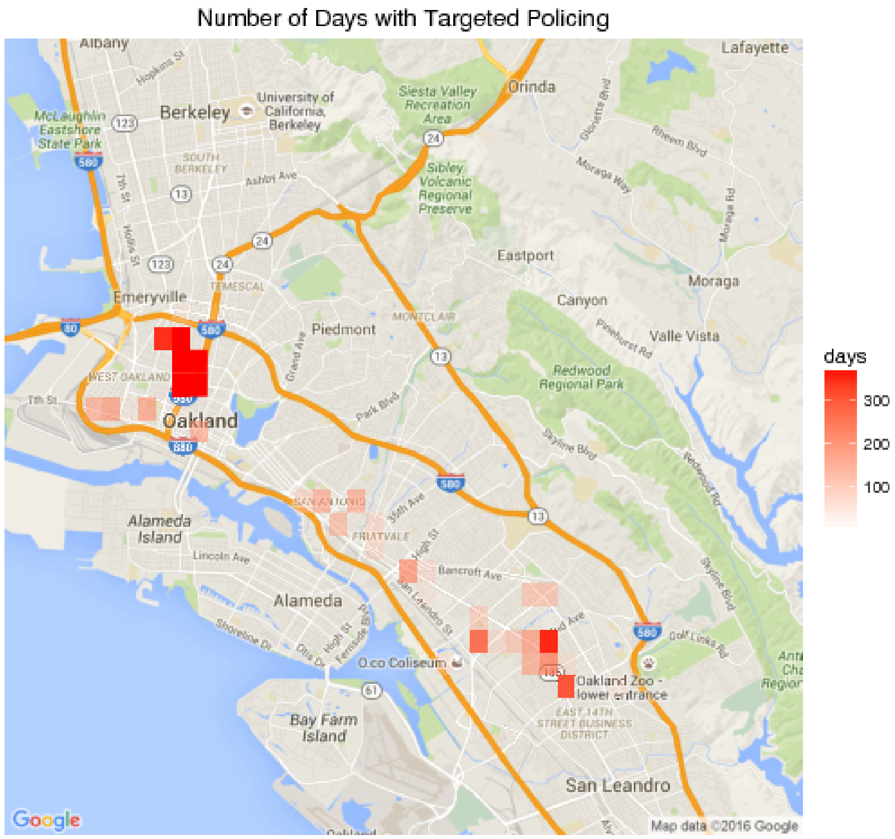


Source: Lum & Isaac (2016, pg. 17)

In their study, Lum & Isaac (2016) attempt to address this issue by simulating a scenario of the application of the ETAS model where in addition to the observed Oakland crime data there is an additional 20% chance that additional crimes in the targeted bin are discovered for a given day. In this scenario, the authors find a significant increase in the predicted odds of targeting previous bins versus to non-targeted bins compared to the baseline example (see figure 1.5). This evidence led the authors to conclude the feedback scenario "causes the Predpol algorithm to become increasingly confident that most of the crime is contained in the targeted bins (pg. 18)." In short, the feedback mechanism is selection bias meets confirmation bias.

Overall, while the study has pointed out the potential disparate impact when using crime data

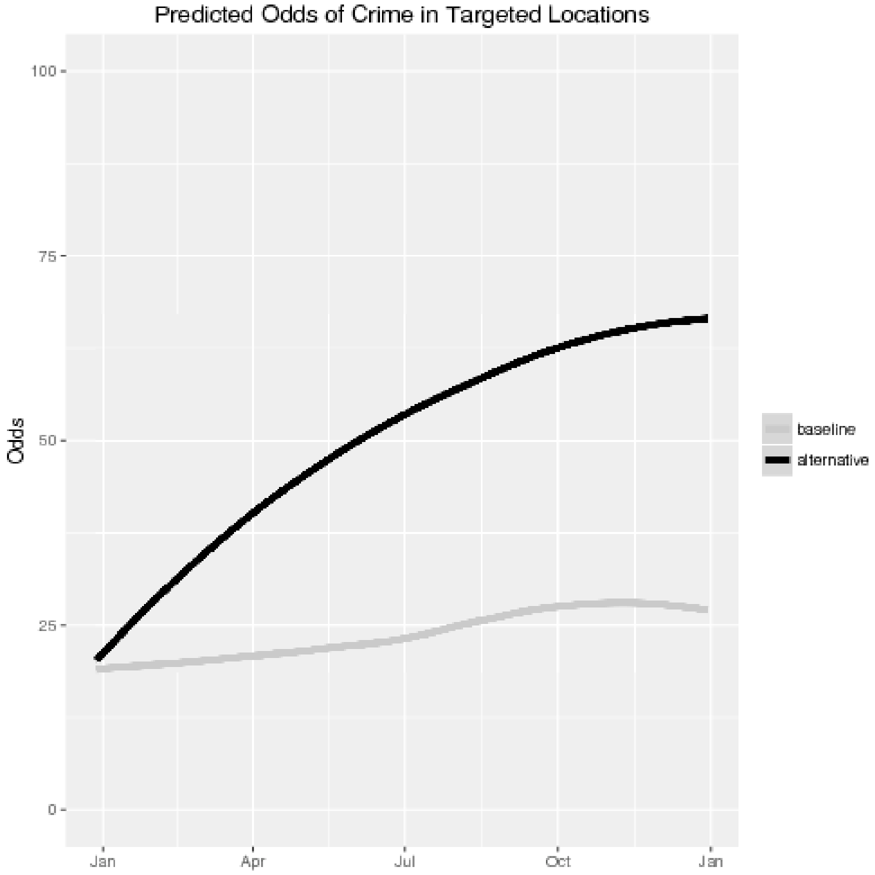
Figure 1.4: Number of days with targeted policing in areas flagged by PredPol analysis of Oakland police data



Source: Lum & Isaac (2016, pg. 18)

in predictive models, critics of the study have questioned the appropriateness of using drug crime data for generating forecasts using Predpol's ETAS algorithm, as Predpol has claimed they do offer departments forecasting for drug crimes (Smith IV 2016). However, this claim fails on many fronts. First, as the authors stated explicitly in the study, drug crimes were used because it allowed for the use of public health data on illicit drug use to serve as a counter-factual against observations recorded by police departments, not because Predpol or other major predictive policing vendors were known to be widely forecasting drug crimes. Second, most major police departments use some form of "Hot Spot" policing tactics (Braga 2005, Weisburd 2018) – which serve as the theoretical foundation for predictive policing – to isolate areas where they believe drug activity to be occurring.

Figure 1.5: Predicted odds of crime in locations targeted by PredPol algorithm, relative to non-targeted locations.



Source: Lum & Isaac (2016, pg. 19)

Lastly, despite predictive policing vendors' claims, governments are interested in using predictive ADS to target drug crimes. For example, the National Institute of Justice included drug crimes as part of their recent $1.2 million crime forecasting challenge (National Institute of Justice 2016). Moreover, local police departments have explicitly proposed using predictive policing for predicting the location of drug crimes and gang activity (Ramunni 2015, Hamden Police Department 2015). In fact, Bond-Graham & Winston (2013) uncovered email exchanges between Predpol and the San Fransisco Police Department from July 2012 to August 2013 where Predpol's lobbyist Donnie Fowler sold the company's ability to predict drug crimes, "The crimes we predict are burglary [residential, commercial, auto], auto theft, theft, robbery, assault, battery, and drug crime." So, as the concern within the Trump administration about the growing opioid crisis (Naylor & Keith 2017, Hirschfeld Davis 2017) increases, it is indeed reasonable to assume that more cities will look to predictive policing or other policing technologies address drug crimes.

1.3 Discussion

Using predictive analytics in the real world is challenging, especially in high-stakes policy areas such as policing. However, this does not mean police departments should abandon the use of analytics or intelligence-led approaches to improving public safety. Instead, it is essential for police departments and other law enforcement agencies to think more broadly about the potential impacts of implementing algorithmic decision-support tools and ensure they create internal and external systems to promote public safety while minimizing disparate impacts. Police departments or agencies that attempt to implement algorithmic decision support tools should take steps to develop internal and external accountability, ensure operational transparency, and be aware of the long-run impact to the community.

The process of accountable and transparent use of algorithmic decision support systems must start with community stakeholders and police departments discussing policing priorities and measures of police performance. Currently, nearly all predictive policing systems aim to identify high-risk neighborhoods or individuals, which departments then use to intervene with additional surveillance or adverse enforcement actions (i.e., arrest or citation).

While some surveillance is needed in order to ensure public safety, recent studies have found that persistent police surveillance leads to worsening mental and physical health outcomes for the underlying neighborhoods over time (Sewell 2017, Sewell et al. 2016). Adverse enforcement actions have also been found to be "contagious" (Lum et al. 2014) in communities targeted by police, leading to a deterioration of police-community relations and perpetuating the mass incarceration crisis in the United States (Western & Wildeman 2009, Alexander 2012, Coates 2015).

The promise of leveraging data to improve public safety also provides an opportunity to reform how we think about policing, and a growing number of innovative departments have started to pursue these alternative approaches. For example, Toronto and other cities in Canada have moved toward the "HUB and COR" model of predictive policing, which seeks to leverage police departments as a conduit for access to other social services rather than leaving the police with a narrow range of tools to address a myriad of social problems.

Under this model, a predictive model uses data collected from multiple governmental agencies to flag at-risk individuals (often minors) in need of urgent intervention by law enforcement. This approach reflects recent experimental evidence which suggests leveraging social services can be an effective strategy for reducing crime, particularly violent crime among juveniles (Bushman et al. 2016, Heller 2014, Barr & Gibbs 2017). However, this approach is not without its challenges. Civil society groups and journalists have rightly raised concerns in about potential abuse of the system due to weak privacy protections such as personally identifiable information of the targeted persons (Munn 2017), new advances in blockchain technology – the algorithm that serves as the foundation to the cryptocurrency bit-coin (Nakamoto 2008) – could potentially allow HUBs to generate predictive risk scores and identify at-risk persons in a manner which respects the privacy of persons in the datasets (Zyskind & Nathan 2015, Kilbertus et al. 2018).

One shortcoming of this approach is that it would be very resource intensive to carry out, and most police departments or city governments cannot rigorously develop or assess these type of tools. However, an alternative approach would be for civil society groups to provide a set of best practices and have a volunteer panel of civil society groups and university experts routinely conduct oversight reports on existing practices. If we are successful in developing better guidelines and techniques for ADS tools, cities will both improve the quality of the data they collect and implement more transparent and inclusive processes to build safer communities for all of their residents.

CHAPTER 2

THE NUMBERS GAME: GOODHART'S LAW AND THE USE OF THE INDEX CRIMES AS A PERFORMANCE METRICS

It was a great idea that has been corrupted by human nature.

— Robert Zink*, Second Vice President, New York Patrolmen's Benevolent Association*

As various segments of the media (Brustein 2017, Winston 2018), academic scholars (Kim et al. 2014, Joh 2017), and public officials (Podesta et al. 2014, Winston 2015, Tchekmedyian 2016*b*) have noted the future of criminal justice lies in the rapid expansion of Algorithmic Decision Systems (ADS) and other data-driven policy tools. Virtually all of these tools center around the use police recorded data – which criminologists and political scientists have questioned (Morrison 1897, Robison 1936, Beattie 1941, Levitt 1998, Wormeli 2018) – to optimize specific metrics which it deems critical to society or the agencies' mission. These also means that much of the potential societal gains from these tools hinge on the accuracy of administrative data often tightly connected to internal incentive structures and the consistent application of bureaucratic discretion. While previous articles have focused on potential gaming of ADS or other public machine learning applications by external actors (Veale 2017, Lazer et al. 2014), few studies have examined the potential impact that ADS have on gaming by internal actors.

This essay will explore the problematic nature of linking policy outcomes to data, and how it can alter institutional and individual behavior. It will focus on introducing and defining a commonly discussed but often ambiguous concept known as Goodhart's law (Goodhart & Courakis 1981, Goodhart 2006, Manheim & Garrabrant 2018). Next, using a novel application of methods developed in the political science literature on election monitoring (Beber & Scacco 2012), this essay will examine the impact that the implementation of ADS have agent incentive structures and behavior. Specifically, the paper will focus on the impact of that the deployment of algorithmic decision systems – namely the predictive policing platform Predpol (Mohler et al. 2015, Brantingham et al. 2018) – has on the reporting of property crimes by the Los Angeles Police Department (LAPD). The final section will conclude with a discussion of implications of the findings and about potential directions for future research.

2.1 Performance Metrics and Goodhart Effects

2.1.1 Defining Goodhart Effects

2.1.1.1 Goodhart's Law

It is rare for any topic within the social sciences to have laws in the same manner that fields such as Physics, but the human capacity to undermine the validity of quantitative measures has the special designation of having two. The first of these self-defined laws emerged in a series of papers written by economist CAE Goodhart, who investigated the failure of new central bank policies on the growth rate of the money supply, and found that the failure stemmed from Central Bank economists' assumption that the behavior market participants under a previous regime would remain constant under a new one. This lead to Goodhart to remark, "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes (Goodhart & Courakis 1981, Goodhart 2006)." Evans (1985) defined the Goodhart's law (GHL) more formal as equation 2.1, where y is a goal variable sought by policymakers to remain as stable as possible and m serving as a monetary variable which can be controlled indirectly through its policy instruments r over time period t = 1,2,...,T. Goodhart's Law states that if use it uses r to reduce the variance of m (which would allow y to stabilize) at period T include could lead the variance of m to increase, leading to a disconnect between the policy target and measure.

$$y = \alpha m_t + e_t \tag{2.1}$$

While Goodhart's Law was conceived for applications in monetary policy (Goodhart & Courakis

1981, Evans 1985, Chrystal et al. 2003, Goodhart 2006), economists and policy researchers eventually began to apply this more broadly to cases where interventions made by policymakers (Eskeland & Feyzioglu 1997, Elton 2004, Newton 2011, Reynaert & Sallee 2016, Edwards & Roy 2017) – often in the form of benchmarks or policy targets – lead to a deterioration of the metric used for guiding that particular policy due to gaming or cheating. However, there a common misconception within the literature about the interpretation of the law. Goodhart's law does not assume that every policy intervention will lead to gaming or data manipulation, just that the consistency of the relationship between measure and goal will deteriorate over time. So, the aim is to separate the general law from its potential policy effects. The impact as expected from the original definition of GHL can be categorized as a **benign Goodhart effect**. These are instances when a policy-induced optimization causes a collapse of the statistical relationship between a goal (which the focus of the optimization) and the proxy used for that goal. In the next section, this type of distinction is needed to understand other types of effects on measurement due to policy interventions such as gaming.

2.1.1.2 Campbell's Law

Campbell's Law (Campbell 1979) could be considered as an adversarial form of Goodhart's Law. Rather than an intervention merely leading to a disconnect between a quantitative metric and underlying policy target, Campbell's Law stipulates that "[t]he more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor (Campbell 1979, pg. 85)." In other words, an **adversarial Goodhart effect** (AGE) is distinct from cases where the policy-induced optimization because it alters agents' incentive structure, leading to an intentional erosion of the statistical relationship between a goal and proxy. Some economists and technologists refer to AGEs as "cobra effects" (Siebert 2001, Vann 2003, Dubner & Levitt 2012, Manheim & Garrabrant 2018) because of an unsubstantiated story from colonial India where British authorities offered a reward for dead cobras. Instead of hunting cobras, residents bred and killed their cobras to receive the reward, which ultimately led to a purported increase in the cobra population.

The literature on the persistence of AGE has emerged in recent years, particularly in the domain of public policy. For example, in education policy, which experienced a surge of scholarly interest in this topic in the wake of the No Child Left Behind (NCLB) mandate for state-wide standardize testing (Connelly et al. 2016). In their now famous study, Jacob & Levitt (2002) examined erasure marks on standardized tests in Chicago public schools and found that the teacher or administrator cheating occurred in a minimum of 4-5 percent of elementary school classrooms annually.

Outside of education policy, Bevan & Hood (2006) investigated the infamous “targets and terror” reforms of the United Kingdom’s National Health Service (NHS) by the Blair administration. The aim of was a creation of a system of governance built around annual performance (star) ratings of NHS organizations, to improve the efficiency and performance of the myriad of NHS entities. While the authors did find meaningful reductions in hospital wait times due to the introduction of the star ratings, they also uncovered a significant amount of gaming by the various NHS trusts (or hospitals systems). For example, to reduce the emergency room wait times, some hospitals required patients to wait inside of ambulances outside the hospital until they were confident that that patient could be seen within four hours. Further, some trusts deliberately adjusted the records of roughly 6000 patients to ensure their waiting lists conformed to administrative requirements.

A much less discussed AGE is raised by Amodei et al. (2016), who suggest that in applied machine learning contexts where the proposed intervention creates a “positive feedback loop” between a goal and its proxy which can amplify the correlation between the two over time by such a degree that it drowns out or severely distorts the value of the measure. Perhaps the most suitable example of this phenomenon is the impact of recorded crime data in predictive policing algorithms (Lum & Isaac 2016, Ensign et al. 2017) where the introduction of additional crimes in a particular location led to increased clustering of targeting by Predpol’s ETAS (Mohler et al. 2011, 2015) into historically over-policed neighborhoods over time.

Overall, this section sought to clarify some of the ambiguity surrounding Goodhart’s Law and its potential effects. First, contrary to the typology of Manheim & Garrabrant (2018), any instance of Goodhart’s Law must have an intervention into the data generation process, generally through a policy change or rule. This requirement is needed to distinguish genuine Goodhart effects from questions of a particular metric's validity and consistency. Second, while the term is commonly associated with gaming and cheating as a result of a proposed new policy, the formal definition of Goodhart's Law never stipulated such an effect. Goodhart's Law merely stipulates that the "statistical regularity" will collapse under policy control. These are important distinctions if – as will be demonstrated in the next sections – the aim is to statistically assess the potential of Goodhart effects in given policy implementation.

2.2 Goodhart Effects and Policing

While Goodhart effects emerge in a wide range of policy areas, to date there is scant mention of it in the criminology literature [1]. Conversely, there is a robust body of academic, government, and media reports regarding the presence of gaming and manipulation by police departments as they have begun the transition into a greatly expanded role for data-driven policing.

The bulk of the literature on gaming in policing centers not around modern tools such as predictive policing, but on earlier technologies such as the widely deployed compare statistics or Compstat geographic crime mapping tools developed in the early 1990's by former New York Police Department Commissioner Bill Bratton and Deputy Commissioner Jack Maple (Firestone 1996, Martin 2001, Walsh 2001). At the core of the Compstat process is a computerized information dissemination system, which is expected to process, map and analyze weekly crime statistics that are then presented and debated by operational managers at weekly oversight meetings to ensure progress on achieving the stated goals (Walsh 2001, Buntin 2009). This approach Bratton argued allowed the department to blend traditional models of policing with the more flexible community and problem-oriented policing strategies (Willis et al. 2010, White 2013, Roeder et al. 2015). While

[1] Ironically, the example Campbell (1979) used to illustrate AGEs was the rise and demise of the use of clearance rates as a metric of police productivity. He noted that while many police departments viewed their clearance rates (the number of crimes recorded divided by the number of crimes solved) as validation of departmental productivity, Campbell noted that a variety of factors such as failing to record citizen complaints, postponing documenting a complaint until it's solved, and leveraging a plea-bargain for an offender to include multiple unrelated cases as examples of policy-induced gaming

Compstat was initially met with skepticism within the NYPD, it was quickly heralded by many as the key for the dramatic turn around in New York City crime rates during the 1990's. In the first year after CompStat's implementation, crime dropped 12 percent, and between 1994 and 2012, there was a 63 percent decrease in crime reported to the police compared to a decrease of 27.2% Nationwide during the same period (Roeder et al. 2015).

Despite the public acclaim for Compstat, its proponents often omit the well-documented flaws of the system. In his excellent essay on the use of quotas in law enforcement, Bronstein (2015) articulates the dilemma facing officers under a quota-based system such as CompStat. Police officers inherently have a high level of discretion in the execution of their duties (Lipsky 1971, Evans & Harris 2004, Lipsky 2010, Tummers & Bekkers 2013). However, from a management perspective, it is a difficult task to enforce compliance and uniformity of policing practices when each subordinate has the relative autonomy to achieve their goals. A quota system (or a performance-based incentive system) provides supervisors a mechanism to decrease an officer's discretion by forcing the officer to take enforcement action in what would otherwise be discretionary situations.

A quota system, depending on the metrics used, can reduce the dimensions an officer has to consider when weighing between options in the execution of their duties (i.e., choosing between policing public space and improving police-community relations). If there is a limit or quota to the number of stop and frisks one can conduct in a given period, then the officer would be forced to choose an alternative strategy to resolve a particular situation. As the author points out a quota can also skew the motivations which underlie enforcement action, in particular, they can: 1) an increase in misconduct and corruption and (2) de-prioritize community-police relations. The next few paragraphs will discuss in detail how these particular effects played a role in the implementation of CompStat.

The most common criticism of Compstat is that it boils policing down into a "numbers game" and dramatically increases the incentive for officers to game the system by altering how crimes are classified or recorded (Eterno & Silverman 2010, Bornstein 2015, Bronstein 2015, Roeder et al. 2015). Systems such as Compstat provided a high degree of incentives to officers and operational managers who were able to achieve the performance goals on critical strategic initiatives determined by the department. However, the pressure to achieve the metric-driven goals are also linked to police officer corruption within the NYPD before Compstat, driving police officers to engage in acts of corruption such as false arrest and falsifying documents to hit their performance goals (Purdum 1987, Knafo 2016).

However, it may also be the case where the aggressive accountability and surveillance of the officers' performance prevents them from trying to game the system. Eterno & Silverman (2010) attempt to address this precise question through conducting qualitative interviews with former NYPD officers and a survey of those who served and left the force before and after the implementation of Compstat. The study examines the negative consequences of the Compstat model, including its shift to greater reliance on aggressive practices and official sanctions, its increasing centralization of power, and its overemphasis on generating numbers rather than problem-solving, community buy-in and cooperation. In one of their more memorable quotes, Robert Zink, Second Vice President of New York Patrolmen's Benevolent Association explains in detail how easy it is for officers to manipulate crime statistics,

> "It [Compstat] was a great idea that has been corrupted by human nature. The Compstat program that made NYPD commanders accountable for controlling crime has degenerated into a situation where the police leadership presses subordinates to keep numbers low by any means necessary. The department's middle managers will do anything to avoid being dragged onto the carpet at the weekly Compstat meetings...So how do you fake a crime decrease? It's pretty simple. Don't file reports, misclassify crimes from felonies to misdemeanors, under-value the property lost to crime so it's not a felony, and report a series of crimes as a single event. A particularly insidious way to fudge the numbers is to make it difficult or impossible for people to report crimes — in other words, make the victims feel like criminals so they walk away just to spare themselves further pain and suffering (pg. 428)."

To test these assertions empirically, the authors conducted a self-reported, anonymous, mail-based survey of retired members of the NYPD in the ranks of Captain and above. The survey sample consisted of 166 (33.9%) respondents who retired before 1995 (the first full year of Compstat in New York City) and 323 (66.1%) who retired 1995 and after. The survey questions sought to measure the level of pressure officers faced complying with various aspects of the Compstat culture, ranging from pressure to change the crime statistics to the aggressive application of broken windows policing.

In table 1 of their paper, the authors present the difference in means for each of the variables in their survey, where responses lie on a 1 (corresponds to 'least pressure') to 10 ('10' corresponds to 'most pressure') scale representing the level of pressure they faced to conduct a particular activity. The variable with the highest pre-post difference (2.605) was pressure to decrease the rate of index crimes or crimes which count for the FBI's annual uniform crimes statistics report. Pressure to downgrade index crimes to non-index crimes also had a statistically significant increase as well (1.365).

The concerns about the pressure of Compstat leading to manipulation of index crime statistics followed Bratton into his tenure as Chief of the Los Angeles Police Department (LAPD). In the series of focus groups conducted by Stone et al. (2009), some officers complained the organizational focus on crime reduction led to a reduced focus on the elements of policing and lead to data manipulation. As the report noted (emphasis mine),

> "Many senior officials with whom we spoke seemed concerned that Compstat may focus so heavily on crime reduction that other goals are neglected. As one told us, 'as long as you just push on crime, other stuff will go by the wayside...The Chief may not fully appreciate how Compstat and the constant push on crime squeezes out space for supervisory oversight in the organization.' Another problem with this push on crime data—of which all managers are keenly aware — is **the risk that crime recording will be manipulated by police officers trying to game the Compstat process**. Indeed, one officer suggested to us that he had personal knowledge of officers recording burglaries as vandalism in order to produce reductions in burglary numbers. New audit procedures,

> well beyond what the consent decree requires, have been implemented to detect and prevent just this sort of manipulation (pg.40)"

Most worrying is that, despite heightened concerns expressed by many regarding the ability of cops to fudge the numbers, cases still seem to occur across the country. Hart (2004) found that the Atlanta police department had significantly underreported violent crimes in an attempt to boost the perception of public safety in advance of the 1996 Summer Olympics. A report by the New Orleans Inspector General (Quatrevaux 2014) found that the New Orleans Police Department had an "institutional problem" with underreporting rape crimes as sexual assaults to minimize their impact on the city's Index crime reporting.

In response, cities such as New York and Los Angeles have tried to quell public concerns by creating audit teams to review reported crime statistics. Despite this additional oversight, allegations of manipulating crime statistics remain. For example, Parascandola (2010) reported of complaints by an NYPD officer who alleged that his precinct recorded felonies as misdemeanors and refused to take complaints from victims — all to drive down the crime rate. In early 2018, NYPD Capt. Marash Vucinaj alleged that commanders in multiple precincts were altering their crime statistics, despite the use of algorithms to detect anomalies in reporting, to advance their careers within the department (Rayman 2018).

The LAPD has also had multiple reports of data manipulation of crime data. For example, Poston & Rubin (2014) uncovered over 1,200 felony violent crimes that were downgraded to misdemeanors to reduce the Index crime numbers reported due to pressures related to Compstat. Poston et al. (2015) extended this investigative reporting and found that the LAPD had downgraded over 14,000 violent crimes dating back to 2005. In addition, violent crime in the city was 7% higher than the LAPD reported in the period from 2005 to fall 2012, and the number of serious assaults was 16% higher than previously reported.

More recently, Van Nuys Station Capt. Lillian Carranza filed a lawsuit against the LAPD in which she claims the department retaliated against her for whistle-blowing on the continued practice of underreporting violent crime statistics (Chou 2018). These repeated claims of data manipulation can only serve to raise further concerns about the validity of using index crimes as a measure or indicator of the state of public safety in the big data era of policing. Further, there also appears to be a clear need to determine a more routine and systematic way to quantify and assess this phenomenon beyond the existing auditing practices used by police agencies.

Overall, it seems clear from previous studies and the qualitative investigations that administrative data are a frequent venue for manipulation by agents seeking to achieve their internal bureaucratic goals. While this may historically have a limited impact on a specific case, this issue becomes more critical in an era where machine learning and artificial intelligence will increasingly rely on historical data to generate predictions to allocate public resources or aid institutional decision-making. Unlike the Compstat process, which was designed to incorporate alternative information, Predpol's ETAS algorithm is solely dependent on police recorded data for generating predictions of potential crime hotspots. If the input data used in the algorithm were subject to manipulation, these biases would merely perpetuate within subsequent predictions. In addition to the concern regarding the input data influencing the predictions generated, there is the new concern about the reciprocal impact of the data influencing officer judgment and behavior in the field (Lum & Isaac 2016, Isaac 2017, Isaac & Lum 2018).

As the experience with Compstat demonstrated, new technology can induce adversarial Goodhart effects by amplifying the incentive to manipulate reporting procedures to artificially reduce crime. But, will newer algorithmic tools such as predictive policing have the same effect? The subsequent sections will seek to answer these questions by examining police recorded property crime data collected by the LAPD before and during a pilot deployment of the Predpol Platform. According to Brantingham et al. (2018) & Mohler et al. (2015), Predpol's software deployed in three participating LAPD divisions: Foothill (FH), North Hollywood (NH) and Southwest (SW) between the period of November 2011 and January 2013 [2]. The pilot aimed to compare the performance of

[2] Because of the unique randomization method, the pilots had varying lengths of implementation.Brantingham et al. (2018) attributed the variation to the inability of a human analyst present to generate a control map for a given day. Overall, in the Foothill Division a total of 124 test days with a successful random assignment, was divided into 62 control and 62 treatment days. The North Hollywood division had a total 152 test days, split into 82 control and 70 treatment days. Lastly,

Predpol against traditional human crime analysts using Compstat (Mohler et al. 2015). To do this, the shift commander would receive a randomly assigned patrol map generated by either a human analyst or a map Predpol generated using their proprietary algorithm before the beginning of each of the three daily patrol shifts.

Each map contained twenty target areas marked as 500 x 500-foot boxes. The shift commanders were not informed of the origin of the crime maps and were only instructed to inform the officers to patrol the areas in their discretionary time (Brantingham et al. 2018). The deployment had a stated focus on targeting burglary, car theft and burglary theft from vehicle (BTFV) because they can account for as much as 60% of police recorded crimes in the City of Los Angeles (Brantingham et al. 2018).

Overall, the Police patrols using ETAS forecasts led to an average 7.4% reduction in crime volume, while patrols based upon analyst predictions showed no significant effect on crime volume across the three test divisions. The authors also noted that while the findings are suggestive of potential crime deterrence, the evidence of a drop in crime due to ETAS was not significant (Brantingham et al. 2018). While this study is not interested in re-assessing whether ETAS led to a significant drop in crime volume, we are interested in a related question: did the introduction of Predpol led to gaming or data manipulation by LAPD patrol officers? Also, if there is evidence of gaming did the Predpol's deployment spark a "feedback loop" of amplified gaming or manipulation over time?

2.2.1 Methods & Data

The methods used to identify potential adversarial Goodhart effects in the reporting of crime data is a digit-based test employed in election and accounting fraud (Durtschi et al. 2004, Beber & Scacco 2012, Badal-Valero et al. 2018) that seeks to exploit human biases in number generation. Psychology research from a wide variety of domains from sports and standardized testing (Pope &

the Southwest Division had a total of 234 test days, which were divided evenly into 117 control and 117 treatment days.

Simonsohn 2011), to stock market pricing (Kandel et al. 2001, Osler 2003), tipping (Lynn et al. 2013, Azar et al. 2015), and survey results (Crawford et al. 2015) that humans have an attraction to round numbers (i.e., 5 or 10). These tendencies lead to "heaping" number patterns (Crawford et al. 2015, Kobak et al. 2016, Rozenas 2017) within human manipulations datasets as people seek to use these numbers as reference points.

It also reflects a human inability to reproduce random patterns of digits, as Beber & Scacco (2012) noted that humans have difficulties reproducing random patterns of equifrequent last digits, even when they induced with incentives to do so. For example, Rath (1966) asked 20 university students to each produce 2500 random digits by filling in 10 one-page grids of 250 rectangles each. The students were told that they could leave the experiment as soon as they filled in all the sheets. The author observed that the students exhibit a strong preference for small numbers (1, 2, and 3) over both larger numbers (5, 7, 8, and 9) and zero.

Moreover, he found strong biases against repetitive pairs of digits but strong biases in favor of adjacent pairs of numbers (such as 12 and 23). In a similar experiment, 458 university students were each asked to produce random sequences of 25 single digits. Much like the previous study, the authors found that subjects had a strong preference for small digits (1,2,3) over larger ones and (somewhat counter-intuitively) had an active avoidance of repetition as 70% of subjects failed to repeat a single digit in their 25-digit sequence.

It appears from the literature that efforts to manipulate data tend to follow consistent patterns around the distributions of digits, but some numbers (0 or 5) may represent a unique "heaping" behavior when seeking to use a metric as a reference point. The analysis in the next section will test whether counts of police recorded crime data exhibits any of these digit patterns. Boland & Hutchinson (2000) utilize the last-digit test based on the notion that, in normal circumstances, any last digit between 0 and 9 should occur with equal likelihood (Beber and Scacco 2012).

Assuming that the number of crimes is of a sufficiently large range and the mean is larger than the standard deviation, if some digits occur significantly more often than others, it can serve as evidence of fraudulent behavior. The null hypothesis of no fraud can then be tested using the chi-squared test that compares the observed distribution of last digits with the null uniform distribution. The last digit method serves as a conservative test of the hypothesis of gaming because the method tends to produce more false negatives than positives (Rozenas 2017). In theory, a "clean" crime dataset should have digit frequencies that reflect a near random digit process if it does reflect a latent propensity for patterns in criminal behavior. Thus, our null hypothesis is that crime reporting patterns will be unchanged by institutional interventions, such as the introduction of Predpol in a particular district.

To measure the magnitude of distortion from the equifrequent distribution within a particular district, a modified version of the mean absolute deviation (MAD) measure developed by Nigrini (1996) and outlined in 3.1 and sum of squares difference (SSD) measure proposed by Kossovsky (2014) is used. MAD scores are an empirically-based whole-test measure that takes the average of the absolute deviation of each digit's frequency from the ideal equifrequent distribution. They are calculated as

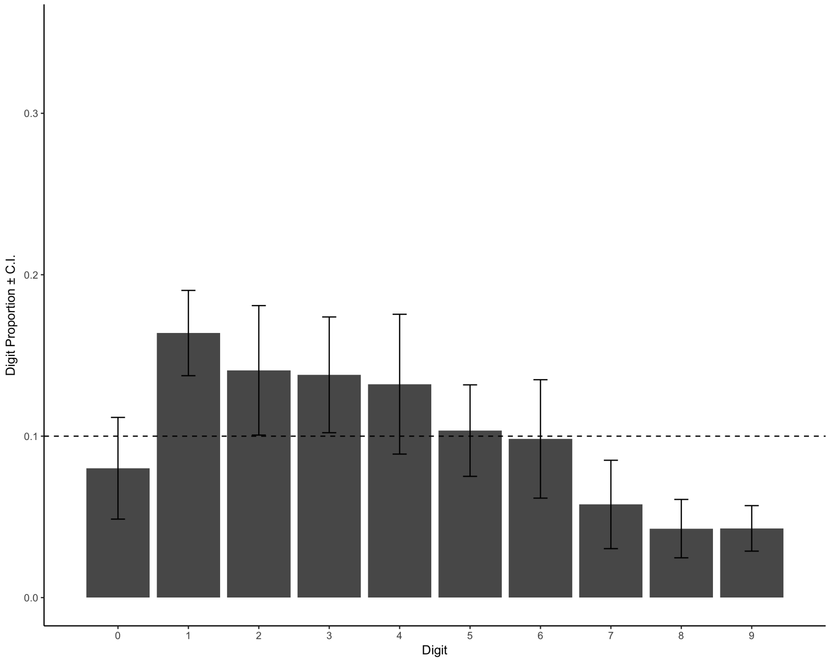
$$MAD = \frac{\sum_{d=1}^{K} |AP_d - EP_d|}{K} * 10^2 \tag{2.2}$$

where K is defined as the number of leading digit bins (9 for first leading digit; 90 for first two leading digits), AP is the actual proportion observed, and EP is the expected proportion (.11). Moreover, while MAD scores do not have an analytically-derived critical value, we can use repeated measures across time to detect significant changes in the degree of distortions from a “clean” distribution over time. Our hypothesis is a "clean" period of reporting will have a consistent degree of distortion over time. In districts with anomalous reporting, we should expect to see the MAD scores increase over the period of intervention.

2.2.1.1 Data

For this study, I used observed crime incident (also known as crime report) data file for the city of Los Angeles [3] which was acquired from the city's open data portal.

Figure 2.1: Cumulative Digit Distribution for Reported Property Crimes, Foothill District

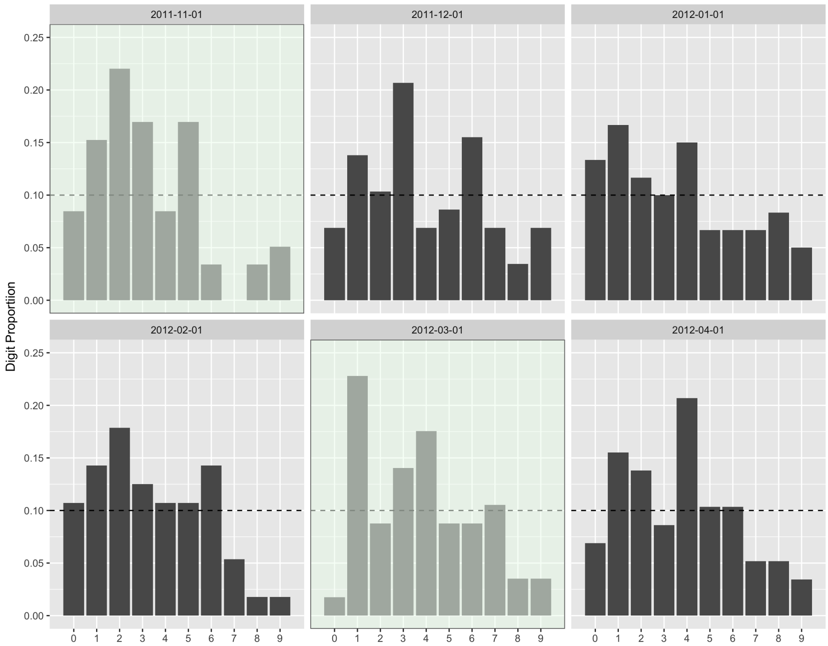


In terms of the data, the raw crime incident data file consisted of 1,684,782 observations with a unique record id, location information for the reported crime (either with lat/long coordinates or street address), date and time the crime occurred, and other relevant metadata such as the race and gender of the victim, the specific uniform crime report label, and whether an arrest was made for the crime.

In ease the computation and interpretation of the results, the hundreds of various criminal violations will be simplified into four mutually exclusive categories: quality of life crimes (i.e. drug possession violations, curfew violations, obscenity, public intoxication, etc.), property crimes (theft, larceny, robbery), and violent crimes (i.e. assault, battery, murder), sex and domestic & sex

[3] The Specific dataset I used was the Los Angeles Police Department Crime Incident Data from January 1, 2010 to March 2018 and was downloaded on March 8th 2018. The URL to the dataset is: https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq

Figure 2.2: Monthly Digit Distribution for Reported Property Crimes, Foothill District



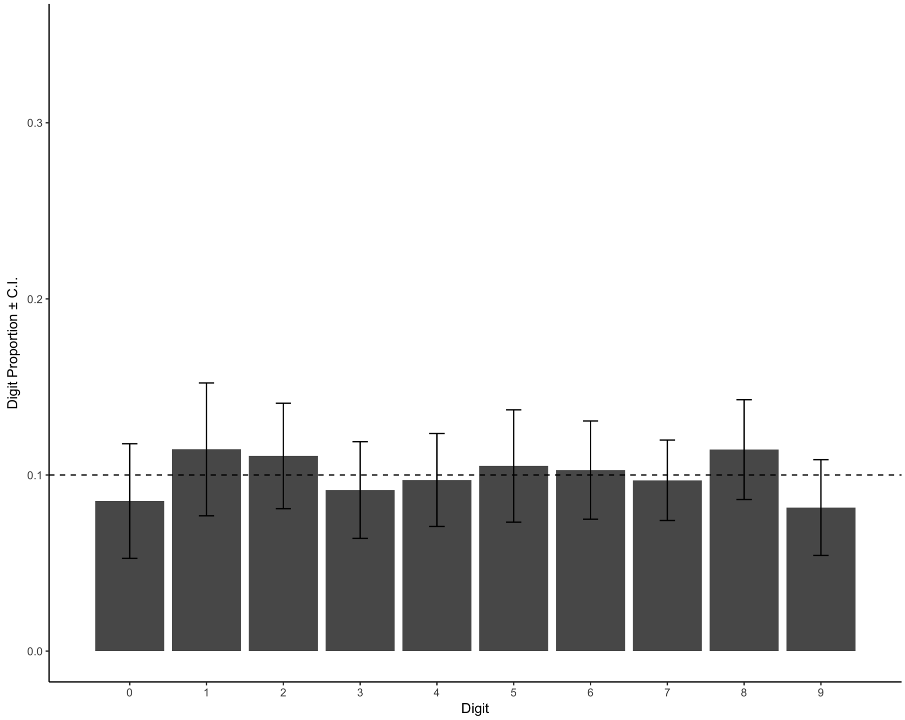
crimes (rape, sexual assault, etc). In addition to the classification by crime type, I also classified crimes as an "index crime" or "non-index crime" to denote whether the reported incident was considered a part I or part II crime based on the FBI's uniform crime reporting guidelines. After classifying each crime, I aggregate annual counts each crime type and UCR classification for each of the LAPD's 1,135 reporting districts, which are the administrative units that form the basis of crime reporting, organizing crimes into specific categories for reporting purposes.

2.2.2 Results

The analysis began by examining plots of the digit distribution of reported property crimes across each of the three deployed districts and performed a chi-square test to asses whether the deviations from the uniform distributions are statistically significant. Figure 2.1 plots the cumulative digit distribution across the entire intervention period (November 2011 - April 2012) for the Foothill District, where the X-axis represents the single digit categories and the Y-axis is the proportion of the reporting district recording the specific digit.

The dashed line represents the ideal equifrequent proportion of the digits (.10) for comparison.

Figure 2.3: Cumulative Digit Distribution for Reported Property Crimes, Southwest District



Interestingly, as predicted in the Beber & Scacco (2012) and Pope & Simonsohn (2011), the anomalous digits in the Foothill district appear to center on low number digits (1,2,3) rather than higher digits such as 8 or 9 as they comprise over 50% of the cumulative distribution during the period.

Figure 2.2 breaks down the digit distribution for each month of the deployment period. For this barplot, the X-axis again represents the single digit categories, and the Y-axis is the proportion of the reporting district recording the specific digit. Months highlighted light green are months which have distributions with chi-square values above the critical value at the 95% confidence level.

A quick visual inspection of the data shows that every year in the study period has multiple digits that are significantly above the uniform threshold, particularly low-digit values (1,2,3,4). However, despite the consistent presence of digits over ideal distribution, only two months November 2011 ($\tilde{\chi}^2$ = 28.62, p-value = .0007) and March 2012 ($\tilde{\chi}^2$ = 22.47, p-value = .0007) were found to be statistically significant deviations of null hypothesis of equifrequent digit distributions.

Moving to the Southwest district, we see a slightly different outcome. Figure 2.3 is the

Figure 2.4: Annual Digit Distribution for Reported Property Crimes, Southwest District

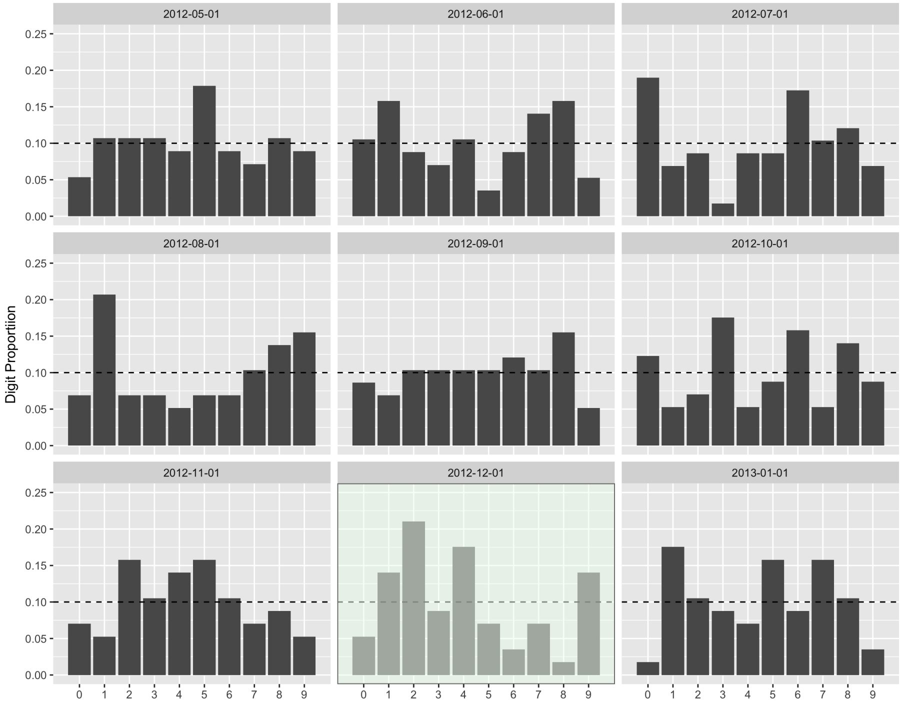
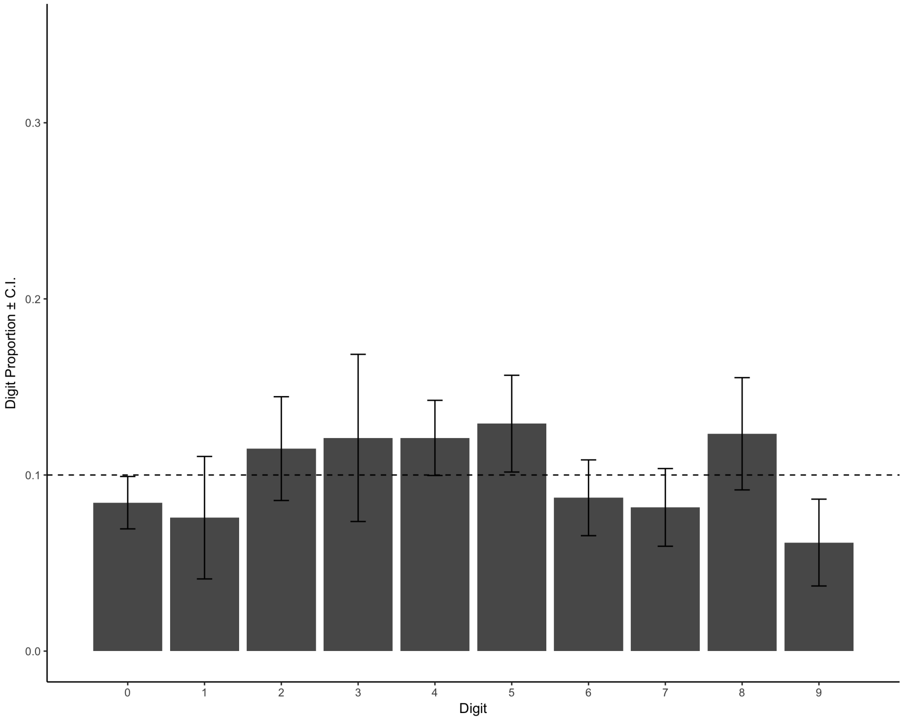
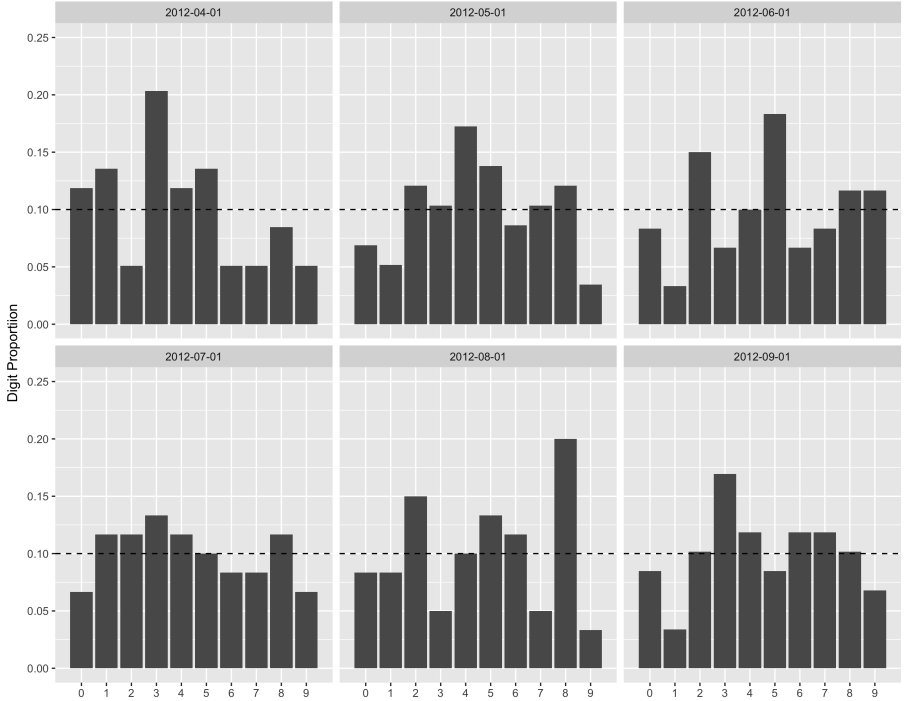


Figure 2.5: Cumulative Digit Distribution for Reported Property Crimes, North Hollywood District



cumulative digit plot for the district during the deployment period, and unlike the Foothill district, the Southwest district has a reasonably uniform digit distribution during the period, as digits with values above the ideal distribution are within the 95% confidence interval.

Figure 2.6: Monthly Digit Distribution for Reported Property Crimes, North Hollywood District



In the monthly breakdown, there is much more variation in the digit distributions than the cumulative totals would suggest. Similar to the foothill district, we consistently see low digit values with outsized proportions of the distribution. What appears to differ from the foothill district is the presence of higher digit values (7,8,9) with higher proportions in multiple months. Despite these deviations of both ends, there is only one month – December 2012 – that has a statistically significant deviation ($\tilde{\chi}^2$ = 20.72, p-value = .014) from the ideal distribution.

Finally, Figures 2.5 and 2.6 break down the results of the North Hollywood District. Comparable to the Southwest district and has a cumulative distribution that is relatively uniform in distribution. Also, with the expectation of the 5 digit, it appears that cumulative distribution is within the 95% confidence intervals.

In the monthly breakdown, there again is more variation in the distributions, as we see both low and high-value digits with large proportions of the distributions during the deployment period. However, unlike the Southwest and Foothill districts, North Hollywood had no districts with anomalous distributions during the deployment period.

2.2.2.1 Feedback Effects

In addition to assessing the prevalence of fraudulent crime reporting, the second question in this analysis was whether the deployment of Predpol would lead to further distortions in the digit distribution over time. One approach to test this hypothesis would be to examine the change in the digit frequencies over time. To examine this question, I ran a segmented regression analysis on the monthly MAD scores for each district, and using the methodology proposed by Wagner et al. (2002), I include a dummy variable for the deployment period, a time variable which counts the periods in the time series, and a post-intervention variable which counts the number months after the conclusion of the deployment.

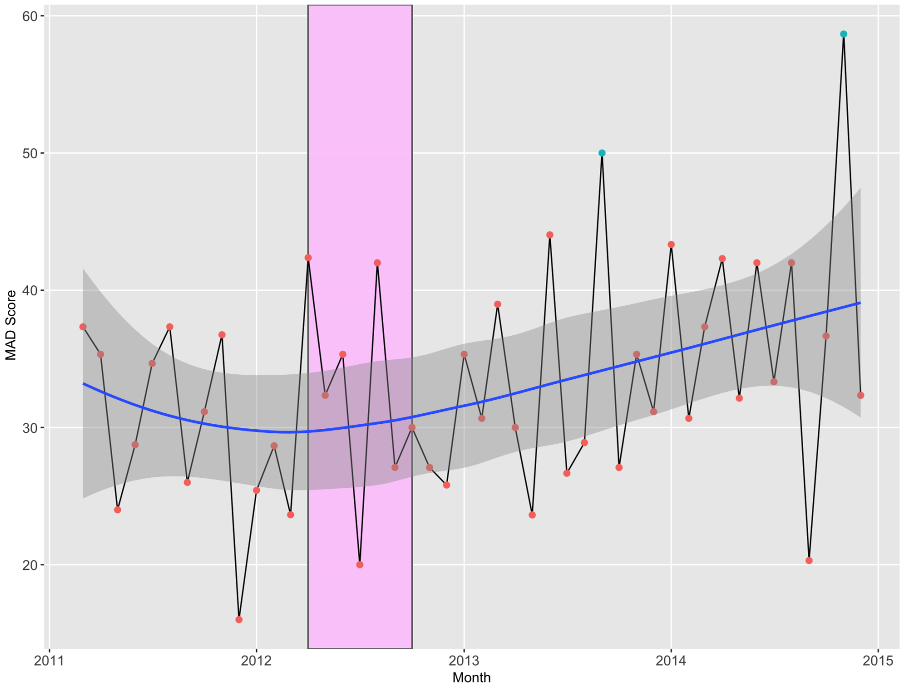
Table 2.1: Segmented Regression Table of MAD Values on Predictive Policing Deployment

	Monthly MAD Score		
	Foothill	Southwest	North Hollywood
Predpol Deployment	4.002	2.521	3.560
	(4.522)	(4.040)	(4.326)
Baseline Trend	0.167	0.107	−0.199
	(0.196)	(0.147)	(0.154)
Post-Intervention	0.039	0.195	0.587^{*}
	(0.324)	(0.363)	(0.309)
Constant	38.467^{***}	31.600^{***}	35.690^{***}
	(3.623)	(3.179)	(3.185)
Observations	59	59	59
R^2	0.120	0.111	0.076
Adjusted R^2	0.072	0.063	0.026
Residual Std. Error (df = 55)	9.231	9.146	8.744
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 2.1 shows the results of the models. For each of the police districts, the Predpol Deployment appears to have increased the monthly MAD scores, with the Foothill district experiencing

the largest effect. However, the effects were not statistically significant at the 95% level. While the initial intervention failed to influence the trend, it is possible the impact occurs after the deployment period. The post-intervention captures this potential impact, and across all three districts, the trend is positive, with North Hollywood having the largest impact and statistically significant at the .05 level.

Figure 2.7: Monthly Time Series of MAD Scores, North Hollywood District



To see this change visually, figure 2.7 plots a monthly time series of MAD scores for the North Hollywood district, where the pink shaded region denotes the initial deployment period and the blue dots represent months were the district had anomalous reporting above the .05 level. As the smooth blue line shows, the initial trend for the district was downward until the Predpol deployment which began an upward trend that continued undistributed through 2015. In addition to the upward trend, the district experienced two periods of anomalous reporting in 2013 and 2014 after Predpol's deployment concluded in 2012. Overall, that despite the lack of gaming during the deployment period, Predpol's deployment ushered in an upward trend in the anomalous reporting in the North Hollywood district that was not found in the Foothill and Southwest district.

2.3 Conclusion

Overall, this analysis found evidence the Foothill and Southwest district experienced anomalous reporting during their deployment periods. Moreover, the North Hollywood District experienced a post-deployment increase in anomalous reporting and deviations in reporting patterns. The finding collectively gives us partial evidence to reject the null hypothesis that the deployment of policing technology does not alter the measurement of reported crime. However, there are some definite limitations to this study. Perhaps the most critical limitation is the measures used in the analysis. While digit based methods for detecting fraud are standard, they are only designed to capture a specific type of anomalous activity (Rozenas 2017).

It could also be possible that gaming could be more widespread or less prevalent that what was found in this analysis if the focus was on examining the proportion of index to non-index crimes reported in a given district rather than a digit based method. Moreover, the analysis also limited in that it only focused on a singular type of crime (property) because that was the intended target of Predpol's initial deployment (Mohler et al. 2015, Brantingham et al. 2018). Investigative reporting by the Los Angeles Times (Poston & Rubin 2014, Poston et al. 2015) and a report from the LAPD Inspector General's office (Bustamante 2015) suggest that most of the gaming by LAPD officers occurred by downgrading violent crimes such as Aggravated Assault to misdemeanors, so they were classified as part II crimes.

Looking ahead there are unresolved research questions in a wide range of areas such as measurement and machine-human interactions. In addition to the measurement and crime type questions mentioned above, it is also important to consider the impact on underrepresented groups which were not discussed in this study. This is a bit surprising because if we assume that officers are intentionally suppressing crimes in specific neighborhoods, algorithms such as PredPol's ETAS could potentially underestimate the risk that a given neighborhood poses. And, if this misreporting had a disproportionate focus on particular neighborhoods (i.e., more affluent), then it could be another pathway by which institutional norms shape and potentially bias input data used in machine learning tools.

Lastly, it will also be important to continue to explore further how ADS and other forms of artificial intelligence are impacted when dealing with humans and human institutions. Much of the literature – rightfully – to date centers around the potential harms of automated tools and their likelihood of perpetuating existing institutional prejudices. However, if we begin to look forward to widespread adoption of "fair" ADS, these new tools still don’t address the very important questions about their impact on human behavior and incentive structures. There is currently a dearth of research on the potential costs and benefits of large-scale human and AI interactions that will be the underpinning for one the most significant revolutions in human history (Daugherty & Wilson 2018). Hopefully, this study will encourage more expansive and creative endeavors to explore these critical questions.

CHAPTER 3

BACK TO THE FUTURE: PREDICTIVE POLICING AND RESIDUAL DISCRIMINATION FROM RACE-NEUTRAL CLASSIFIERS

> *Every technology is an expression of human will. Through our tools, we seek to expand our power and control over our circumstances—over nature, over time and distance, over one another."*
>
> — Nicholas Carr*, The Shallows: What the Internet Is Doing to Our Brains*

In recent years, the rapid expansion and heralding of algorithmic decision systems (ADS) in criminal justice and policing as a panacea for decades of discriminatory policing have given rise to a public backlash over concerns over biased input data and its potential for disparate impacts for people of color. Much of this backlash has been animated by pioneering interdisciplinary research into areas such as pre-trial risk assessment tools (Angwin et al. 2016), predictive policing (Lum & Isaac 2016, O'Neil 2016), and facial recognition (Buolamwini & Gebru 2018). These initial findings have prompted government agencies to begin to take these concerns seriously.

In a recent report on algorithmic systems and civil rights (Munoz et al. 2016), the Obama White House confirmed the potential for harms and rights violations to underrepresented groups through predictive algorithms could stem from both algorithm design and the underlying input data which powers the model. In particular, they found that poor quality input data can lead to "skewed algorithmic systems that effectively encode discrimination" and a "feedback loop [that] causes bias in inputs or results of the past to replicate itself in the outputs of an algorithmic system" (pg.7-8). A report by the United States Federal Trade Commission (2016) also noted that data, "inaccuracies and biases might lead to detrimental effects for low-income and underserved populations" (pg.9), such as exposure of their sensitive information for ad targeting or reinforcing existing biases by excluding them from employment or housing opportunities.

In response, some scholars (Ferguson 2017, Miller 2018) and vendors (Smith IV 2016, Wood 2018) have suggested that even if the potential for input data bias exists there is scant empirical evidence to corroborate these concerns, and the potential benefits of these tools outweigh the purported risks. Others scholars (Selbst & Barocas 2017, Citron & Pasquale 2014) have also lamented the difficulty in identifying harms associated with algorithms. However, they argue the lack of evidence is due to the "black box" design of many predictive models, which limit access to the underlying source code or data to assess the tools rigorously.

Given the scant empirical evidence and mounting public concerns, there is a clear need to understand the connection between algorithms and institutional behavior better. At present government agencies and software vendors still consider their predictive algorithms in a vacuum, independent of their decision-making and behavior. However, this paper will argue that the deployment of ADS create what engineering researchers refer to as cyber-physical-human systems (CPHS), which consist of interconnected systems (computers, cyber-physical devices, and people) "talking" to each other across space and time (Schirner et al. 2013, Holzinger 2016, Sowe et al. 2017). Thinking of algorithmic systems as a CPHS is an essential change in framing because it means that human and institutional behavior can *be* influenced or augmented (i.e., intelligence augmentation or amplification (Engelbart 2001, Skagestad 1993, Carter & Nielsen 2017, Matarić 2017, Niforatos et al. 2017, Schmidt 2017)) by biased input data rather than merely be the provider of it. Further, it means that rather than merely targeting a model's inputs and output, we can utilize a broader pool of available data to examine the potential for disparate impacts or other forms of algorithmic bias.

In this spirit, the subsequent sections will seek to provide a broader historical and theoretical connection of how technology can impact human decision-making and cognition, with a focus on current applications of ADS in high-stakes policy settings. The paper will proceed as follows. Section two will briefly review current research and literature on the impacts of technology and cognition and decision-making. Section three will focus on linkages between institutional norms in policing and technology. The final section will examine a case study of the deployment of predictive policing in Los Angeles, California and examine its impact on officer decision making.

3.1 This is Your Brain on Data

At the beginning of his book *The Signal and the Noise*, Silver (2015) discusses the impact of arguably the single most significant advancement of intelligence augmentation in human history: Johannes Gutenberg's invention of the printing press in 1440. As the author is quick to point out, Gutenberg did not invent the book, but instead fundamentally altered the economics of knowledge. Much like cloud computing has done for digital records in modern times, the printing press made it possible to inexpensively (book prices dropped between 80% and 99% after the press was introduced) and permanently (previous books were hand-written meaning books often prone to damage or decay) accumulate knowledge at previously unfathomable levels. In fact, the printing press increased the number of books produced by roughly 30 times the previous rate during the first century of its existence and continued exponentially in the following centuries.

Moreover, while the early years of the printing press provided humans global scale access to higher quality maps as well as the works of Shakespeare and Galileo, it also spurred many of the same struggles modern technology is presenting us today. Misinformation was rampant during the early days of the printing press, as typos and spelling errors (such as the *wicked* Bible that stated that thou *shalt* commit adultery) and an overwhelming surge of new and conflicting ideas produced mass confusion and social unrest. In reflecting on the surge of information in the aftermath of the printing press, Silver notes,

> "The amount of information was increasing much more rapidly than our understanding of what to do with it, or our ability to differentiate the useful information from the mistruths. Paradoxically, the result of having so much more shared knowledge was increasing isolations along national and religious lines. The instinctual shortcut that we take when we have 'too much information' is to engage with it selectively, picking out the parts we like and ignoring the remainder, making allies with those who have made the same choices and enemies of the rest (pg.3)."

In short, while the printing press did significantly augment human intelligence by allowing us to preserve and archive information inexpensively, its emergence also triggered very primitive and instinctual cognitive tendencies to triage new information to bits which validate our preferred version of reality and to discard or discount the remainder. Half a millennia later, these questions and concerns have emerged again when considering how the rapid adoption of algorithmic decision systems designed to augment human perception and behavior will alter humanity and our institutions.

The idea of using technology to augment human abilities has been a consistent part of our evolution as a species. Carr (2011) outlines four categories of technology/human augmentations: 1) innovations such as the plow and the steam engine which extend our physical strength, dexterity, or resilience; 2) technologies such the microscope and Geiger counter which aim to extend the range or sensitivity of our senses; 3) genetically modified plants and other tools to reshape the natural world to better serve humanity's needs; and 4) "intellectual technologies" such as the maps and clocks, which aim to "extend or support our mental powers—to find and classify information, to formulate and articulate ideas, to share know-how and knowledge, to take measurements and perform calculations, to expand the capacity of our memory." This fourth group is the focus of the present study.

As Schmidt (2017) notes, the idea of using technology to augment intelligence has ebbed and flowed, with the industrial revolution primarily focused on using technologies, such as the steam engine and hydraulic lift, to replace or augment humans' physical strength and stamina and moved away from technologies which center around intelligence. However, the beginning of the 20th century saw a demand for new tools and mechanisms to further enhance memory and intelligence as innovations such as broadcast radio, television, and terrestrial telephones exponentially increased the amount of information available to the world.

In response, early scholars such as Bush (1945) wrote about the idea of a "memex," a portmanteau of memory and index which

> "stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate

> supplement to his memory. It [memex] consists of a desk, and while it can presumably be operated from a distance, it is primarily the piece of furniture at which he works. On the top are slanting translucent screens, on which material can be projected for convenient reading. There is a keyboard, and sets of buttons and levers. Otherwise it looks like an ordinary desk."

While the memex seemed like science fiction in 1945, it represented a desire for tools to store and organize with the deluge of information emanating from the variety of new and sometimes conflicting sources. Later technologies such as the personal computer and later the Internet realized this concept of dramatically augmenting human intelligence by providing a convenient and always-on platform to organize vast amounts of information. While this may have effectively reduced the cost of extending our collective memory to near zero (Carr 2011), it has also created some unexpected externalities.

3.1.1 Principles of Modern Technology and Human Behavior

In particular, this section will focus on two specific types of influences that the Internet and Internet-enabled platforms (i.e., social media or search engines) have on human cognitive functioning and behavior. Specifically, the ability of engagement with Internet platforms to alter our behavior in the real or physical world and to create a digital incentive system. Perhaps the most documented study of online/offline real-world behavioral transference is the voter mobilization experiment by Bond et al. (2012) on the social media site (SNS) Facebook during the 2012 Presidential General Election.

The study randomly assigned 60 million Facebook users over the age of 18 into one of three treatment groups: 1) a "social message” group who received a reminder to vote, polling information, as well as a random selection of profile pictures from the user’s network who had already indicated they voted, 2) an "information message” group who were shown the reminder to vote and the polling information but not the assortment of profile pictures from their network, or 3) a control group who received no indicators whatsoever. The authors found a 2.08% increase in users who reported voting in the social messages group compared to the information message group (as measured by clicking on the ’I voted’ button required.) Jones et al. (2017) replicated the 2012 study and extended it by assessing turnout using validated voter turnout data from users in 13 states with publicly available voter files. In the study, the authors confirmed the positive influence in voter turnout, albeit at a much smaller percentage (.24% vs. 2.08%).

Social networking sites also can impact real-world behaviors outside of the domain of political mobilization. For example, Althoff et al. (2017) examine the introduction of an online SNS to a smartphone application (Argus) designed to track the daily physical activity of its users. To measure the SNS impact, the authors used a matched sample (6,076 matched pairs from 211,383 users on the network) and measured the difference in average daily steps 20 weeks after the intervention. The results suggest the SNS introduction led to an initial 7% (406) increase in daily steps which sustained for a three month period. However, by 20 weeks the effect appears to diminish to the point where the treatment and control group activity are indistinguishable.

In a separate domain and using a different methodology, researchers have qualitatively explored how SNSs have also played a critical role in the growth of terror groups such as Islamic State (IS). Gates & Podder (2015, pg. 109) notes the uniqueness of the media infrastructure developed by IS to recruit new members,

> "IS has developed an effective virtual propaganda machinery. Its media arm Al Hayat has been releasing videos showing different sides of the militant group. On the one hand is its face of cold terror such as of children holding decapitated heads; on the other are more Western friendly videos of IS militants posing with Nutella jars to demonstrate familiarity with Western lifestyles. While the online propaganda is increasingly important, offline traditional recruitment methods such as writing letters to prisoners and organizing at or around mosques are also being used, often hand in hand with social media campaigns.”

Thus, with a far more insidious and disturbing endgame, terror groups seem to operate similarly to political parties and silicon valley start-ups: converting online engagement into specific off-line behaviors. Also, while the evidence suggests they have been successful in doing so, questions remain about the magnitude and duration of these effects.[1]

The second key influence of the Internet and Internet platforms are their ability to create digital incentive systems which alter our behavior. As Ward (2013*b*) notes, Internet platforms "provide a high speed system for delivering responses and rewards–'positive reinforcements,' in psychological terms – which encourage the repetition of physical and mental actions." The bing of a smartphone notification indicating a new email, social media message, the near instantaneous reaction to our search queries when conducting a Google search, or the ease of the one-click transaction on Amazon alters the neurological pathways in our brain.

However, how exactly does something like the Internet alter our brains? Perhaps the first study to focus on the impact of activities on the Internet on our cognitive abilities is the work of Small et al. (2009). In the study, the authors assigned a sample of 24 subjects, aged 55-78, to either search the Internet for a specific topic or read a book while a magnetic resonance imaging (MRI) is being performed. The authors then assessed the cognitive difference between subjects categorized as "Net Naive" (minimal self-reported search engine experience) and "Net Savvy" (higher levels of self-reported prior experience with computers and the Internet). For the text reading task, both Net Naive and Net Savvy participants appeared to have activation in the same cluster of brain regions (hippocampus, visual cortex, posterior cingulate).

On the Internet search task, the Net Savvy Group had greater than twofold spatial activation (more areas of the brain engaged). Particularly engaged regions of the brain were ones which control complex reasoning and decision making. For the Net Naive group, in addition to having fewer regions activated, the Internet search tasks failed to activate the regions related to memory (posterior cingulate and hippocampus). The findings suggest that with repeated experience conducting searches through the Internet, the brain changes by engaging more areas related to critical thinking and decision making, likely to parse through the deluge of information generated by a Google search query. Given the limitations of this study (small, non-representative sample, the emergence

[1] For example, other scholars Ferrara (2015), Zekulin (2018) argue that the impact of these recruitment techniques can be mitigated if counter-narratives are promoted on the platform.

of smartphones) it is crucial to review subsequent research to see if these findings hold.

Sparrow et al. (2011) follow up on this question by examining "cognitive offloading” – the use of physical action to alter the information processing requirements of a task to reduce cognitive demand (Risko & Gilbert 2016) – exhibited after prolonged use of Internet search engines. The authors’ paper, which consists of four separate studies present a series of new insights. First, the authors used a modified Stroop task (a color naming task with words presented in either blue or red ink) to measure reaction times on paired subjects and found that computer terms were more accessible (quicker reaction time on Stroop) than general words after participants encountered questions where they were uncertain of the answer. Next, subjects were tasked with memorizing and recalling a series of trivia facts and told they would either be able to refer to hand-typed notes in a computer or that they would have their notes erased. The findings suggest that when told a computer would not be an available resource, subjects had higher information recall. Interestingly, when the experiment was modified to ask subjects the location of their typed information (into one of six randomly named folders), participants who were told they would receive their notes were less able to recall facts but better able to identify if and where the information could be accessed.

While the findings from Sparrow et al. (2011) provide substantial evidence of cognitive offloading, the author does not provide details regarding why and how this phenomenon occurs. Ward (2013*b*) sought to fill this void by introducing two important theories. First, the author introduces the concept of transactive memory (TM) and transactive memory systems (TMS) (Wegner 1987). TM theory was proposed initially by psychologists to understand how intimate partners establish cognitive divisions of labor to solve information problems but were eventually generalized to explore the knowledge organization processes for any group or organization (Peltokorpi 2008). TMS, then, is the term often used to refer to these collective systems described within the context of TM theory. In both TM and TMS, the cognitive division of labor falls into two components. First, is *internal* memory or knowledge that a person encodes and stores for their retrieval.

Alternatively, *external* memory is information encoded and stored within other members of the defined group which can be located and retrieved from other external storage devices (Peltokorpi

2008, Brandon & Hollingshead 2004). A critical component of TMS is that in order to maximize the collective knowledge of the group, members must identify and assign an area of expertise for each member. As Ward (2013*b*, pg. 342) points out, "Group members intuitively offload responsibility for information to those individuals with the highest levels of relative expertise and/or access to information in a relative domain and assume responsibility for the domains in which they are experts and/or insiders." If properly utilized, TMS allow individual members to efficiently acquire increased depth of knowledge in their domains of expertise while simultaneously having access to the board range of information held by the other expert members of the group (Wegner 1987).

However, the introduction of the Internet has likely fundamentally altered our social cognition and how we utilize the resources within TMS. First, scholars have argued that the Internet can be viewed as a unique member of a TMS (Loh & Kanai 2016, Ward 2013*b*, Sparrow & Chatman 2013). As Ward (2013*b*, pg. 343) notes, "the Internet may be more than just another memory partner; it may be treated as an informational catch-all, reducing the amount of information stored both in other external sources (e.g., human transactive memory partners) and internally (i.e., in an individuals' memories)." Thus, the author defines the Internet as a "supernormal stimuli," one which hijacks the cognitive process underlying the formation of the TMS because it is a novel stimulus that outperforms all naturally occurring stimuli in domains related to selection. However, because the Internet is omnipresent and possesses an unrivaled scale of domain expertise, members within a TMS will not just intuitively defer to the Internet as the expert in most domains but may not even be aware the process is happening. This deference is due in part to the unobtrusive nature of the Internet as a memory partner, and in the era of the smartphone, is so ubiquitous and tightly integrated into the user interfaces that it is virtually invisible to most users.

This unique dynamic between technology use and human cognition does not just end with information selection. Recent research is beginning to explore further the range of impacts that the emergence of the Internet (as well as smartphone use of the web) has on social cognition. One area of concern consistently identified by researchers has been that the repeated usage of the Internet as a memory partner can lead to user overestimating their intelligence because of an inability to distinguish between the limits of one's intelligence and the Internet (i.e., source memory). The concern about overconfidence was evident even within traditional TMS, as Wegner (1987, pg. 198) notes, "[i]ndividuals can suffer from transactive memory, though, even when it is well established and running smoothly. This happens when they over-estimate its capabilities. Just as a person's metamemory offers information about what the person knows, a transactive memory system provides individuals information about what knowledge they may access in the group. This may result in a brand of the "feeling of knowing" that yields overconfidence in one's own ability to access knowledge."

This "feeling of knowing" appears to be even more exaggerated when the Internet serves as the memory partner. For example, Fisher et al. (2015) conducted a series of nine experiments which sought to test how the use of Internet search impacted users self-assessed levels of knowledge. In most of the trials, users were asked to confirm the details of a subset of basic explanatory questions such as "How does a zipper work?" by using a search engine and supplying a URL that was most helpful in confirming the question. However, the "no Internet" group was merely asked to rate their ability to answer the questions without using "outside sources." After the search task, subjects were asked to rate their ability to answer questions on a series of unrelated areas. In five of the seven search related trials, users who were provided access to the Internet rated their self-knowledge higher in all six of the unrelated domains.

In another study, Ward (2013*a*) conducted a series of six experiments where subjects were asked to complete a trivia quiz either with or without the help of the Internet. Following this short quiz, they were asked to complete a scale assessing Cognitive Self-Esteem (CSE)— which measures one's self-perceived ability to remember and process information— and also predict how well they would do on another quiz of similar difficulty, completed without any external resources (including the Internet). The author found that participants who used the Internet to complete the first quiz reported higher levels of CSE than those who had not, suggesting that they were more likely to view attributes associated with the Internet as being self-descriptive.

Another critical area where the Internet has an impact on social cognition is in evidence evaluation. In their study on the impact of technology on evidence evaluation, Guadagno et al. (2013) found that presenting information on a computer screen engendered more favorable evaluations of the material at hand (a football scout's evaluation of a potential recruit) and that this effect was more pronounced for expert scouts. Sparrow & Chatman (2013) found that the sense of control over how information is retrieved induced greater skepticism immediately after information was retrieved, but that after multiple repeated trials this effect was flipped, with reduced skepticism and increasing belief. These findings led authors to suggest that " a dynamic of epistemic evaluation where the choice that one has when performing searches may result in greater skepticism initially but serve as a cue of credibility when an individual is fatigued after many trials (pg.352)."

Overall, it appears that technologies such as the Internet have a bit of a what Mäntymäki & Islam (2016) refer to as the Janus effect - after the Roman god of transitions, beginnings, and doorways - in that technology engenders both positive and negative influences simultaneously. It is clear that the Internet is changing our behavior in the real world and even altering our cognition and construction of reality. Moreover, while technologies do present an ability to efficiently enhance the range of our indirect expertise in a range of domains, they also create a false sense of overconfidence in our level of knowledge and impact our ability to evaluate information critically. These impacts do raise serious questions, as will be discussed in upcoming sections of this paper, about the expanded use of algorithmic decision systems (and artificial intelligence broadly speaking) which require humans to quickly interpret digital information and prediction and perform decision-making in the real world.

3.1.2 Technology-Behavior Interaction in Policing

Advocates of "data-driven" or "evidence-based" policy approaches argue that the emerging predictive systems will allow governments and businesses to augment institutional behavior and targeting of resources using objective data, thereby better serving underrepresented groups and helping to overcome historical inequalities (Podesta et al. 2014, Chowdhry et al. 2016, Goel 2017, Miller 2018). For example, Goel et al. (2016) propose an algorithm with the goal of potentially reducing racial bias in New York City’s stop-and-frisk policy by creating variable precinct thresholds that would prohibit an officer from detaining a suspect unless he posed a sufficient public threat. In the United Kingdom, Oswald et al. (2017) has sought to generate risk scores to identify individuals to avoid future arrests (and subsequent court hearings) and, instead, be diverted into a "checkpoint” program which aims to reduce future re-offending.

However, in the age of artificial intelligence, the extensive use of data to supplant officers’ intuition in hopes of racial-neutral policing may backfire. As noted in previous sections of this paper, the heavy dependency of historical input data required for machine learning algorithms and the consistent tendency of humans to use technology as an epistemological memory partner and "cognitively off-load” task-related knowledge could potentially lead to construction of “perceived” realities of the groups represented by data (Hersh 2015, pg. 28-33) which can reinforce and amplify disparate impacts embedded within the data.

Much like the Internet, artificial intelligence (AI) and ADS can be viewed as a superstimuli, a tool that humans will rely on regardless of their alternative sources of information expertise (Logg 2017). Place-based predictive policing systems, such as Predpol, are explicitly designed to yield intense focus on a narrow group of blocks or neighborhoods within an assigned region, requiring officers to conduct extended periods of patrols in the aim of deterring crime. As an epistemological partner, extended exposure would in theory yield a higher concentration of crimes and enforcement actions being conducted in these areas as officers begin to off-load their cognitive intuition as to the location of crimes to the AI tool.

In addition to altering the focus of institutional actors, AI’s ability to reduce the cost of uncertainty could lead to altered decision-making in areas unrelated to the input data. Logg (2017) found that algorithms fail to attenuate a person’s confidence in their predictions, and in some cases may increase it. Overconfidence in human behavior is not just observed in psychology or neuroscience (Griffin et al. 1992, Moore & Healy 2008, Johnson & Fowler 2011), but could serve as an explanation for a wide range of irrational actions in political behavior (Wilson 2011, Ortoleva & Snowberg 2015). On the contrary, Proeger & Meub (2014) suggests that overconfidence is strategic and evolutionary. The authors argue that confidence can be a means of a gaining an advantage over your peers by appearing more competent, and can be viewed as favorable in social settings if it leads to net benefits for the individual and population at the aggregate level. This advantage perhaps may partially explain the rush to acquire and adopt AI tools - it can instantly signal more competency than your peers and may lead to positive net benefits at the individual (promotion, financial reward) or organizational (more resources and accolades) level.

In the policing context, this overconfidence can be exhibited in the behavior of officers in designated hotspots. In his now seminal book *Street-level Bureaucracy*, Lipsky (2010) argues that contrary to the popular "top-down” views of policy implementation, everyday government employees such as teachers, police officers, and social workers serve as important policymakers. The author argues that these street-level bureaucrats (SLBs) serve as influential policymakers because, "they have considerable discretion in determining the nature, amount, and quality of benefits and sanctions provided by their agencies (pg.13).” The reason that street-level bureaucrats (SLBs) have such a significant degree of autonomy is because they are often placed in direct contact with citizens with complex issues that need to be resolved expediently with limited resources and knowledge (Evans & Harris 2004). However, this level of autonomy gives SLBs a great deal of influence in the policy implementation process, such as shaping the social construction of particular groups of clients, the ability to raise the salience of particular policies to clients (Tummers & Bekkers 2013), or varying the level of responsiveness to client demands.

However, the discretion provided to officers also could lead to distortions of service delivery and amplification of disparate impact. For example, McEvers (2016) documented this wide variability of policing practices on Skid Row – an infamous homeless camp in Los Angeles, California – depending on which officers were assigned to the day and night shifts on a given day. One officer who grew up near skid row and lived through bouts of homelessness early in his life, viewed his role as an officer to primarily maintain stability and order within the camp and facilitate access to public services, rather than detain or cite the residents with citations.

During the night, a second officer was a firm believer in the "broken windows” policing philosophy (Kelling & Wilson 1982) and took a much more aggressive stance on his patrol through strict enforcement of so-called nuisance laws, such public intoxication, and loitering. Both officers used their broad discretion as SLBs to approach their jobs in the manner they believed helps the community and their clients, but from a data perspective, it tells diametrically opposing views about the level of crime in this particular area.

Predictive policing is likely to exacerbate this type of behavior because at their core the current era of ADS can be best characterized as tools to reduce the cost of prediction (Agrawal et al. 2018). These "prediction machines" aim to provide you with useful information (predict your estimated time of arrival or detect fraudulent credit card transactions) from input data that would be missing from human insight alone. However, if the input data is encoded biased or unrepresentative, it could also inaccurately validate negative stereotypes of neighborhoods or groups. Because the potential amplification of discriminatory policing due to predictive policing is problematic, many stakeholders have proposed inserting a "human in the loop" when cases of biased data or discriminatory outcomes have emerged when deploying ADS within the criminal justice system and elsewhere (Brynjolfsson & Mitchell 2017, Sowe et al. 2017, World Economic Forum 2018).

However, this approach fails to question whether humans with repeated exposure to the insight generated by the tool could disentangle insight generated by the model (and its embedded biases) from insight generated by the person (in theory to counter the biased predictions). In practice, situating a human in the loop may lead to an amplification of bias data because of the inability of humans to separate sources of information when a superstimulus such as an AI tool serves as an epistemological partner. The upcoming section will attempt to address these questions and concerns about the interaction between people and data. In particular, the focus will be the application of ADS in policing and their potential real-world impacts on officer behavior in the field.

3.2 Case Study: LAPD Deployment of Predpol

In the waning years of the Chief Bill Bratton's tenure of the LAPD, Bratton began to be convinced that the Compstat era of policing he began in the 1990's with the New York Police

Department (NYPD) (Buntin 2009, Hayden 2013) was ending and that a new form of data-driven policing would emerge: Predictive Policing. Growing out the community policing demands for more proactive (rather than reactive) forms of policing, Bratton was again an early adopter of this new form of data-centric policing, and in 2007 applied for a National Institute of Justice grant that would help his precinct implement a model of forecasting crime used initially by the military to track insurgents in Iraq (Bond-Graham & Winston 2013). And, while Bratton would resign before receiving it, the grant funding would enable the LAPD (and its residents) to become the beta-testers for what would become one of the largest predictive policing companies in the country: PredPol (Cantu 2014). Predpol's origins date back not to law enforcement, but rather to the Pentagon. In the midst of the Iraq war, the military awarded a series of grants [2] to the Pure and Applied Mathematics Department at the University of California, Los Angeles (UCLA) to develop mathematical algorithms for use on the battlefield [3]. The researchers, anthropology professor Jeffery Brantingham, math professor Andrea Bertozzi, and math postdoc George Mohler specifically sought to re-purpose the spatio-temporal branching or a "self-exciting" Poisson process, commonly referred to as a Hawkes Process based on the seminal research by Hawkes (1971) into using seismographic activity to predict earthquake aftershocks, to accurately forecast battlefield causalities and insurgent activities.

After the conclusion of his postdoc, Mohler took a position near Silicon Valley at the University of Santa Clara. It was in Santa Clara where he met a trio of political lobbyists (Ryan Coonerty, Caleb Baskin, and Zach Friend) who convinced Mohler and Brantingham to convert their academic research into a private company in January 2012 (Bond-Graham & Winston 2013). After receiving of $1.3 million from investors, they sought to find a venue to demonstrate the viability of their tool

[2] Based on the historical database of federal grants compiled by USASpending (https://www.usaspending.gov//award/39367873) and disclosed by a UCLA Press Release (http://newsroom.ucla.edu/releases/predictive-policing-substantially-reduces-crime-in-los-angeles-during-months-long-test), Predpol's co-founders received over $2 million dollars in federal grant funding between 2008 and 2014 for Predpol related projects.

[3] As Economist Editorial Board (2014) notes, this is not unusual as statistics and mathematics have a long and connected history to war and conflict.

over conventional crime analyst tools such as Compstat. They found a willing partner in two rising stars within the LAPD: Lt. Sean Malinowski (Now Deputy Chief) and Deputy Chief Charlie Beck (now Chief of Police) (OMalley 2013, Bond-Graham & Winston 2013). Beck and Malinowski wanted to continue Bratton's mission of transforming the LAPD, which had long been resistant to outside ideas on policing, into a lab for police innovation (Hayden 2013, Cantu 2014).

Predpol teamed up with Malinowski and Beck to conduct a series of pilot programs in three LAPD police districts. According to Brantingham et al. (2018) & Mohler et al. (2015), Predpol's software was deployed in three participating LAPD divisions: Foothill (FH), North Hollywood (NH) and Southwest (SW) between the period of November 2011 and January 2013 [4]. The pilot study aimed to compare the performance of Predpol against traditional human crime analysts using Compstat (Mohler et al. 2015). To do this, the shift commander would receive a randomly assigned patrol map generated by either a human analyst or a map Predpol generated using their proprietary algorithm before the beginning of each of the three daily patrol shifts.

Each map contained twenty target areas marked as 500 x 500-foot boxes. The shift commanders were not informed of the origin of the crime maps and were only instructed to inform the officers to patrol the areas in their discretionary time (Brantingham et al. 2018) [5]. The stated focus of the pilot was on targeting burglary, car theft and burglary theft from vehicle (BTFV) because they can

[4] Because of the unique randomization method, the pilots had varying lengths of implementation. Brantingham et al. (2018) attributed the variation to the inability of a human analyst present to generate a control map for a given day. Overall, in the Foothill Division had a successful random assignment on a total of 124 test days, which was evenly divided with 62 control and 62 treatment days. The North Hollywood division had a total of 152 test days which were split into 82 control and 70 treatment days. Lastly, the Southwest Division had a total of 234 test days, which were divided evenly into 117 control and 117 treatment days.

[5] Anecdotally evidence would suggest the shift commanders had a different impression of the map's intent. For example, Capitan Jorge Rodriguez of the Foothill division described the process, "Every morning, we get a report from PredPol for which 20 boxes are going to be where crime is most likely to happen," and that, "[i]f your resources are diminished, then you want to focus on those boxes with the highest rate of crime (OMalley 2013)." Further, the districts did not always adhere to patrol instructions, in addition to just distributing the hotspot maps, the Foothill Division division, for example, deployed four officers to patrol the crime hotspots in clusters. These teams would either operate as a uniformed deployment or in plainclothes, depending on how they are being used on that particular day (Bond-Graham & Winston 2013).

account for as much as 60% of police recorded crimes in the City of Los Angeles (Brantingham et al. 2018). Overall, the Police patrols using ETAS forecasts led to an average 7.4% reduction in crime volume, while patrols based upon analyst predictions showed no significant effect on crime volume across the three test divisions. The authors also noted that while the findings are suggestive of potential crime deterrence, the evidence of a drop in crime due to ETAS was not significant (Brantingham et al. 2018).

While proponents of predictive policing have viewed this trend as a significant step towards transparency and pragmatic, data-driven policymaking, the use of predictive policing within police departments has also raised grave concerns among activists and scholars regarding this new intersection between statistical learning and public policy. Civil liberties advocates have argued the growth of predictive policing means that officers in the field are more likely to stop suspects who have yet to commit a crime under the guise of historical crime patterns that are not representative of all criminal behavior. In their excellent report on predictive policing, Robinson & Koepke (2016) point out that reported crime data are "greatly influenced by what crimes citizens choose to report, the places police are sent on patrol, and how police decide to respond to the situations they encounter." Legal scholars such as Joh (2017, pg. 3) notes that, "Police are not simply end users of big data. They generate the information that big data programs rely upon. Crime and disorder are not natural phenomena. These events have to be observed, noticed, acted upon, collected, categorized, and recorded — while other events aren't."

There is also the questions of their actual effectiveness. To date, only three empirical studies of predictive policing have been published. Saunders et al. (2016) assess the Chicago Police Department's Strategic Subject's List (SSL), a person-based predictive policing system created internally in 2013. The authors use ARIMA time series models to estimate the impacts of the deployment of the SSL in 2013 on city-level homicide trends. While the authors find a decline in city-level homicides overall, the introduction of the SSL failed to have a measurable impact. As the authors note, " the statistically significant reduction in monthly homicides predated the introduction of the SSL, and that the SSL did not cause further reduction in the average number of monthly homicides above and beyond the pre-existing trend." Hunt et al. (2014) conducted a randomized control trial on the deployment of a predictive policing system in Shreveport, Louisiana, and found was no statistically significant change in property crime in the experimental districts that applied the predictive models compared with the control districts.

Moreover, while Brantingham et al. (2018) acknowledges their initial study failed to demonstrate meaningful crime reduction, Thomas (2016) suggests the LAPD's crime statistics show other divisions that were not using Predpol saw crime reduction as high as 16 percent during the same period. Given these inconclusive peer-reviewed findings and anecdotal evidence, some vendors point to internal testing done by departments themselves as evidence of the efficacy of predictive policing. However, Robinson & Koepke (2016) note that although "system vendors often cite internally performed validation studies to demonstrate the value of their solutions, our research surfaced few rigorous analyses of predictive policing systems' claims of efficacy, accuracy, or crime reduction."

In addition to the concern regarding the input data influencing the predictions generated, there is the new concern about the reciprocal impact of the data influencing officer judgment and behavior in the field. As outlined in detail over the previous chapters, many police departments use metrics such as arrests or citations for internal promotion (Knafo 2016), and the implementation of a predictive AI tool which claims to identify areas where crimes hotspots create a perverse incentive for police officers to increase these punitive and aggressive enforcement actions. Often with disparate Impacts for communities of color. For example, Saunders et al. (2016) discovered in their assessment of the implementation of the Strategic Subjects List (SSL) for the Chicago Police Department (CPD), the individual-level ADS tool was rarely used to prevent victimization from violent crimes, instead the authors found that CPD officers often used the SSL as a way to generate leads for unsolved shooting cases.

O'Neil (2016) adds further that the introduction of predictive policing could induce a further amplification of broken windows policing strategies,

> "Raised on the orthodoxy of zero tolerance, many have little more reason to doubt the

> link between small crimes and big ones than the correlation between smoke and fire. When police in the British county of Kent tried out PredPol, in 2013, they incorporated nuisance crime data into their model. It seemed to work. They found that the PredPol squares were ten times as efficient as random patrolling and twice as precise as analysis delivered by police intelligence. And what type of crimes did the model best predict? Nuisance crimes. This makes all the sense in the world. A drunk will pee on the same wall, day in and day out, and a junkie will stretch out on the same park bench, while a car thief or a burglar will move about, working hard to anticipate the movements of police. Even as police chiefs stress the battle against violent crime, it would take remarkable restraint not to let loads of nuisance data flow into their predictive models. More data, it's easy to believe, is better data. While a model focusing only on violent crimes might produce a sparse constellation on the screen, the inclusion of nuisance data would create a fuller and more vivid portrait of lawlessness in the city. And in most jurisdictions, sadly, such a crime map would track poverty. The high number of arrests in those areas would do nothing but confirm the broadly shared thesis of society's middle and upper classes: that poor people are responsible for their shortcomings and commit most of a city's crimes (pg. 89)."

Other significant concern is the impact of the technology of the officers themselves. In 2016, the city of Burbank, CA surveyed rank-and-file officers and found 75% indicated that morale at the department as low or extremely low, and one of the primary reasons for their dissatisfactions was the recent introduction of the predictive policing firm Predpol in 2014 (Tchekmedyian 2016*a*,*b*). In the years since it's adoption officers felt it often directed them to obvious areas such as sports stadiums or shopping centers or locations such as a nearby police station (if there was a surge in reports at the station) which often a inefficient use of officers time as they were required to spend extended time (45 minutes) in each of the three assigned patrol boxes. The overall effect of was one of confusion as patrol officers struggle to reconcile the conflict between their intuition and the machine-generated predictions. Regardless of the rank-and-file officers' objections, there are clear signals from department leadership that predictive policing is here to stay. Burbank Police Chief Scott LaChasse said the department would continue to deploy Predpol because, "the future in law enforcement is not going to be random patrol, it's going to be predictive analytics."

This anecdotal evidence also reflects the empirical evidence comparing the differences between human and machine perceptions of crime and how the interaction can alter patterns of individual behavior. Ratcliffe & McCullagh (2001) compared how accurately 65 Nottinghamshire police officers could predict the crime "hotspots" generated by a GIS algorithm for three types of property crimes (residential burglary, non-residential burglary, and vehicle theft). Overall, while the police officers for residential burglary coincided with the hotspot analysis over 60 percent of the time in each of the police subdivisions, but the correlation between hotspots and officer perceptions were on average 20% lower across the divisions for vehicle theft and non-residential burglary. Overall, the authors note that "the point remains that in a majority of cases the perception of operational police differed significantly from the computerized hotspot generation process." However, a critical unanswered question from this study is when predictive policing tools such as Predpol move into patrol cars and smartphones, can the human in the loop be able to effectively triage the variety of information flows to administer police procedures in a fair and impartial manner?

It could also be possible that the combination of poor input data and misaligned incentive structures could amplify the historically bad practices as officers utilize the data and predictions to validate further their subjective determinations about neighborhoods they patrol and their use of more punitive police actions. Again, evidence from the partnership between the LAPD and Predpol seem to suggest this dynamic may occur. For example, Deputy Chief Malinowski found that after the implementation of Predpol in 2012, officers noted they were "hyper-alert" when deployed to one of the designated hotspots. Also, Predpol further designed the smartphone and tablet systems to track their movement during their patrols, with the bins changing color when the risk of crime increased and more police presence was needed in a particular area. These efforts to rank and focus police attention is intentional, as co-founder Brantingham believes that Police officers operate in a fashion similar to hunters (OMalley 2013).

This idea was at the center of Brantingham et al. (2018) (BVM), which sought to examine whether predictive policing lead to biased patterns of arrests in the field. Their study outlined three null hypotheses: (1) arrest of minority individuals did not differ between control and treatment conditions in test divisions; (2) arrest rates overall did not differ between control and treatment conditions in test divisions; (3) the rate of arrests per crime was unchanged across treatment and control conditions. Of these three, Hypothesis 2 stands out as needing further scrutiny as their implications for substantiating the claims of many activists and scholars who have suggested predictive may induce racial biases (O'Neil 2016, Lum & Isaac 2016).

The authors' evidence for rejecting the null hypothesis (2) was presented in table 1 & 3 in the Table and Figures section. Table 1 breaks down the total number of arrests between treatment and control days in each of the three LAPD police districts by race. While Table 3 which provides a breakdown of overall arrests for each of the three police districts by race for the treatment and control days. In table 3, arrests increase for eight of the nine treatment cells, with Black arrests in the Southwest district (122% increase) having the most significant percentage increases.

Table 1 was slightly different in that arrests on treatment days (overall not just within the predicted boxes) seem to fall for Blacks and Latinos in two out of three districts. Despite these considerable differences, the Cochran-Mantel-Haenszel test – which tests for significant differences across strata (i.e., race groups)– failed to reject the null hypothesis under both instances. These findings led the authors to conclude, "the current study is only able to ascertain that arrest rates for black, and Latino individuals were not impacted, positively or negatively, by using predictive policing. Future research could seek to test whether the situational conditions surrounding arrests and final dispositions differ in the presence of predictive policing (pg.9)."

If the findings held, it would suggest the successful deployment of a race-neutral classifier in a high stakes setting. However, Brantingham et al. (2018) suffered from significant methodological shortcomings. First, the ETAS algorithm Predpol used during the trial was designed to target three types of crimes: burglary, car theft and burglary theft from vehicle (BTFV). However, in their analysis, the authors arbitrarily choose to use an aggregate measure of arrests. The choice of the measure would be defensible if the bulk of the arrests consisted of crimes related to the input data selected by Predpol, but they presented no evidence this was the case. Second, the authors chose to use raw counts of arrests for each racial group, which could substantively alter the findings if there large discrepancies between the demographic compositions of the districts. Lastly, multiple studies from statistics and psychology (Rozeboom 1960, Gill 1999, Nickerson 2000, Aczel et al. 2017) that failing to reject a null hypothesis is not the same as proving the alternative to be true as other factors could play a role. For example, a null hypothesis could also be the result of an insufficient sample size, poor experimental controls, or the presence of confounding variables.

These concerns merit a more thorough investigation of the case. Two publicly available datasets were used to conduct this analysis: 1) Los Angeles Police Department Crime Incident Data from January 1, 2010 to March 2018 and 2) Los Angeles Police Department Arrest Data from January 1, 2010 to March 2018. After acquiring each file, custom R scripts were created to geolocate the individual records to a specific U.S. census jurisdiction using the private subscription service Askgeo [6]. If a record failed to include a set of coordinates, a second R script accessing the Google maps API [7] was used to acquire the coordinates from the street address provided. In addition to appending the census geography, a series of demographic and economic variables such as the proportion no minority residents, per capita income, and educational attainment at the census block group level.

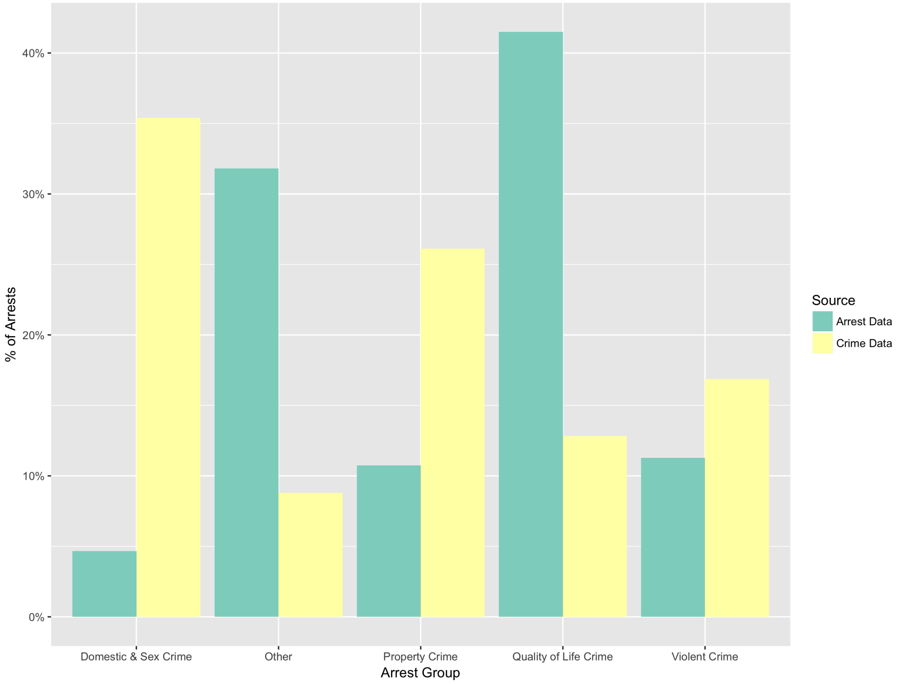
The analysis began by examining the distribution of arrests for the three (Foothill, Southwest, and North Hollywood) LAPD patrol districts mentioned in BVM during the stated deployment periods. For comparison, the distribution of arrests collected from the crimes database is also included. Figure 3.1 displays the breakdown by arrest charge type for the Foothill district. At first glance, it is clear the two datasets, collected by the same police department during the same period, have very different interpretations of crime patterns. For the arrests dataset, quality of life or "broken windows" arrests –including charges ranging from loitering, drug possession, to illegal

[6] Details on the Askgeo's API Geolocation service can be found at https://askgeo.com

[7] Google Maps API was accessed through the R package GGmaps. Details of the package can be found at the following URL: https://github.com/dkahle/ggmap

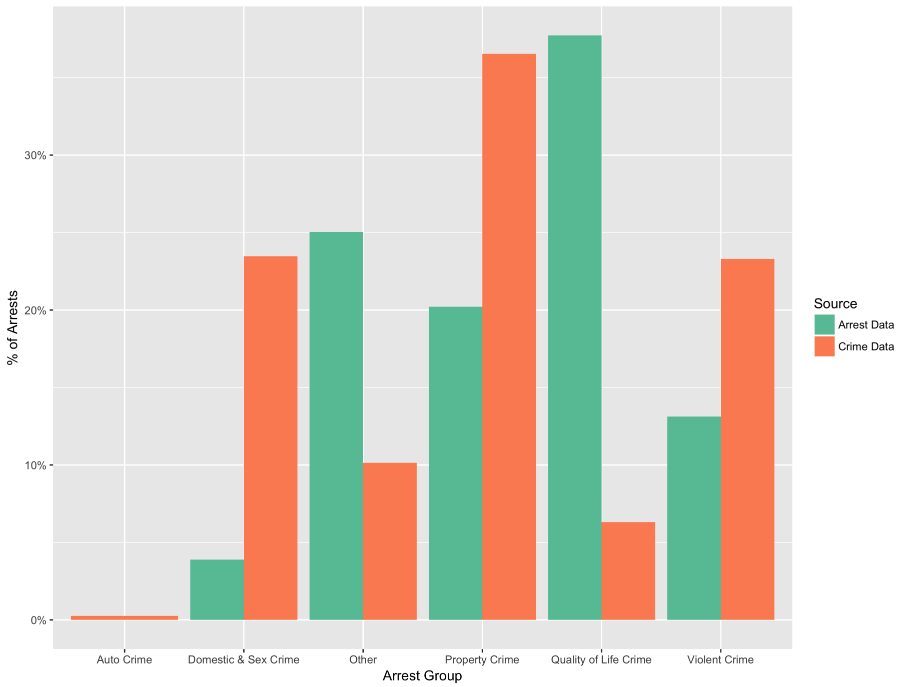
ownership of a shopping cart – comprise over 41% of the arrests during the deployment period compared to only 12% when evaluating the crimes dataset. Property crimes are the opposite as they comprise over 25% of arrests made in the crimes database but only 10.7% in the arrest database.

Figure 3.1: Percentage of Arrests by Crime Type and Data Source, Foothill District 2010-2017



Moving to figures 3.2 and 3.3, quality of life arrests are again the majority of arrests for both the Southwest (36.5%) and North Hollywood (34.9%) districts in the arrests database. Property crimes represented 36.5 and 33.8% of the arrests made, respectively, in the crimes database. Miscellaneous crimes range from undocumented arrests to trespassing represented the one of the highest category of arrests at 31%, 25%, and 42.67% respectively. The large differences between datasets do raise some interesting substantive questions. First, as noted earlier, police departments use these datasets in different ways. The crimes database is often the dataset sent to the FBI as part of the Uniform Crime Reporting process, and the metric that is primarily used to measure crime nationally, the key input data for hotspot tools such as Predpol or Compstat. By utilizing these data as the consensus measure of crime, it perhaps alters the department's perception of both the level and composition of crime in a given jurisdiction. These findings also suggest that the significant number of quality of life arrests in each of the districts during their deployment periods suggest a continuation of the "broken windows" model is indeed a prominent feature in the Predpol era. Finally, the evidence above would suggest that the bulk of arrests BVM are using as their key metric consist of crimes unrelated to the crimes used in their input data of the ETAS algorithm. In addition to concerns

Figure 3.2: Percentage of Arrests by Crime Type and Data Source, Southwest District 2010-2017



about the distribution of the arrest charges, there is a question of whether using raw counts of arrests is appropriate for measuring the impact of predictive policing on arrest patterns. Much like the concerns raised by (Brantingham et al. 2018) regarding the patterns of higher crimes serving as a confounder in the change in cumulative arrests, using raw counts may not be an appropriate measure if there is a large degree of variation between group sizes within the patrol districts. To assess the possibility of disparities in population between racial groups, the population of each group within the three LAPD patrol districts in BVM was calculated. This was achieved by intersecting GIS spatial map files (or shapefiles) of LAPD districts with census block groups to identify the unique census block group IDs within each of the respective districts. Next, using the geolocation scripts mentioned in the previous section, the 2010 census population counts for each racial group as well as the total population were collected and aggregated by census block to generate the district level population counts.

Figure 3.3: Percentage of Arrests by Crime Type and Data Source, North Hollywood District 2010-2017

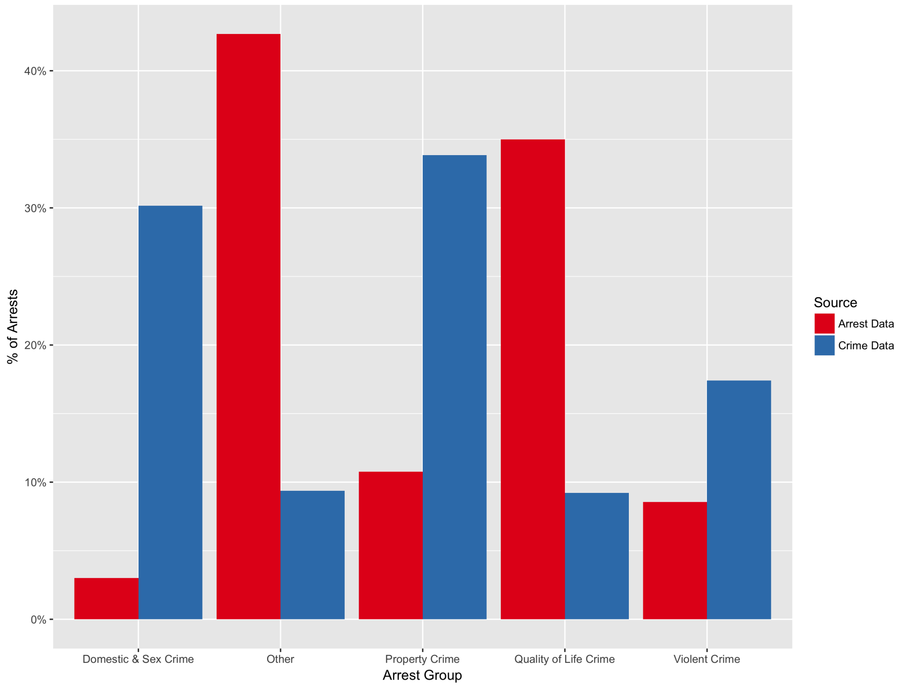


Figure 3.4 below outlines the demographic proportion of each patrol district by racial group. As expected, Latino residents represent the largest demographic category in each of the three districts, with 63% (Foothill), 40% (North Hollywood), and 51% (Southwest) of the populations respectively.

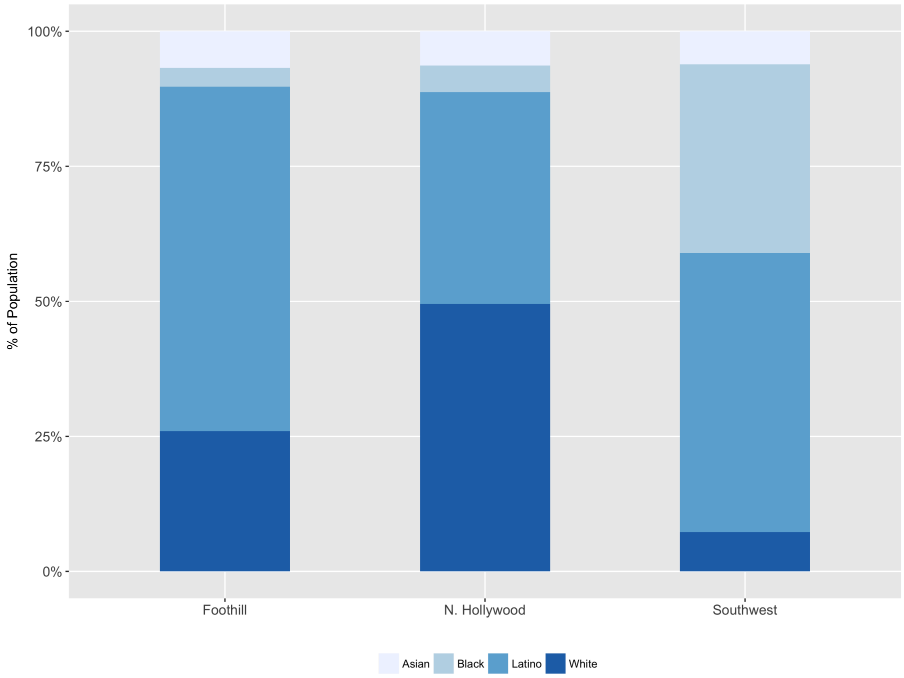
Black Residents have the most variable population, comprising only 3.4% and 4.9% population in the Foothill and North Hollywood districts but over 35% in the Southwest district. Much like Black residents, White residents also vary in population between the districts, ranging from 49% of the population in North Hollywood to 7.3% in the Southwest district. This considerable heterogeneity in the district populations suggests that the inferences of CMH need to be re-visited as the assumption that the arrest count of each racial group being equal within the districts in invalid. To create a more fair comparison, we need to calculate arrest rates estimates which adjust for differences in population counts (Humphreys et al. 2017). Arrest rates are generated by using the arrest counts within the predicted areas provided by the authors with equation 3.1, where index i represents the racial group and index j the deployment district.

$$arrests_{percapita} = (arrests_{ij}/population_{ij}) * 10^5 \tag{3.1}$$

Figure 3.5 below plots the per capita arrests by racial group and district within the predicted boxes. In figure 3.5, the red bars denote the per capita arrests on control days or days where LAPD patrols used maps provided by human analysts. The blue bars denote treatment days where patrols received maps provide by Predpol's ETAS algorithm. Once adjusting the arrests for the population, it is clear that for most groups Predpol's ETAS algorithm increases the number of arrests for each racial group, which could be explained by the authors' suggestion their algorithm was more efficient in identifying hotspot areas. However, that fails to explain the distribution of these increase across racial groups. Black residents in each of the three districts saw a dramatic increase in arrests, with the Foothill district seeing a 3-fold increase in arrests and the southwest district experiencing a 122% increase in arrests per capita on treatment days. Moreover, while the Whites and Latinos experience sizable gains in arrests (White per capita arrests increased by 50% in North Hollywood), it is important to denote that Black arrests have a much higher baseline. Moreover, assessing the table with the Cochran-Mantel-Haenszel (CMH) test with the null hypothesis of no difference in racial group arrest rates between treatment groups is rejected (M^2 = 12.66, df = 2, p-value = 0.001).

While the CMH test focuses on across strata differences, further testing is needed to identify

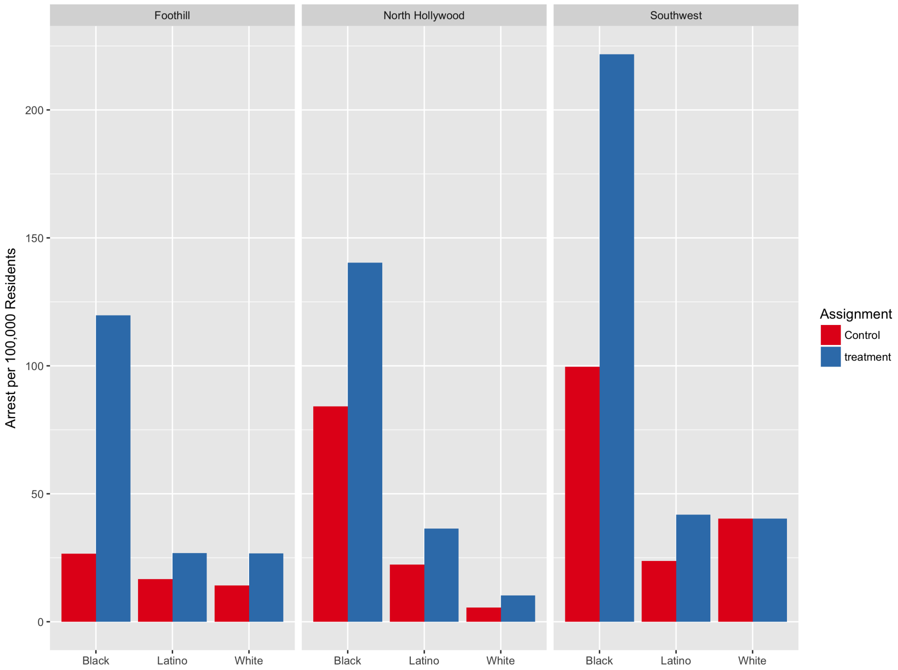
Figure 3.4: Demographic Composition of LAPD Districts



significant differences within each strata (racial group) (McDonald 2015). This was achieved by performing a Chi-squared test on each strata within the table. The differences in treatment and control group for Black arrests rates are significant ($\tilde{\chi}^2$ = 15.73, df = 2, p-value = 0.0003), but the Latino ($\tilde{\chi}^2$ = 0.063, df = 2, p-value = 0.968), and white arrests rates ($\tilde{\chi}^2$ = 3.07, df = 2, p-value = 0.214) groups were not significant. These findings suggest that contrary to the findings in BVM, the deployment of Predpol may actually have amplified Black arrests in all three of the LAPD districts. Further, the increases for white and Latino arrests while substantial were not statistically distinguishable from zero.

One potential confounder in the findings could be that the differences in arrest proportions are reflective of biased policing patterns overall in the district, rather than the impact of the algorithm

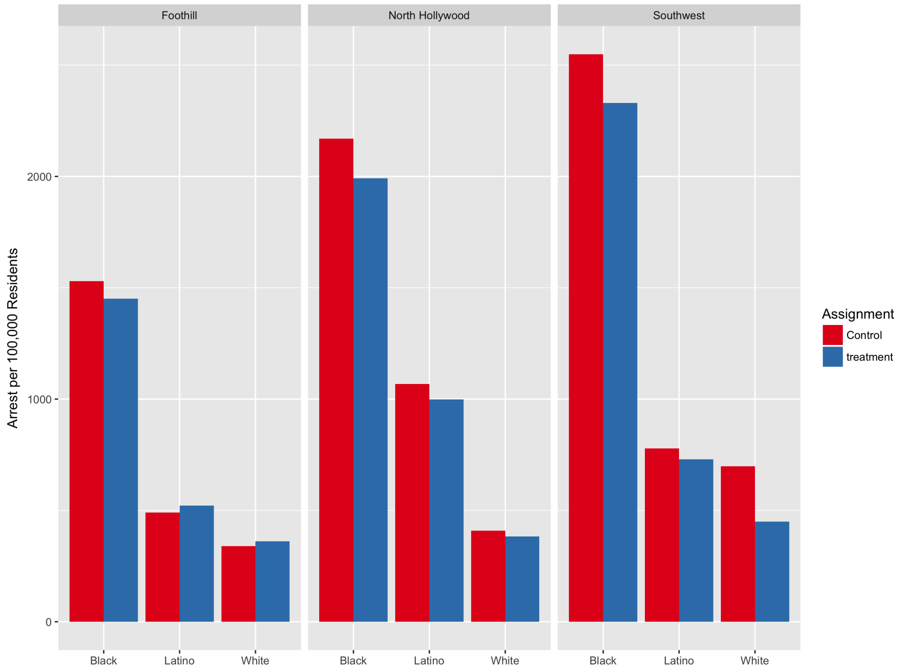
Figure 3.5: Arrests Rates in Control and Treatment Boxes for Predpol Deployed LAPD Divisions



per se. To examine this counter-factual, we can assess the arrest rates by race for areas not within the predicted areas [8]. Figure 3.6 plots the arrest rates per 100,000 residents for each LAPD district, with the blue bars representing the control deployment days for each district and the red bars denoting the arrest rate for treatment deployment days. In total, trends appear to be different than within the predicted boxes. Comparable to arrest rates within the predicted hotspots, the rate at which Blacks are arrested are dramatically higher than for other racial groups. In the case of the Foothill district arrest rate is 2-fold higher than the rate for whites and Latinos. Despite the higher baseline for Blacks the difference between control and treatment days are minimal. One interesting point from this plot is white arrest rates, wherein districts such as Southwest arrests declined by over 35%. Overall, the arrest rate declines during the treatment days in all three districts. Surprisingly, when

[8] These values were calculated by taking the difference of the cumulative number of arrests for treatment and control days and the total treatment and control arrest during the respective deployment period

Figure 3.6: Arrests Rates Non-Predicted Boxes in Predpol Deployed LAPD Divisions



testing the differences across strata, the CMH test is statistically significant (M^2 = 11.16, df = 2, p-value = 0.004). On the Chi-squared test on each stratum (race) within the table, the results again differ from the outcomes for the predictive boxes. The differences in treatment and control group for white arrests rates are significant ($\tilde{\chi}^2$ = 31.28, df = 2, p-value = 1.609e-07), but the Latino ($\tilde{\chi}^2$ = 3.140, df = 2, p-value = 0.208) and Black arrests rates ($\tilde{\chi}^2$ = 0.667, df = 2, p-value = 0.716) were not significant. In total, the two tables suggest that Predpol's deployment had the dual effect of increasing Black arrest rates in the predicted areas and declining white arrests in the areas outside of the predicted boxes.

But, this analysis has some limitations. First and perhaps most important are that the tables used in this analysis are only cross-sectional sums and fail to adjust for the temporal dynamics which always underlie the data generation process and may lead to spurious claims of disparate impact. Second, one of the common theories for explaining or justifying discriminatory policing is the high correlation between reported crimes and arrests. It could also be plausible that the differences in arrests could be explained by a greater emphasis on high crime neighborhoods. The next section will present an original analysis which seeks to address these theoretical and methodological concerns.

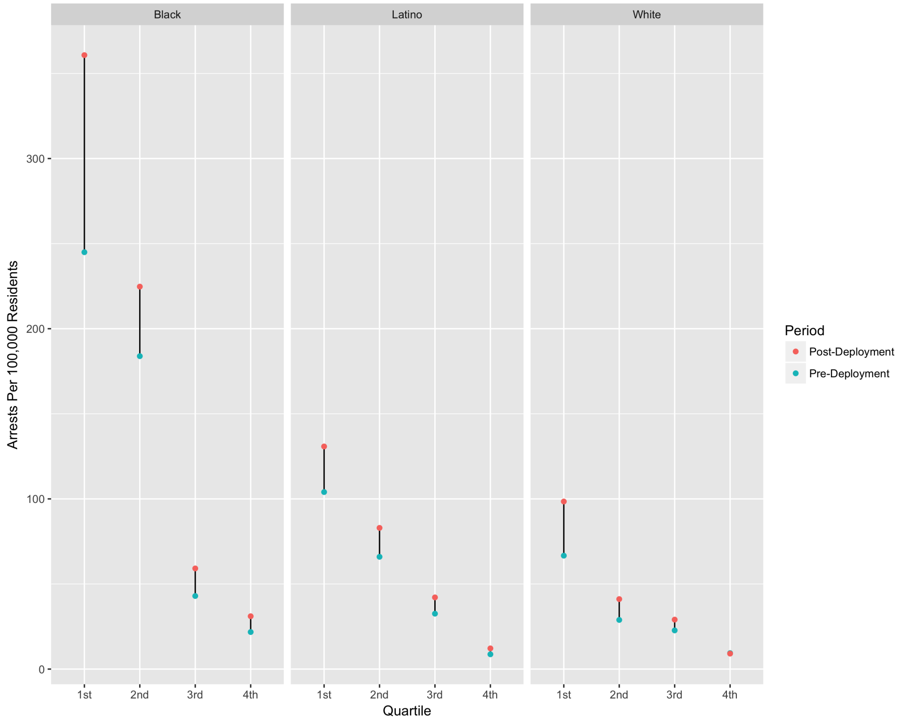
To address the concerns of differences in crime reporting serving as a potential confounder, a separate analysis which combines the modified crime and arrests datasets used in the previous section. Also, the analysis sought to determine if the biases embedded within historical crime data translate into a disparate impact on arrest rates in the future. The process started by subsetting the crimes database to crimes occurring at the beginning of the dataset (December 31, 2009) to the month before the district Predpol deployment [9]. Next, the census block groups are divided into four quartiles based on the distribution of report crimes occurring during the period. Finally, for each race discussed in Brantingham et al. (2018) (Black, Latino, White) the monthly sum of arrests are counted. To adjust for the potential dynamics of the time series counts, seasonally adjusted counts are generated using a loess decomposition method (Cleveland et al. 1990). The census estimates for each racial group are then used to create the population adjusted arrest rates used in the previous analysis.

The seasonally-adjusted arrest rates for each race now allow us to make a more direct comparison of the impact of Predpol's deployment. Figure 3.7 provides the pre and post average monthly arrest rate for each of the four crime quartiles. For example, within these series of figures, quartile 1 signifies the arrest rate for census block groups with the number of crimes in the 75th to 100th percentile. Each panel denotes the average monthly arrest rate for each specific racial group. Moving from the leftmost panel to the right, we can see the pre-post Black arrest rates for the Foothill district, and it seems that the census block groups in quartile one experience a 47.5% arrest rate increase after the deployment of Predpol, from a monthly average of 427 arrests per 100,000 residents to 644 arrests. This sharp increase in the arrest rate is highly concentrated as the other quartiles fail to experience the same uptick. This same pattern of large spikes in the historically high reported crime census block groups was also experienced to a lesser extent in the Latino (25% increase in Q1). White residents seem to have a higher percentage increase (48.4%) but with a much lower baseline. What is also striking is that for quartiles 1-3, the baseline ("pre")

[9] For the foothill district this was October 2010, Southwest district was April 2012, and North Hollywood was February 2012

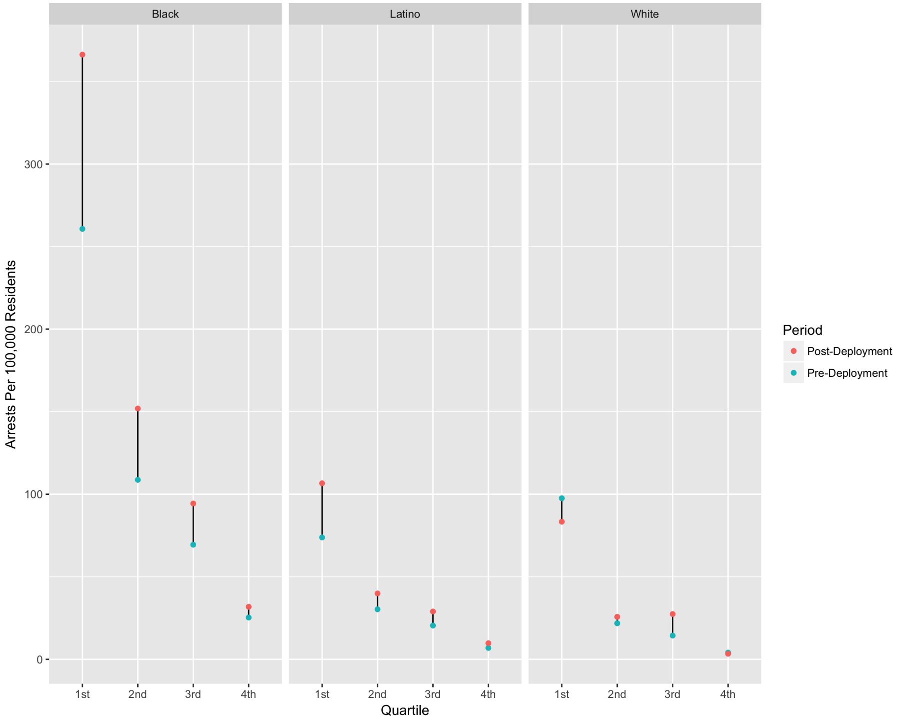
arrest rates for Blacks are higher than for both Latinos and whites. In fairness to Predpol, this elevate baseline perhaps indicate institutional biases towards Black people rather than an artifact of Predpol's deployment. Figure 3.8 plots the pre-post quartile plot for the Southwest district.

Figure 3.7: Pre-Post Monthly Arrest Rates by Race and Quadrant, Foothill District



The Southwest district appears to differ from the Foothill district for how Predpol's deployment amplified (or regressed) LAPD's arrest patterns. First, the most striking aspect is that contrary to the other Black and Latino arrests, white arrests declined after the deployment of Predpdol in Quartile 1 and Quartile 4. This compares to a 38% (Black) arrest rate and 45% (Latino) arrest rate increase in census blocks at the top quartile. Second, unlike the Foothill district, Blacks and Latino arrest rates in every quartile experience at least a 20% increase during the Predpol deployment, and the Black arrest rate seems to have an elevated baseline for all of the Much of the same trends

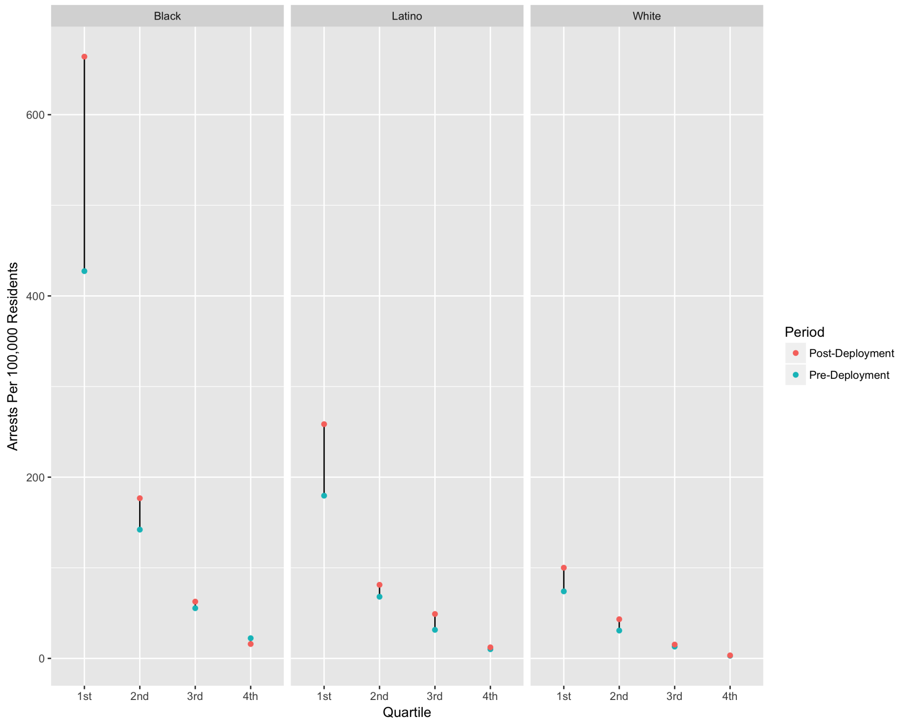
Figure 3.8: Pre-Post Monthly Arrest Rates by Race and Quadrant, Southwest District



seem to apply for Figure 3.9, which plots the pre-post arrests rates by race for the North Hollywood district. The left panel plots pre-post Black arrest rates for the North Hollywood district and again the census block groups in quartile one experience a 55.5% arrest rate increase after the deployment of Predpol, from a monthly average of 427 arrests per 100,000 residents to 644 arrests. This sharp increase in the arrest rate is highly concentrated as the other quartiles fail to experience the same uptick. This same pattern of large spikes in the historically high reported crime census block groups was also experienced to a lesser extent in the Latino (44.1% increase in Q1) and white (36.9%).

Overall, the evidence seems to confirm a few trends from the previous analysis section. First, the LAPD appears consistent in applying the "broken windows" style policing strategy of focusing on targeting high crimes areas with a high level of arrests. Moreover, it appears that Predpol seems to amplify this these patterns by increasing the arrest rate for all groups but specifically Black residents in areas they believe are "high crime." The evidence also seems to confirm previous reports (Lum & Isaac 2016, Robinson & Koepke 2016) which suggest that ETAS would merely amplify and

Figure 3.9: Pre-Post Monthly Arrest Rates by Race and Quadrant, North Hollywood District



concentrate existing biases within the input data. The findings also appear to refute the idea that patrol officers can somehow use their independently acquired knowledge to counter the embedded biases within the system. Instead, the human in the loop appears to merely use the tool to improve their bureaucratic output (arrests). In short, the findings raise serious questions of how individuals and institutions can find a balance between technology and institutional behavior.

3.3 Conclusion

The findings of this case study suggest that despite claims of race neutrality in the model, the implementation of Predictive policing in Los Angeles may have lead to a disparate impact on arrests patterns in multiple police districts by allowing biased input data to augment officer behavior. Further, as predicted by Lum & Isaac (2016), the discriminatory targeting was almost exclusively focused in areas with the highest historical rate of reported crimes, which would suggest that the officers were responding to previous patterns of historical crimes rather than new spikes in crime. Overall, these findings add to the claims raised by civil society groups and academics about the tangible real-world impacts of algorithmic decision systems in criminal justice and concur with Brantingham et al. (2018) in a call for robust auditing of all potential ADS adopted by public agencies.

What would the auditing process look like? The process to accountable and transparent use of algorithmic decision support systems must start with a full rebuke of "technological solutionism"(Thornhill 2018), or the belief that mere application of an AI tool or other technology will address critical public issues without meaningful policy reforms or institutional changes. As a recent study by Mummolo (2018) on the impact of departmental reforms on stop and frisk in New York City showed, the catalyst for changing police-citizen interactions needs to be real institutional reform rather than technology focus alone. Further, those most impacted by these potential reforms should be key stakeholders and be allowed to collaborate with police departments in discussing the potential deployment of policing technology.

Most likely, the impetus for generating institutional reform in the acquisition and deployment of policing technologies will be the passage of new regulatory guidelines that can hold officials accountable and remove complex ethical decisions out of the hands of software developers, whose interests may not always be in alignment with the needs of the communities affected. If a serious movement toward evidence-based policymaking is to take hold, it will be critical to elucidate how interventions into the data generation process (DGP) by institutional actors shape the data used for prediction and analysis of policy decisions; as well as ensuring that predictive tools do not violate the civil and human rights of historically underrepresented groups.

What would a potential regulatory framework for ADS tools look like? Shneiderman (2016) has outlined a three prong approach that could serve as a potential blueprint. In the article, the author outlines three kinds of AI oversight mechanisms, 1) a review board model where vendors or agencies should submit their tool or algorithm before any real world implementation, 2) continuous monitoring or auditing oversight reminiscent of what companies and non-profit foundations are required to do for financial due diligence, and 3) retrospective analysis of "disaster" scenarios much like the NTSB does after a plane crash by reviewing the black box data and internal governance.

Aligned with this framework Selbst & Barocas (2017) argues for legislation requiring police agencies draft “algorithmic impact statements” (AIS) modeled after the environmental impact statements of the National Environmental Policy Act (NEPA). The goal of these statements are not curtail the use of new predictive policing technologies, but rather ensure “the agency in reaching its decision, will have available, and will carefully consider, detailed information concerning significant discriminatory impacts,” as well as ensure this information is shared with the broader potential stakeholders of the respective technology’s deployment and implementation.

Perhaps anticipating these pending reforms, many agencies have sought to move away from third-party commercial vendors and opt for tools built in-house or in collaboration with universities (Hvistendahl 2016, Shapiro 2017). Newer predictive policing companies such as Civicscape have committed to algorithmic transparency by publishing a version of their source code on the online code repository Github (Brustein 2017) and pledge to not use their tools to predict drug crimes because of concerns that the bias present in crime data are too difficult to model out of their predictions (Gershgorn 2017). These efforts are certainly a laudable move toward transparency, but neither the AISs or these voluntary disclosures actually ensure institutional accountability in the deployment of ADS or other police technologies.

For example, a question that arises from the Civicscape transparency efforts is whether vendors should be responsible for defining what constitutes transparency, fairness, and oversight before policymakers set firm guidelines. Under ideal circumstances, it would be ideal for vendors or police departments to disclose their code for public scrutiny after each major release, but there is little incentive to continue their attempts at transparency as future iterations of the software are released, perhaps allowing biases to creep back in as more features or different data are included. Rather than inadequate self-regulatory efforts, rigorous independent audits of the tools and their impact are needed. These audits will require researchers with a diverse range of skills to effectively assess the broad impact of ADS. If we are successful in developing better guidelines for ADS tools, cities will implement more transparent and inclusive processes to build safer communities for all of their residents.

BIBLIOGRAPHY

BIBLIOGRAPHY

ACLU (2016), 'Statement of concern about predictive policing by aclu and 16 civil rights privacy, racial justice, and technology organizations'.
URL: *https://www.aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice*

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M. & Barnabas, S. (2017), 'Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation'.

Agrawal, A., Gans, J. & Goldfarb, A. (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business School Press.

Alexander, M. (2012), *The new Jim Crow: Mass incarceration in the age of colorblindness*, The New Press.

Althoff, T., Jindal, P. & Leskovec, J. (2017), Online Actions with Offline Impact, *in* 'the Tenth ACM International Conference', ACM Press, New York, New York, USA, pp. 537–546.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. & Mane, D. (2016), 'Concrete Problems in AI Safety'.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016), 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks', *ProPublica* .
URL: *https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*

Azar, O. H., Yosef, S. & Bar-Eli, M. (2015), 'Restaurant tipping in a field experiment: How do customers tip when they receive too much change?', *Journal of Economic Psychology* **50**, 13–21.

Bacry, E., Mastromatteo, I. & Muzy, J. F. (2015), 'Hawkes processes in finance', *Market Microstructure and Liquidity* **01**(01), 1550005.

Badal-Valero, E., Alvarez-Jareño, J. A. & Pavía, J. M. (2018), 'Combining Benford's Law and machine learning to detect money laundering. An actual Spanish court case', *Forensic Science International* **282**, 24–34.

Baraniuk, C. (2015), 'Pre-crime software recruited to track gang of thieves', *New Scientist* .
URL: *https://www.newscientist.com/article/mg22530123-600-pre-crime-software-recruited-to-track-gang-of-thieves/*

Barr, A. & Gibbs, C. (2017), 'Breaking the Cycle? Intergenerational Effects of an Anti-Poverty Program in Early Childhood', pp. 1–49.

Baumer, E. P. & Lauritsen, J. L. (2010), 'Reporting crime to the police, 1973–2005: a multivariate analysis of long-term trends in the National Crime Survey (NCS) and National Crime Victimization Survey', *Criminology* **48**(1), 131–185.

Beattie, R. H. (1941), 'The Sources of Criminal Statistics', *The Annals of the American Academy of Political and Social Science* .

Beber, B. & Scacco, A. (2012), 'What the Numbers Say: A Digit-Based Test for Election Fraud', *Political Analysis* **20**(02), 211–234.

Bevan, G. & Hood, C. (2006), 'What's Measured is What Matters: Targets and Gaming in the English Public Health Care System', *Public Administration* **84**(3), 517–538.

Boland, P. J. & Hutchinson, K. (2000), 'Student selection of random digits', *Journal of the Royal Statistical Society: Series D (The Statistician)* **49**(4), 519–529.

Bond-Graham, D. & Winston, A. (2013), 'All Tomorrow's Crimes: The Future of Policing Looks a Lot Like Good Branding', *San Fransisco Weekly* .
URL: *https://archives.sfweekly.com/sanfrancisco/all-tomorrows-crimes-the-future-of-policing-looks-a-lot-like-good-branding/Content?oid=2827968*

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E. & Fowler, J. H. (2012), 'A 61-million-person experiment in social influence and political mobilization', *Nature* **489**(7415), 295–298.

Bornstein, A. (2015), 'Institutional racism, numbers management, and zero-tolerance policing in new york city', *North American Dialogue* **18**(2), 51–62.

Braga, A. A. (2005), 'Hot spots policing and crime prevention: A systematic review of randomized controlled trials', *Journal of Experimental Criminology* **1**(3), 317–342.

Brandon, D. P. & Hollingshead, A. B. (2004), 'Transactive memory systems in organizations: Matching tasks, expertise, and people', *Management Science* **15**(6), 633–644.

Brantingham, P. J., Valasik, M. & Mohler, G. O. (2018), 'Does Predictive Policing Lead to Biased Arrests? Results from a Randomized Controlled Trial', *Statistics and Public Policy* **2**(0), 1–19.

Bronstein, N. (2015), 'Police management and quotas: Governance in the CompStat Era', *Columbia Journal of Law and Social Problems* **48**(4).

Brustein, J. (2017), 'The ex-cop at the center of controversy over crime prediction tech', *Bloomberg Businessweek* .

Brynjolfsson, E. & Mitchell, T. (2017), 'What can machine learning do? Workforce implications', *Science* **358**(6), 1530–1534.

Buntin, J. (2009), 'Did Bill Bratton Succeed in Changing LAPD's Culture?', *Governing* .

Buolamwini, J. & Gebru, T. (2018), Gender shades: Intersectional accuracy disparities in commercial gender classification, *in* 'Conference on Fairness, Accountability and Transparency', pp. 77–91.

Bush, V. (1945), 'As we may think', *The Atlantic monthly* **176**(1), 101–108.

Bushman, B. J., Newman, K., Calvert, S. L., Downey, G., Dredze, M., Gottfredson, M., Jablonski, N. G., Masten, A. S., Morrill, C., Neill, D. B., Romer, D. & Webster, D. W. (2016), 'Youth violence: What we know and what we need to know.', *American Psychologist* **71**(1), 17–39.

Bustamante, A. A. (2015), Los angeles police commission: Review of crime classification practices, Technical report.

Campbell, D. T. (1979), 'Assessing the impact of planned social change', *Evaluation and Program Planning* **2**(1), 67–90.

Cantu, A. M. (2014), 'Beyond Drones and Stop-and-Frisk', *Truth-Out* .
URL: *http://www.truth-out.org/news/item/24396-beyond-drones-and-stop-and-frisk*

Carr, N. (2011), *The Shallows: What the Internet Is Doing to Our Brains*, W. W. Norton & Company.

Carter, S. & Nielsen, M. (2017), 'Using artificial intelligence to augment human intelligence', *Distill* .

Cederman, L.-E. & Weidmann, N. B. (2017), 'Predicting armed conflict: Time to adjust our expectations?', *Science* **355**(6324), 474–476.

Chou, E. (2018), 'Valley LAPD captain sues city, alleges she was denied chances at promotion for voicing concern over crime reporting'.
URL: *https://www.dailynews.com/2018/01/19/valley-lapd-captain-sues-city-alleges-she-was-denied-promotion-for-voicing-concern-over-crime-reporting/*

Chowdhry, B., Das, S. & Hartman-Glaser, B. (2016), 'How Big Data Can Make Us Less Racist', *Zocalo Public Square* .
URL: *http://www.zocalopublicsquare.org/2016/04/28/how-big-data-can-make-us-less-racist/ideas/nexus/*

Chrystal, K. A., Mizen, P. D. & Mizen, P. (2003), 'Goodhart's law: its origins, meaning and implications for monetary policy', *Central banking, monetary theory and practice: Essays in honour of Charles Goodhart* **1**, 221–243.

Citron, D. K. & Pasquale, F. A. (2014), The Scored Society: Due Process for Automated Predictions, Technical report.

Cleveland, R. B., Cleveland, W. S., McRae, J. E. & Terpenning, I. (1990), 'Stl: A seasonal-trend decomposition', *Journal of Official Statistics* **6**(1), 3–73.

Coates, T.-N. (2015), 'The black family in the age of mass incarceration', *The Atlantic* .
URL: *https://www.theatlantic.com/magazine/archive/2015/10/the-black-family-in-the-age-of-mass-incarceration/403246/*

Connelly, R., Playford, C. J., Gayle, V. & Dibben, C. (2016), 'The role of administrative data in the big data revolution in social science research', *Social Science Research* **59**(c), 1–12.

Crawford, F. W., Weiss, R. E. & Suchard, M. A. (2015), 'Sex, lies and self-reported counts: Bayesian mixture models for heaping in longitudinal count data via birth-death processes', *The annals of applied statistics* **9**(2), 572.

Daugherty, P. R. & Wilson, H. J. (2018), *Human+ Machine: Reimagining Work in the Age of AI*, Harvard Business Press.

Desmond, M., Papachristos, A. V. & Kirk, D. S. (2016), 'Police violence and citizen crime reporting in the black community', *American Sociological Review* **81**(5), 857–876.

Du, N., Farajtabar, M., Ahmed, A., Smola, A. J. & Song, L. (2015), Dirichlet-Hawkes Processes with Applications to Clustering Continuous-Time Document Streams, *in* 'the 21th ACM SIGKDD International Conference', ACM Press, New York, New York, USA, pp. 219–228.

Dubner, S. J. & Levitt, S. D. (2012), 'The cobra effect'.
URL: *http://freakonomics.com/podcast/the-cobra-effect-a-new-freakonomics-radio-podcast/*

Durtschi, C., Hillison, W. & Pacini, C. (2004), 'The effective use of Benford's law to assist in detecting fraud in accounting data', *Journal of Forensic Accounting* **5**, 17–34.

Economist Editorial Board (2014), 'They also served: Statisticians in World War II', *The Economist* .

Edwards, M. A. & Roy, S. (2017), 'Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition', *Environmental Engineering Science* **34**(1), 51–61.

Elton, L. (2004), 'Goodhart's law and performance indicators in higher education', *Evaluation & Research in Education* **18**(1-2), 120–128.

Engelbart, D. C. (2001), *Augmenting human intellect: a conceptual framework (1962)*, Multimedia: From Wagner to Virtual Reality, WW Norton & Company, pp. 64–90.

Ensign, D., Friedler, S., Neville, S., Scheidegger, C. & Venkatasubramanian, S. (2017), 'Runaway feedback loops in predictive policing'.

Eskeland, G. S. & Feyzioglu, T. (1997), 'Rationing can backfire: the "day without a car" in mexico city', *The World Bank Economic Review* **11**(3), 383–408.

Eterno, J. A. & Silverman, E. B. (2010), 'The NYPD's Compstat: Compare Statistics or Compose Statistics?', *International Journal of Police Science & Management* **12**(3), 426–449.

Evans, P. (1985), 'Money, output and goodhart's law: The us experience', *The Review of Economics and Statistics* pp. 1–8.

Evans, T. & Harris, J. (2004), 'Street-level bureaucracy, social work and the (exaggerated) death of discretion', *The British Journal of Social Work* **34**(6), 871–895.

Federal Trade Commission (2016), 'Big Data: A tool for inclusion or exclusion? Understanding the issues (FTC Report)'.

Ferguson, A. G. (2014), 'Big data and predictive reasonable suspicion', *University of Pennsylvania Law Review* **163**(327), 327–410.

Ferguson, A. G. (2017), 'The truth about predictive policing and race', *The Appeal* .
URL: *https://theappeal.org/the-truth-about-predictive-policing-and-race-b87cf7c070b1/*

Ferrara, E. (2015), 'Manipulation and abuse on social media', pp. 1–9.

Firestone, D. (1996), 'The Bratton Resignation: The Overview; Bratton Quits Police Post; New York Gains Over Crime Fed A Rivalry With Giuliani', *New York Times* .

Fisher, M., Goddu, M. K. & Keil, F. C. (2015), 'Searching for explanations: How the Internet inflates estimates of internal knowledge.', *Journal of Experimental Psychology: General* **144**(3), 674–687.

Gates, S. & Podder, S. (2015), 'Social Media, Recruitment, Allegiance and the Islamic State', *Perspectives on Terrorism* **9**(4), 107–116.

Gershgorn, D. (2017), 'Software used to predict crime can now be scoured for bias'.
URL: *https://qz.com/938635/a-predictive-policing-startup-released-all-its-code-so-it-can-be-scoured-for-bias/*

Gill, J. (1999), 'The insignificance of null hypothesis significance testing', *Political Research Quarterly* **52**(3), 647–674.

Goel, S. (2017), 'Creating Simple Rules for Complex Decisions', *Harvard Business Review* .

Goel, S., Rao, J. M. & Shroff, R. (2016), 'Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy', *The Annals of Applied Statistics* **10**(1), 365–394.

Golub, A., Johnson, B. D. & Dunlap, E. (2006), 'Smoking marijuana in public: the spatial and policy shift in new york city arrests, 1992–2003', *Harm Reduction Journal* **3**(1), 22.

Golub, A., Johnson, B. D. & Dunlap, E. (2007), 'The race/ethnicity disparity in misdemeanor marijuana arrests in new york city', *Criminology & public policy* **6**(1), 131–164.

Goodhart, C. A. (2006), 'The ECB and the Conduct of Monetary Policy: Goodhart's Law and Lessons from the Euro Area', *JCMS: Journal of Common Market Studies* **44**(4), 757–778.

Goodhart, C. A. & Courakis, A. (1981), *Problems of Monetary Management: The UK Experience*, Inflation, Depression and Economic Policy in the West, Barnes & Nobel Books, pp. 111–143.

Griffin, D., Tversky, A. & 1992 (1992), 'The weighing of evidence and the determinants of confidence', *Cognitive Psychology* **24**(3), 411–435.

Guadagno, R. E., Okdie, B. M. & Muscanell, N. L. (2013), 'Have We All Just Become "Robo-Sapiens"? Reflections on Social Influence Processes in the Internet Age', *Psychological Inquiry* **24**(4), 301–309.

Hamden Police Department (2015), '**Edward Byrne Memorial Justice Assistance Grant (JAG) Program**'.
URL: *http://www.hamden.com/content/219/228/9315/default.aspx*

Harcourt, B. E. (2009), *Illusion of Order: The False Promise of Broken Windows Policing*, Harvard University Press.

Hart, A. (2004), 'Report Finds Atlanta Police Cut Figures On Crimes', *New York Times* .

Hawkes, A. G. (1971), 'Spectra of some self-exciting and mutually exciting point processes', *Biometrika* **58**(1), 83–90.

Hayden, T. (2013), 'Dismantling the Myth of Bill Bratton's LAPD', *The Nation* .
URL: *https://www.thenation.com/article/dismantling-myth-bill-brattons-lapd/*

Heller, S. B. (2014), 'Summer jobs reduce violence among disadvantaged youth', *Science* **346**(6214), 1219–1223.

Hersh, E. D. (2015), *Hacking the Electorate: How Campaigns Perceive Voters*, Cambridge University Press.

Hirschfeld Davis, J. (2017), 'Trump declares opioid crisis a 'health emergency' but requests no funds', *New York Times* .
URL: *https://www.nytimes.com/2017/10/26/us/politics/trump-opioid-crisis.html*

Holzinger, A. (2016), 'Interactive machine learning for health informatics: when do we need the human-in-the-loop?', *Brain Informatics* **3**(2), 119–131.

Human Rights Watch (2018), 'China: Big data fuels crackdown in minority region'.
URL: *https://www.hrw.org/news/2018/02/26/china-big-data-fuels-crackdown-minority-region*

Humphreys, D. K., Gasparrini, A. & Wiebe, D. J. (2017), 'Evaluating the Impact of Florida's "Stand Your Ground" Self-defense Law on Homicide and Suicide by Firearm', **177**(1), 44–7.

Hunt, P., Saunders, J. & Hollywood, J. S. (2014), Evaluation of the Shreveport predictive policing experiment, Technical report.

Hvistendahl, M. (2016), 'Crime forecasters', *Science* **353**(6307), 1484–1487.

Isaac, W. S. (2017), 'Hope, hype, and fear: The promise and potential pitfalls of artificial intelligence in criminal justice', *Ohio State Journal of Criminal Law* **15**, 543.

Isaac, W. S. & Lum, K. (2018), 'Setting the record straight on predictive policing and race', *The Appeal* .
URL: *https://theappeal.org/setting-the-record-straight-on-predictive-policing-and-race-fe588b457ca2/*

Jacob, B. A. & Levitt, S. D. (2002), 'Rotten apples'.

Joh, E. E. (2014), 'Policing by numbers: Big data and the fourth amendment', *Washington Law Review* **89**(35), 35–68.

Joh, E. E. (2017), ‘Feeding the machine: Policing, crime data, & algorithms’, *William Mary Bill of Rights Journal* **26**(2), 287–302.

Johnson, D. D. P. & Fowler, J. H. (2011), ‘The evolution of overconfidence’, *Nature* **477**(7364), 317–320.

Jones, J. J., Bond, R. M., Bakshy, E., Eckles, D. & Fowler, J. H. (2017), ‘Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election’, *PLoS ONE* **12**(4).

Kandel, S., Sarig, O. & Wohl, A. (2001), ‘Do investors prefer round stock prices? evidence from israeli ipo auctions’, *Journal of banking & finance* **25**(8), 1543–1551.

Kelling, G. L. & Wilson, J. Q. (1982), ‘Broken Windows’.

Kilbertus, N., Gascón, A., Kusner, M. J., Veale, M., Gummadi, K. P. & Weller, A. (2018), ‘Blind justice: Fairness with encrypted sensitive attributes’, *arXiv preprint arXiv:1806.03281* .

Kim, G.-H., Trimi, S. & Chung, J.-H. (2014), ‘Big-data applications in the government sector’, *Communications of the ACM* **57**(3), 78–85.

Knafo, S. (2016), ‘A Black Police Officer’s Fight Against the N.Y.P.D.’, *New York Times* .

Kobak, D., Shpilkin, S., Pshenichnikov, M. S. et al. (2016), ‘Integer percentages as electoral falsification fingerprints’, *The Annals of Applied Statistics* **10**(1), 54–73.

Kossovsky, A. E. (2014), *Benford’s law: theory, the general law of relative quantities, and forensic fraud detection applications*, World Scientific.

Lazer, D., Kennedy, R. & Vespignani, A. (2014), ‘The parable of Google Flu: traps in big data Analysis’, **343**(6176), 1203–1205.

Levitt, S. D. (1998), ‘The Relationship Between Crime Reporting and Police: Implications for the Use of Uniform Crime Reports’, *Journal of Quantitative Criminology* **14**(1), 61–81.

Lipsky, M. (1971), ‘Street-level bureaucracy and the analysis of urban reform’, *Urban Affairs Quarterly* **6**(4), 391–409.

Lipsky, M. (2010), *Street-Level Bureaucracy, 30th Ann. Ed.: Dilemmas of the Individual in Public Service*, Russell Sage Foundation.

Logg, J. (2017), ‘Theory of Machine: When Do People Rely on Algorithms?’, pp. 1–96.

Loh, K. K. & Kanai, R. (2016), ‘How Has the Internet Reshaped Human Cognition?’, *The Neuroscientist* **22**(5), 506–520.

Lum, K. & Isaac, W. (2016), ‘To predict and serve?’, *Significance* **13**(5), 14–19.

Lum, K., Swarup, S., Eubank, S. & Hawdon, J. (2014), ‘The contagious nature of imprisonment: an agent-based model to explain racial disparities in incarceration rates’, *Journal of The Royal Society Interface* **11**(98), 1–12.

Lynn, M., Flynn, S. M. & Helion, C. (2013), 'Do consumers prefer round prices? evidence from pay-what-you-want decisions and self-pumped gasoline purchases', *Journal of Economic Psychology* **36**, 96–102.

MacDonald, Z. (2002), 'Official crime statistics: Their use and interpretation', *The Economic Journal* **112**(477), F85–F106.

Manheim, D. & Garrabrant, S. (2018), 'Categorizing Variants of Goodhart's Law'.

Mäntymäki, M. & Islam, A. K. M. N. (2016), 'The Janus face of Facebook: Positive and negative sides of social networking site use', *Computers in Human Behavior* **61**(C), 14–26.

Martin, D. (2001), 'Jack Maple, 48, a Designer of City Crime Control Strategies', *New York Times* .

Matarić, M. J. (2017), 'Socially assistive robotics: Human augmentation versus automation', *Science Robotics* **2**(4), 5410–5413.

McDonald, J. H. (2015), Cochran–Mantel–Haenszel test, *in* 'Handbook of Biological Statistics', biostathandbook.com, Baltimore, pp. 94–100.

McEvers, K. (2016), 'When It Comes To Policing LA's Skid Row, What Tactics Work?'.
URL: *http://www.npr.org/2016/04/21/474849734/when-it-comes-to-policing-las-skid-row-what-tactics-work*

Miller, A. P. (2018), 'Want less-biased decisions? use algorithms', *Harvard Business Review* .
URL: *https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms*

Mohler, G. (2013), 'Modeling and estimation of multi-source clustering in crime and security data', *The Annals of Applied Statistics* **7**(3), 1525–1539.

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L. & Brantingham, P. J. (2015), 'Randomized controlled field trials of predictive policing', *Journal of the American Statistical Association* **110**(512), 1399–1411.

Mohler, G., Short, M. B., Brantingham, J., Schoenberg, F. P. & Tita, G. E. (2011), 'Self-exciting point process modeling of crime', *Journal of the American Statistical Association* **106**(493), 100–108.

Moore, D. A. & Healy, P. J. (2008), 'The trouble with overconfidence.', *Psychological Review* **115**(2), 502–517.

Morrison, W. D. (1897), 'The interpretation of criminal statistics', *Journal of the Royal Statistical Society* **60**(1), 1–32.

Mosher, C. J., Miethe, T. D. & Hart, T. C. (2010), *The Mismeasure of Crime*, SAGE Publications.

Mummolo, J. (2018), 'Modern police tactics, police-citizen interactions, and the prospects for reform', *The Journal of Politics* **80**(1), 1–15.

Munn, N. (2017), 'Canada's Pre-Crime Model of Policing Is Sparking Privacy Concerns', *Motherboard* .
URL: *https://motherboard.vice.com/en*$_{u}s/article/mg7w4x/canada - hub - and - cor - policing - privacy - police$

Munoz, C., Smith, M. & Patil, D. (2016), *Big data: A report on algorithmic systems, opportunity, and civil rights*, Executive Office of the President.

Nakamoto, S. (2008), 'Bitcoin: A peer-to-peer electronic cash system'.

National Institute of Justice (2016), 'Real-Time Crime Forecasting Challenge'.
URL: *https://www.nij.gov/funding/Pages/fy16-crime-forecasting-challenge.aspx*

Naylor, B. & Keith, T. (2017), 'Trump Says He Intends To Declare Opioid Crisis National Emergency', *National Public Radio* .
URL: *http://www.npr.org/2017/08/10/542669730/trump-says-he-intends-to-declare-opioid-crisis-national-emergency*

Newton, A. C. (2011), 'Implications of goodhart's law for monitoring global biodiversity loss', *Conservation Letters* **4**(4), 264–268.

Nickerson, D. W. & Rogers, T. (2014), 'Political campaigns and big data', *Journal of Economic Perspectives* **28**(2), 51–74.

Nickerson, R. S. (2000), 'Null hypothesis significance testing: a review of an old and continuing controversy.', *Psychological methods* **5**(2), 241.

Niforatos, E., Vourvopoulos, A. & Langheinrich, M. (2017), Amplifying human cognition, *in* 'the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2017 ACM International Symposium on Wearable Computers', pp. 673–680.

Nigrini, M. J. (1996), 'A taxpayer compliance application of benford's law', *The Journal of the American Taxation Association* **18**(1), 72.

OMalley, N. (2013), 'The numbers men of LA', *Sydney Morning Herald* .
URL: *https://www.smh.com.au/technology/the-numbers-men-of-la-20130330-2h0bg.html*

O'Neil, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown.

Ortoleva, P. & Snowberg, E. (2015), 'Overconfidence in Political Behavior', *American Economic Review* **105**(2), 504–535.

Osler, C. L. (2003), 'Currency orders and exchange rate dynamics: an explanation for the predictive success of technical analysis', *The Journal of Finance* **58**(5), 1791–1819.

Oswald, M., Grace, J., Urwin, S. & Barnes, G. C. (2017), 'Algorithmic risk assessment policing models: Lessons from the durham hart model and 'experimental'proportionality'.

Parascandola, R. (2010), ‘Cops urged to prove stats clean by Citizens Crime Commission ’, *New York Daily News* .

Peltokorpi, V. (2008), ‘Transactive memory systems.’, *Review of General Psychology* **12**(4), 378–394.

Perry, W. L. (2013), ‘Predictive policing: The role of crime forecasting in law enforcement operations’.

Podesta, J., Pritzker, P., Moniz, E., Holdren, J. & Zients, J. (2014), *Big data: Seizing Opportunities, Preserving Values*, Executive Office of the President.

Pope, D. & Simonsohn, U. (2011), ‘Round numbers as goals: Evidence from baseball, sat takers, and the lab’, *Psychological science* **22**(1), 71–79.

Poston, B. & Rubin, J. (2014), ‘LAPD misclassified nearly 1,200 violent crimes as minor offenses’, *Los Angeles Times* .

Poston, B., Rubin, J. & Pesce, A. (2015), ‘LAPD underreported serious assaults, skewing crime stats for 8 years’, *Los Angeles Times* .

Powles, J. & Hodson, H. (2017), ‘Google deepmind and healthcare in an age of algorithms’, *Health and technology* **7**(4), 351–367.

Proeger, T. & Meub, L. (2014), ‘Overconfidence as a social bias: Experimental evidence’, *Economics Letters* **122**(2), 203–207.

Purdum, T. S. (1987), ‘Transit Scandal: Do Arrest Incentives Motivate the Police or Invite Abuse?’, *New York Times* .

Quatrevaux, E. R. (2014), **A Performance Audit of the New Orleans Police Department’s Uniform Crime Reporting of Forcible Rapes** , Technical report.

Ramunni, K. (2015), ‘Hamden police using app, eyeing software to help fight crime - New Haven Register’, *New Hampshire Register* .
URL: *http://www.nhregister.com/connecticut/article/Hamden-police-using-app-eyeing-software-to-help-11345181.php*

Ratcliffe, J. H. & McCullagh, M. J. (2001), ‘Chasing Ghosts? Police Perception of High Crime Areas’, *British Journal of Criminology* **41**(2), 330–341.

Rath, G. J. (1966), ‘Randomization by humans’, *The American Journal of Psychology* **79**(1), 97–103.

Rayman, G. (2018), ‘NYPD captain claims some cops are fudging crime stats to advance their careers, igniting police investigation’, *New York Daily News* .

Reynaert, M. & Sallee, J. M. (2016), ‘Corrective policy and goodhart’s law: The case of carbon emissions from automobiles’.

Risko, E. F. & Gilbert, S. J. (2016), ‘Cognitive Offloading’, *Trends in Cognitive Sciences* **20**(9), 676–688.

Robinson, D. & Koepke, L. (2016), Stuck in a pattern stuck in a pattern: Early evidence on “predictive policing" and civil rights, Technical report, Upturn.

Robison, S. M. (1936), *Can Delinquency Be Measured?*, Periodicals Service Company.

Roeder, O., Eisen, L.-B., Bowling, J., Stiglitz, J. & Chettiar, I. (2015), ‘What Caused the Crime Decline?’.

Rozeboom, W. W. (1960), ‘The fallacy of the null-hypothesis significance test.’, *Psychological bulletin* **57**(5), 416.

Rozenas, A. (2017), ‘Detecting Election Fraud from Irregularities in Vote-Share Distributions’, *Political Analysis* **25**(01), 41–56.

Saunders, J., Hunt, P. & Hollywood, J. S. (2016), ‘Predictions put into practice: a quasi-experimental evaluation of Chicago’s predictive policing pilot’, *Journal of Experimental Criminology* **12**(3), 1–25.

Schirner, G., Erdogmus, D., Chowdhury, K. & Padir, T. (2013), ‘The future of human-in-the-loop cyber-physical systems’, *Computer* **46**(1), 36–45.

Schmidt, A. (2017), ‘Augmenting Human Intellect and Amplifying Perception and Cognition’, *Computing* (January-March), 6–10.

Selbst, A. & Barocas, S. (2017), Regulating Inscrutable Systems, Technical report.

Sewell, A. A. (2017), ‘The Illness Associations of Police Violence: Differential Relationships by Ethnoracial Composition’, *Sociological Forum* **43**(2), 407–23.

Sewell, A. A., Jefferson, K. A. & Lee, H. (2016), ‘Living under surveillance: Gender, psychological distress, and stop-question-and-frisk policing in New York City’, *Social Science & Medicine* **159**, 1–13.

Shapiro, A. (2017), ‘Reform predictive policing’, *Nature* **541**(7638), 458–460.

Shneiderman, B. (2016), ‘Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight.’, *Proceedings of the National Academy of Sciences of the United States of America* **113**(48).

Siebert, H. (2001), *Der Kobra-Effekt: wie man Irrwege der Wirtschaftspolitik vermeidet*, Dt. Verlag-Anst.

Silver, N. (2015), *The Signal and the Noise*, Why So Many Predictions Fail, But Some Don’t, Penguin.

Skagestad, P. (1993), ‘Thinking with machines: Intelligence augmentation, evolutionary epistemology, and semiotic’, *Journal of Social and Evolutionary Systems* **16**(2), 157–180.

Small, G. W., Moody, T. D., Siddarth, P. & Bookheimer, S. Y. (2009), 'Your brain on google: patterns of cerebral activation during internet searching', *The American Journal of Geriatric Psychiatry* **17**(2), 116–126.

Smith IV, J. (2016), 'Predictive policing only amplifies racial bias, study shows', *Mic* .
URL: *https://mic.com/articles/156286/crime-prediction-tool-pred-pol-only-amplifies-racially-biased-policing-study-shows*

Sowe, S. K., Simmon, E., Zettsu, K., de Vaulx, F. & Bojanova, I. (2017), 'Cyber-Physical-Human Systems: Putting People in the Loop', *IT Professional* **18**(1), 10–13.

Sparrow, B. & Chatman, L. (2013), 'Social Cognition in the Internet Age: Same As It Ever Was?', *Psychological Inquiry* **24**(4), 273–292.

Sparrow, B., Liu, J. & Wegner, D. M. (2011), 'Google effects on memory: Cognitive consequences of having information at our fingertips', *Science* **333**(6043), 776–778.

Stone, C., Foglesong, T. & Cole, C. M. (2009), Policing Los Angeles Under a Consent Degree: The Dynamics of Change at the LAPD, Technical report, Boston MA.

Tchekmedyian, A. (2016*a*), 'Burbank police implement changes following survey indicating low morale in department', *Los Angeles Times* .

Tchekmedyian, A. (2016*b*), 'Police push back against using crime-prediction technology to deploy officers'.

Tench, S., Fry, H. & Gill, P. (2016), 'Spatio-temporal patterns of IED usage by the Provisional Irish Republican Army', *European Journal of Applied Mathematics* **27**(3), 377–402.

Thomas, E. (2016), 'Why oakland police turned down predictive policing'.
URL: *https://motherboard.vice.com/en*$_{u}s/article/ezp8zp/minority-retort-why-oakland-police-turned-down-predictive-policing$

Thornhill, J. (2018), 'The March of the Technocrats', *Financial Times* .

Tummers, L. & Bekkers, V. (2013), 'Policy Implementation, Street-level Bureaucracy, and the Importance of Discretion', *Public Management Review* **16**(4), 527–547.

Vann, M. G. (2003), 'Of rats, rice, and race: The great hanoi rat massacre, an episode in french colonial history', *French Colonial History* **4**(1), 191–203.

Veale, M. (2017), 'Logics and practices of transparency and opacity in real-world applications of public sector machine learning'.

Wagner, A. K., Soumerai, S. B., Zhang, F. & Ross-Degnan, D. (2002), 'Segmented regression analysis of interrupted time series studies in medication use research', *Journal of clinical pharmacy and therapeutics* **27**(4), 299–309.

Walsh, W. F. (2001), 'Compstat: an analysis of an emerging police managerial paradigm', *Policing: An International Journal of Police Strategies & Management* **24**(3), 347–362.

Ward, A. (2013*a*), One with the Cloud: Why People Mistake the Internet's Knowledge for Their Own, PhD thesis.

Ward, A. F. (2013*b*), 'Supernormal: How the Internet Is Changing Our Memories and Our Minds', *Psychological Inquiry* **24**(4), 341–348.

Wegner, D. M. (1987), Transactive Memory: A Contemporary Analysis of the Group Mind, *in* 'Theories of Group Behavior', Springer, New York, NY, New York, NY, pp. 185–208.

Wei, Y., Yildirim, P., Van den Bulte, C. & Dellarocas, C. (2016), 'Credit scoring with social network data', *Marketing Science* **35**(2), 234–258.

Weisburd, D. (2018), 'Hot Spots of Crime and Place-Based Prevention', *Criminology & Public Policy* **17**(1), 5–25.

Western, B. & Wildeman, C. (2009), 'The black family and mass incarceration', *The ANNALS of the American Academy of Political and Social Science* **621**(1), 221–242.

White, M. D. (2013), 'The New York City Police Department, its Crime Control Strategies and Organizational Changes, 1970-2009', *Justice Quarterly* **31**(1), 74–95.

Willis, J. J., Mastrofski, S. D. & Kochel, T. R. (2010), 'The Co-implementation of Compstat and Community Policing', *Journal of Criminal Justice* **38**(5), 969–980.

Wilson, R. K. (2011), 'The Contribution of Behavioral Economics to Political Science', *Annual Review of Political Science* **14**(1), 201–223.

Winston, A. (2015), 'Predictive policing is 'wave of the future,' NY commissioner says', *Reveal News* .

Winston, A. (2018), 'Palantir has secretly been using new orleans to test its predictive policing technology', *The Verge* .
URL: *https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd*

Wood, M. (2018), 'Thoughts on machine learning accuracy'.
URL: *https://aws.amazon.com/blogs/aws/thoughts-on-machine-learning-accuracy/*

World Economic Forum (2018), How to Prevent Discriminatory Outcomes in Machine Learning, Technical report.

Wormeli, P. (2018), 'Criminal justice statistics — an evolution', *Criminology & Public Policy* **17**(2), 483–496.

Zekulin, M. G. (2018), 'More than the medium: how the communication literature helps explain ISIS's success in recruiting Westerners', *Journal of Policing, Intelligence and Counter Terrorism* **13**(1), 17–37.

Zyskind, G. & Nathan, O. (2015), 'Decentralizing privacy: Using blockchain to protect personal data', *Security and Privacy Workshops (SPW . . .* pp. 180–184.