Team GAHSP

(Generating an Algorithm for Hot Spots Policing)

Cleansing Data and Bias within Predictive Policing Algorithms

May 2, 2025

Trina Arellano, Alex Chen, Allen Du,
Andrea Eichstadt, Aaron Lin, Nicole Samuels, Grace Tao,
Zoya Tasneem, and Rios Versace

Mentor: Mohammad Taghi Hajiaghayi
Librarian: Celine McDonald

Thesis submitted in partial fulfillment of the requirements of the
Gemstone Honors Program, University of Maryland, 2025

We pledge on our honor that we have not given or received any
unauthorized assistance on this assignment.

List of Abbreviations

BWP	Broken Windows Policing
BWT	Broken Windows Theory
HSP	Hot Spots Policing
NHSP	Non-algorithmic HSP
PP	Predictive Policing
ZTP	Zero Tolerance Policing

Glossary

Bias Cleansing	Collecting and analyzing data in such a manner that a conclusion derived from the data set does not produce results that differ systematically from the truth.
Broken Windows Policing	Broken Windows Theory as applied to policing, i.e shifting police resources toward quelling signs of disorder.
Broken Windows Theory	The theory that visible indicators of crime and civil disorder create an urban environment encouraging more crimes.
Data Cleansing	Removing data from a data set that is incorrect, unusable, or incomplete.
Hot Spots Policing	Focusing policing resources on limited areas where crime has a higher tendency to occur.
Predictive Policing	Predictive Policing (PP) is a subset of HSP, wherein the Hot Spots are designated by algorithms as opposed to analysts.
Zero Tolerance Policing	A subset of Broken Windows Policing which attempts to increase misdemeanor arrests to decrease violent crime.

Abstract

Hot spots policing—the allocation of police resources toward high-crime areas—has been revolutionized by machine learning. Instead of relying on historical crime hot spots, predictive policing algorithms allow departments to allocate officers to where crime is expected to occur next. This has led to their increasing adoption by especially large police departments, as well as modest reductions in crime.

However, predictive policing algorithms have thus been shown to exhibit similar biases to traditional policing methods. A vast literature has shown that nonwhite areas are more frequently policed, and that laws are disproportionately enforced against nonwhites in these communities. This creates a problem for predictive policing; since these algorithms are trained on historical crime data which reflects these racial biases, predictions come to perpetuate racial bias into the future.

As such, our team has built a new predictive algorithm which not only uses more contemporary machine learning techniques, but directly accounts for demographic fairness in its predictive judgments. Using real crime data, we then tested our model against PredPol, a state-of-the-art predictive policing software, comparing them on predictive accuracy and on racial bias in their predictions. Our results showed that our model outperformed PredPol in both predictive accuracy and fairness, demonstrating that it is possible to make policing more equitable without sacrificing the predictive accuracy of these algorithms.

Acknowledgements

We would like to thank our mentor, Dr. Mohammad Hajiaghayi, for his guidance throughout this project, especially his input on applying machine learning to predictive policing, which pushed us to think creatively in developing a fairer predictive policing algorithm. Thanks to our librarian, Celine McDonald, for helping us locate relevant resources to inform our research on hot spots policing, an important yet under-explored area. We also appreciate former police officer Robert Gorman for sharing his first-hand insights into policing practices and offering perspectives on how our project could have a meaningful impact in police departments. Finally, we are grateful to Gemstone directors Dr. David Lovell and Dr. Allison Lansverk for their continued support.

Contents

Chapter 1: Overview

According to the National Institute of Justice, Hot Spots Policing (HSP) is a common policing strategy used to selectively deploy police officers to geographic areas with high crime rates. As an implementation of HSP, a number of police departments use Predictive Policing (PP) algorithms, basing its predictions on past crime data and trends. The primary goal of PP algorithms, amongst other predictive methods, is to effectively distribute limited police resources. It follows that to distribute such police forces, they must be sent to crime clusters known as Hot Spots.

These Hot Spots may be defined as corners, blocks, or even neighborhoods that appear to be more predisposed to petty and violent crime (Braga et al., 2019a). Petty crime encompasses vandalism, burglary, drug dealing, and other relatively minor legal offenses (Eck et al., 2005). On the other hand, violent crime includes the more extreme side: homicide, assault, rape, and other serious offenses that severely impact victims (Braga et al., 2019a).

HSP appears noble and without major downsides; however, its potential side effects have put HSP under heavy scrutiny. As shown in Chapter 2, HSP's reliance on the flawed Broken Windows Theory—the idea that untended disorder and behavior of citizens precipitates higher police control—raises doubts about the algorithm's integrity. One of the salient critiques of PP algorithms is their alleged reinforcement of racial and socioeconomic stereotypes. Since more police are deployed to areas with historically high crime rates, unjust arrests are statistically more likely in those areas. Specifically, due to the long-term existence of racial profiling, police departments have historically made more arrests in areas with larger Black populations. This means predictive algorithms may forecast greater crime rates in Black communities, designating them as Hot Spots. In this sense, PP may lead to a detrimental self-fulfilling prophecy: greater police attention in Black communities leading to more arrests, which substantiate the initial Hot Spot forecast. Police presence would then further increase in these Hot Spots to cooperate with the predictions of the algorithm. With more arrests and lack of regard for crime severity, the bias against these communities would continue to be reinforced. Arrest rates would not accurately reflect the true crime rates of these areas relative to those not deemed as Hot Spots. This positive feedback loop not only perpetuates the targeting of certain communities, but also diverts resources away from preventing violent crimes occurring elsewhere (Rosenbaum, 2009).

As we will describe in Chapter 2, the results vary widely and the actual effects of PP in regard to racial bias are not well-documented. Most of the literature on this topic focuses on the effectiveness of HSP, but fails to zoom in on the bias within the data that it collects or its possible social consequences.

Team GAHSP aims to help solve this problem by creating a new PP algorithm that

appropriates police resources with a specific eye toward demographic fairness. Our motivation is simple: to reduce the extrinsic biases held within current algorithms. We believe that by using modern machine learning techniques, we can create an algorithm which is both more fair and more accurate, thus decreasing crime while also decreasing unnecessary and unjust arrests. We plan to focus on large cities as they are most often subject to HSP and have large amounts of crime and arrest documentation for use in a model.

As such, our driving research questions are:

1. What are the psychosocial impacts of PP, and racial bias in general?
2. In what ways can machine learning techniques be used to develop predictive algorithms, and what considerations need to be made for policing algorithms specifically?
3. What methods can be used to collect raw crime data that is consistent and exhaustive?
4. In addition to bias in data, what biases exist in police enforcement, and how can a PP algorithm correct for these biases?
5. What is the relationship between fairness and accuracy in a model, and what is the right tradeoff to make between them?

Chapter 2: Review of the Literature

Our project involves two broad topics: criminology and computer science. While our project is an intersection of these two fields, the literature will be partitioned for organizational purposes.

2.1: An Overview of Racial Bias in Policing

This section will establish the history of modern policing since the Broken Windows paradigm took hold, demonstrating the compatibility and throughline between Broken Windows Policing (BWP) and HSP. Racial bias resulting from BWP will be conclusively demonstrated, along with HSP's failure to address these problems. Finally, we will show how algorithm-based HSP fails to address inequities any better than traditional Non-algorithmic HSP (NHSP).

2.1.1: Broken Windows Policing

The origin of BWP is unspectacular, coming from a magazine article by criminologist George L. Kelling and political scientist James Q. Wilson. The article suggests, based on folk psychology and a misrepresented social experiment, that public disorder is a precursor to more violent crime. Thus, the article concludes that addressing public disorder is essential to preventing more serious crimes, and police resources ought to be directed toward that end (Kelling & Wilson, 1982).

Despite lacking empirical evidence of the proposed disorder and petty crime relationship, the article was highly influential. In 1993, New York mayor Rudy Giuliani brought on Bill Bratton, an intellectual admirer of Kelling, as the city's police commissioner (Thompson, 2015). Bratton implemented a form of BWP called Zero Tolerance Policing (ZTP), which aimed to crack down heavily on petty crimes perceived to create public disorder, including littering, panhandling, prostitution, public intoxication, vandalism, truancy, low-level drug violations, and more (Weisburd & Majmundar, 2018, p. 73). Unsurprisingly, such implementation of these ZTP policies led to a sharp increase in the number of misdemeanor arrests in New York City. In the first three years of the program, misdemeanor arrests rose from 133,446 to 205,277, peaking in 2010 with a total of 249,641 arrests (Chauhan et al., 2014, p. 18; Weisburd & Majmundar, 2018, p. 73). This is reflected even when accounting for population change: between 1980 and its peak in 2010, the misdemeanor arrest rate (misdemeanor arrests per 100,000 people) rose from 1.2 percent to 3.8 percent. The rise can be attributed to NYC's implementation of BWP because in the rest of New York State (where these policies were not implemented) the misdemeanor arrest rate remained at a stable 1.8 percent (Chauhan et al., 2014, p. 25).

BWP was nonetheless hailed as a success. Bratton publicly celebrated the efficacy of the program in a 1997 Op-Ed, attributing a 50% reduction in crime directly to BWP (Bratton, 1997). This supposed success sparked a revolution in policing, to the point that by 2011, 78.9% of police departments formally or informally adopted BWP practices (Mastrofski & Fridell, 2011, Figure 1). However, upon reflection, it seems likely that the supposed success of the NYPD had little to do with ZTP. Indeed, crime rates in US urban centers fell across the board, and more steeply so in areas with alternative policing approaches (Greene, 1999). The increase in misdemeanor offenses which are characteristic of ZTP (the most common variant of BWP) also seem to have done nothing: a meta-analysis of experimental and quasi-experimental BWP studies found that aggressive increases in misdemeanor arrests had no significant effect on crime reduction (Braga et al., 2015). Furthermore, it is unclear if the proposed relationship between disorder and crime is even causal. Longitudinal studies on the association between public disorder and crime have found little to no evidence of a causal relationship (Taylor, 2019; Yang, 2010). In truth, BWP had no empirical evidence backing it up when it was implemented, and as demonstrated, there has since been overwhelming evidence to indicate that the most significant impact of BWP is more people in jail.

This shift in policing focus (i.e., the shift in focus from violent crime to petty crime) serves as an important backdrop for all information hereafter. It is in this context that racial bias must be viewed, in the sense that the increased focus on misdemeanors is not race-neutral. As will be demonstrated in the next section, racial bias is most prevalent in these minor misdemeanors. Additionally, the only reason that arresting low-level offenders has become such a high priority is because of BWP, which as previously stated, has no research showing its actual effectiveness.

2.1.2: Bias in Policing

Racial bias in policing is conclusively affirmed by academic literature on the subject. This review will consider some studies on the petty crimes of drug use and traffic violations (which are the primary focus of BWP, as demonstrated earlier).

In the realm of drug use, the Justice Policy Institute (2007) conducted a study of 198 large population counties, home to just over half of the US population. Of these 198 counties, 193 of them were determined to have incredibly large racial disparities in drug admission rates. Across the whole dataset, Black people were 10 times as likely to be admitted on drug charges than White people. In the realm of traffic stops, the story is much the same. An analysis of the DC metropolitan area found that Black people constituted only 46.5% of the population, but over 70% of all traffic stops. In comparison, White people made up 37.1% of the population but only 14.1% of all traffic stops (ACLU-DC & ACLU Analytics, 2020). These results were replicated in a state-wide California report, which found that Black

people made up 17% of all traffic stops despite being only 7% of the population (Elgart et al., 2022). Indeed, these results carry across the whole of the US: an analysis of traffic stops from multiple states found that Black people were 1.43 times more likely than White people to be stopped by both state patrol cops and municipal police (Pierson et al., 2020).

Not only are these results staggering, but the literature makes it clear that racial discrimination, and not merely differential behavior, is the best explanation for racial discrepancies in policing and arrest rates. Given that Black people are 10 times more likely than White people to be admitted on drug charges, one would expect that Black people are 10 times more likely to use drugs; in reality, the rates of drug use between Black and White people are comparable (Justice Policy Institute, 2007). With traffic stops, the best-known test of racial bias is the "veil of darkness" test, wherein differences in traffic stops by race are compared before and after the sunset. Proponents of this method reason that if racial discrimination is a primary factor in why one is stopped by the police, then differential stops should decrease as the sunset goes down, as police are less able to identify the races of drivers (Pierson et al., 2020). Of the aforementioned studies on traffic stops, two conducted a veil of darkness analysis. In both cases, it was found that after dark, Black drivers made up a smaller portion of those subject to traffic stops than they were during the day, in line with the suggestion that racial bias held by officers is a contributing factor to racial differences in traffic stops (Elgart et al., 2022; Pierson et al., 2020). It is also important to remember that these biases are exacerbated by ZTP, which explicitly shifts police resources toward combating misdemeanor offenses. In New York City, the paradigmatic example of ZTP, this manifests in its stop-and-frisk policy. One study found that even when accounting for differential crime rates across racial groups, Black and Hispanic suspects were more likely to be stopped (Gelman et al., 2005). Because stops and frisks could be for any number of reasons, it shows how the prevalence of racial bias in policing extends beyond drug use and traffic stops. Furthermore, it suggests that these racial biases cannot so neatly be separated from the paradigm in which they exist.

While far from comprehensive, these studies heavily suggest that racial bias exists. Moving forward, racial bias in policing will be considered as a default against which different methods of policing are compared.

2.1.3: Negative Social Consequences of Discriminatory Policing

Before moving on to HSP and the potential factors which cause racial bias, it is essential to discuss the relevance of the issue. As such, we will explore ways in which excessive and racially targeted incarceration can have adverse effects (above and beyond the obvious negative experience of being arrested).

The negative consequences of discriminatory policing is extended to not just the community, but the individual. When people are targeted for arrest due to racial and gender

discrimination, they are affected by the consequences following the arrest. Many studies suggest a development of mental health issues, including post-traumatic stress disorder (PTSD) and anxiety (Swencionis & Goff, 2017). These mental health disorders are not just the result of direct interactions with law enforcement, but also tend to stem from chronic stress, fear and trauma that individuals experience when targeted due to their race or ethnicity. Other studies go even further to suggest that once someone is targeted by police based on their race, they have problems accepting themselves, commonly through an ethnic identity crisis (Lee et al., 2010). In other words, one questions their own racial identity. If police target someone based on their race, the victim begins to question if their every action, and their very existence, is in some sense "illegal." Going forward, the victim may have trouble developing socially and feel uncomfortable with their race and ethnic background. Everything they do from then on may warrant hesitance and fear (Lee et al., 2010). Additionally, discriminatory policing perpetuates familial strains and intergenerational trauma. Through research, it was shown that an arrest of a family member deeply impacts children. For one, they often experience feelings of abandonment, confusion and mitrust with police or people of authority. Losing a parent to incarceration induces economic and emotional instability which keeps the corrupt cycle of poverty and criminal involvement alive. Not only does this take an emotional toll on children, instilling grief and trauma, but it increases the likelihood of generational transmission, rebellion against figures of authority, and future arrests. (Lee et al., 2010). Furthermore, discriminatory policing eliminates trust in institutions, specifically law enforcement and the justice system. When certain groups are targeted, the community loses faith in the justice system's impartiality, weakening social cohesion and collective efficacy. This increases the likelihood of social disorder and crime and further exacerbates this cycle of mistrust and conflict between communities and law enforcement.

2.1.4: Bias in Hot Spots Policing

Hot Spots Policing (HSP), whether algorithmic or not, is based on the empirical observation that crime tends to be concentrated in specific locations. The rationale is simple: that directing police resources toward these areas will more efficiently reduce crime. In terms of this intended goal, HSP appears effective; a large-scale meta-analysis by Braga et al. (2019b) found small but statistically significant reductions in crime relative to control areas. However, this success does not make HSP immune to scrutiny – particularly regarding its racial implications.

At face value, there is no reason to think that HSP would be equipped to address inequities in policing. While its goal is more efficient resource allocation, it does not change the underlying tactics or institutional structure of policing itself. Rosenbaum (2009) argues that HSP has potential to increase efficiency, but not reform what police actually do. Further,

he specifically notes the potential for HSP to be used supplementally with Broken Windows Policing (BWP) or Zero Tolerance Policing (ZTP).

This concern is reinforced by several theoretical issues. One issue involves the scale and breadth of Hot Spots. A designated high-crime area might include only a single problematic street, while the rest of the area remains nominally safe (or low-crime). Yet police resources would be deployed throughout the entire zone, increasing the likelihood of over-policing and unjust targeting (Wiesburd, 2016). Another concern has to do with police behavior: deploying officers to an area may expose them to more potential crimes, but whether and how they investigate those crimes remains susceptible to individual racial bias. As discussed in Section 2.1.1, such biases have been demonstrated in traffic stops and arrests (Knox et al., 2020). Additionally, HSP is widespread: one survey found that 90.8% of departments officially or unofficially incorporate HSP practices or its key features (Mastrofski & Fridell, 2011, Figure 1). Given that the aforementioned studies on police bias were conducted after this survey, and were thus likely conducted in areas where HSP is employed, it does not inspire much confidence in the idea that HSP can eliminate racial bias.

Direct empirical evidence assessing racial bias in HSP remains limited, as much of the research emphasizes crime reduction over equity. However, a few studies stand out. Notably, Barnes (2018) examined traffic enforcement within designated Hot Spots to determine whether racial disparities existed in stops and searches. The study used two methods to measure racial bias: a Chi-squared test and a Disproportionality Index (DI) derived from the literature (Barnes, 2018, pp. 49, 52). Barnes found that Black drivers were significantly more likely to be stopped and searched despite having no significant difference in found contraband (2018, Chapter 4). Furthermore, the DI indicated that Black drivers were 48% more likely to be stopped inside Hot Spots than outside Hot Spots relative to their portion of the driving population (Barnes, 2018, Appendix A).

While this study provides valuable insight, it is not conclusive. The Chi-squared test assumed equal racial distribution across all areas and did not control for demographic differences between Hot Spots and non-Hot Spots or for potential differences in driving behavior across racial groups (Barnes, 2018, p. 53). Although the DI method avoids the assumption of uniform population makeup by using the area's actual driving population as a benchmark (Barnes, 2018, p. 49), it still does not address behavioral variables that might influence stop rates. In short, while the DI offers a more population-sensitive metric, it cannot fully isolate bias from other factors. Moreover, the DI method is not a widely used test, and its results cannot be assessed for statistical significance.

Although Barnes' study indicates that Hot Spots may exacerbate racial disparities, the broader empirical record remains inconclusive. More research is needed to determine with certainty whether Hot Spots are more biased than traditional patrol patterns. Still, the

available direct and indirect evidence gives little reason to believe that HSP reduces racial bias. In sum, even as empirical evidence remains scarce, the theoretical and practical concerns surrounding HSP's capacity to address racial inequity are increasingly difficult to miss.

2.1.5: Bias in Predictive Policing

Predictive Policing (PP) is a subset of Hot Spots Policing (HSP), in which Hot Spots are designated by algorithms as opposed to human analysts. Crucially, present algorithms do not put police where crime has already occurred, but rather to where the algorithm predicts crime will occur in the future (Bennett Moses & Chan, 2018). With how sparse research on racial bias is for HSP, the state of research in PP is even more sparse. Nonetheless, certain inferences can still be drawn.

First, all the same theoretical critiques levied at HSP in Section 2.1.2 apply equally to PP. Although PP currently accounts for a minority of HSP implementations, its usage is growing rapidly. According to Reaves (Reaves, 2011), only 13% of departments use algorithms for hot-spot overall, but departments that served larger numbers of people were far more likely to do so. In the sample, 100% of the departments serving between 500,000 to 999,999 people used algorithms for hot-spot identification, as did 92% of departments serving 1,000,000 or more people (Reaves, 2011, p. 22).

Furthermore, there is direct empirical evidence to suggest PP and non-algorithm HSP are equivalent with respect to racial bias. Brantingham et al. (Brantingham et al., 2018) analyzed the results of a randomized control trial conducted in Los Angeles to test for differences in biased arrests between algorithm-created Hot Spots (treatment) and analyst-created Hot Spots (control), making for an apt comparison between PP and traditional non-algorithm HSP. They found no statistically significant difference between the proportion of minority arrests in treatment days and control days. While this does not conclusively establish the equivalence between PP and HSP, the randomized study design and large sample provide strong evidence for that conclusion in the absence of contrary evidence.

Many questions remain unanswered concerning the differences between HSP and PP that might give way to bias, but the evidence currently does not permit the judgment that PP is any less biased than non-algorithm HSP. In light of this, and the existence of evidence indicating racial bias in HSP generally, the conclusion that PP is biased as well is supported.

2.2: Data, Algorithms, and Machine Learning in Law Enforcement Activities

Predictive algorithms are trained to identify patterns and consistencies within datasets. In the context of Predictive Policing (PP), these algorithms may analyze variables such as crime types, geographic distribution of offenses, and demographic information on those who commit crime. The goal is to predict where and when future crimes are likely to occur – particularly through the identification of geographic Hot Spots – or, more precisely, to estimate the relative likelihood of crime occuring in a given area or by a given individual.

Many police departments in the United States rely on information fed to them by these predictive algorithms for resource allocation purposes, particularly in determining where to deploy officers (Braga et al., 2019a). However, because these models are trained on historical crime, court, and arrest data – datasets that often reflect longstanding patterns of systemic bias – they risk becoming self-affirming (Bennett Moses & Chan, 2018). That is, if biased data are used for training, the resulting models will likely reproduce and reinforce those same biases. Therefore, ensuring that the data used are truly representative and free of historical distortions is critical. For our research, the representativeness and integrity of the data will be a foundational standard in evaluating both algorithmic validity and bias.

In theory, machine learning offers a degree of objectivity that human decision-making lacks. Unlike humans, who bring internal beliefs and perceptions into the decision-making process, machines generate predictions solely based on input data. However, this advantage disappears when the data itself is biased. In such cases, machine learning models will mirror those biases. At present, many PP models that are currently in use rely on machine learning, and while these could serve as baselines for our project, most existing systems are proprietary. This makes it difficult to assess their internal logic or correct for specific sources of bias. Even if access were granted, the complexity of these models can make it difficult to pinpoint which components contribute to biased outcomes.

Moreover, the generalizability of machine learning models depend heavily on the quality and diversity of the training data. If datasets are incomplete, outdated, or unrepresentative of the communities to which the model is applied, predictions may not reflect reality. This is a persistent challenge in criminal justice applications, where comprehensive, unbiased datasets are often difficult to obtain.

Relying on crime data alone would not give the clearest picture of where the bias lies. It could be the HSP algorithm enforcing biases in our data set or individual police officers being racially motivated causing one group to be over-represented in a data set; however, we would not be able to determine the exact cause of bias. Some machine learning models have relied heavily on protected characteristics, such as race or gender, while others have

completely erased them from their models (Corbett-Davies et al., 2024). There are issues with both cases, but in either case, it is essential that we define our limitations and the impact of including such classifications on our conclusion.

2.2.1: Algorithm Types, Characteristics, and Functionalities

There are three main types of machine learning algorithms: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Supervised Learning algorithms learn from labeled input-output data to map inputs to corresponding outputs. Unsupervised Learning algorithms, on the other hand, enable a model to analyze input data patterns and arrange them numerically, without the need for labeled data to guide the learning or training process. These algorithms are mainly useful for arranging data inputs into categories or clusters. Reinforcement Learning algorithms are trained through trial-and-error. They learn from the significance of their activities rather than being given explicit training data. Such algorithms improve themselves using the aforementioned trial-and-error techniques until they have reached a level where they make accurate and correct decisions (Dike et al., 2018).

A Simple Linear Regression model is a subsection of Supervised Learning algorithms. This model algorithm is characterized by one independent variable that is related to another dependent variable. In our case, this helps to find a correlation between the number of arrests made and the description of the location where the arrest was made. Furthermore, with a Supervised approach, we would be focusing on regression analysis. This helps clarify the relationship between dependent and independent variables, which, in our project, represent crimes and other unknown variables affecting crime (Delua, 2021).

In the context of Unsupervised Learning algorithms, our algorithm could implement a K-Means model, which classifies data into distinct clusters. For HSP, using this type of algorithm to cluster and classify the types and locations of arrests would organize our data, making it more efficient for further analysis. This format is beneficial to implement since there is no clear understanding of whether an arrest was influenced by race or other factors, and so the data, in that sense, would be unlabeled, allowing us to gain insight into the validity of arrests or the objectivity of current algorithms.

Another means of forecasting is deep neural networks. While neural networks can be extremely effective at forecasting, they are incredibly complex and prone to inhuman reasoning. These networks sometimes seem to perform well despite their outputs incorrectly relying on seemingly irrelevant features (Dressel & Farid, 2018). Consequently, deep neural networks are often called “black boxes”; they produce an output without any explicit reasoning, therefore making it extremely difficult to comprehend how the output came to be produced. Keeping this issue in mind, it can be very easy for neural networks to perpetuate bias that is present in biased data. Thus, it is important to analyze the effects of our neural network and ensure

that the algorithm is being trained on fair data, and the outcomes are also fair and equitable.

2.2.2: Algorithm Fairness

Data cleansing algorithms and the bias that arises within them have been addressed in the past. Research on these algorithms also emphasizes data transparency throughout the cleaning process, since the experiment data is bound to change. If the researcher tampers with the data being fed into the algorithm in any manner, it is important to note the change, as any change in data affects the output of the algorithm, potentially introducing bias into it.

Past research has also examined algorithms that predict recidivism, or the likelihood of a convicted criminal to relapse and commit another crime. According to Dartmouth scientists Julia Dressel and Hany Farid, most criminal justice algorithms represent the IP of private companies and are thus not open-source. The ones that are, however, are not ideal. For example, Dressel and Farid found the open-source COMPAS algorithm for predicting recidivism to be only 65% accurate, performing only as well as a simple linear classifier. Our goal, then, is to create an open-source algorithm that accurately identifies crime hotspots.

There are also different metrics in measuring the accuracy and fairness of an algorithm. For instance, the aforementioned 65% accuracy figure for COMPAS can be broken down by race, and whether the error was an overprediction or underprediction (predicting the defendant as respectively more or less dangerous than they turned out to be). Investigative reporters at ProPublica found that COMPAS disproportionately overpredicted Black defendants and underpredicted White defendants (Angwin et al., 2016). When faced with accusations of bias, COMPAS responded that their risk scores accurately predicted recidivism rates when grouped by race. For example, their average risk score of 5.5 for Black defendants corresponded to the approximately 55% Black recidivism rate.

To explain the apparent discrepancy, Kleinberg et al. (2016) formalized these opposing arguments, creating three metrics of fairness. Well-calibrated predictive algorithms will, for all groups, output an average risk score over a demographic group that is characteristic of that group. In the case of COMPAS, the demographic was race. COMPAS is well-calibrated since its average risk score of 5.5 for Black defendants did correspond accurately to the Black recidivism rate. In contrast to well-calibrated algorithms, positively and negatively balanced algorithms possess the same overpredicting and underpredicting errors, respectively, across all groups. The COMPAS algorithm was found to be positively unbalanced for Black defendants, since it disproportionately overpredicted Black recidivism rates when compared to the rate of overprediction of White defendants. The algorithm was analogously negatively unbalanced for White defendants (Kleinberg et al., 2016).

Building on Kleinberg et al. (2016), Corbett-Davies et al. (2024) demonstrate that it is

impossible to design an algorithm that is perfectly fair over all three previously discussed fairness metrics. In other words, an extremely well-calibrated algorithm is likely to be either positively or negatively unbalanced, and vice versa. However, both studies emphasize that inherent algorithm bias does not mean they should be discounted. Instead, it is the responsibility of the algorithm's designers to make the trade-offs between these metrics in a way that minimizes the negative social impact from their algorithm's inherent bias, as some metrics are situationally more important than others (Corbett-Davies et al., 2024; Kleinberg et al., 2016). In the COMPAS example, the algorithm–willingly or not–traded off positive and negative balance for being well-calibrated. It is thus our responsibility to analyze the three fairness metrics of every algorithm we improve upon or implement, and to consider how we should balance these metrics to minimize the negative impact of said algorithm's biases.

While so far we have discussed biases inherent to algorithms, Corbett-Davies et al. (2024) also discuss external sources of unfairness. COMPAS was designed in part to attempt to eliminate human racial biases in the sentencing process; in fact, their assessment contained no questions even tangentially related to race (such as geographic location, since a lot of Black residents still live in red-lined urban areas). In that case, race would be considered a "protected factor". However, it is possible for an algorithm to be biased against a certain protected factor; the researchers gave the example of literacy tests being used to prevent Black people from voting. Technically, literacy tests do not directly discriminate based on race, but due to systemic barriers to education such as red-lining residential districts and underfunding majority Black schools, the lower average education levels of Black prospective voters meant using literacy tests indirectly discriminated against them (Corbett-Davies et al., 2024).

When considering the notion of the calibration of algorithms, protected factors and their proxies must also be considered. The COMPAS algorithm, as stated above, failed to fully account for the impact of proxies. Proxy use, in the linear regression sense, is defined as computations which are both statistically associated with a random input variable and causally influential on the program's output. Yeom et al. (2018) argue that proxy use (or lack thereof) is a stronger measure of fairness than demographic parity, whereby different demographic groups receive identical outcomes on average, with zero disparate impact. Further, they developed a proxy detection system, accounting for exempt variables which represented proxies that were allowed to be kept (ones that had sufficient justification behind them). The system was tested on Chicago's Strategic Subject List (SSL) and the Communities and Crimes (C&C) datasets, the former of which claims its model avoids discrimination (i.e. has weak proxies, if at all), and the latter of which contains many variables correlated to race. They found proxies in the model trained on SSL data with insignificant influence, and conversely strong proxies for race in the model trained on C&C data; these outcomes substantiate the original predictions.

In the machine learning phase of our project, we must consider which fairness metrics and parity measures to emphasize. If applicable, the fairness of the final algorithm can also be assessed using a similar proxy detection algorithm.

2.3: Predictive Policing Algorithms and PredPol

While there are a variety of PP algorithms, our project is focused very specifically on PredPol. At the time of our project's conception, PredPol was the leading PP algorithm. Our algorithm was mostly conceived in response to it and used its predictions as a benchmark. As such, this section is dedicated to analyzing it in depth.

2.3.1: The Mathematics of PredPol

PredPol uses an epidemic type aftershock model, which is often used in seismology to predict the progression of earthquake aftershocks (Wong, 2022). It's based on the theory that some event, whether it be a crime or an earthquake, will create "aftershocks" in the immediate area that will then die out over time. To start, the area to be policed is divided into small (one implementation used $150m \times 150m$ square) cells. According to PredPol's model, the crime rate $r_h(t)$ on day t in cell h can be modeled as a combination of the inherent background rate of crime plus the shocks of recent crime events. In particular,

$$r_h(t) = \mu_h + \sum_{t_i<t, t_i \in I_h} \theta_h \omega_h \exp(-\omega(t - t_i)), \text{ where:}$$

- μ_h is the background rate, which does not vary with time
- I_h is the data, the set of all crimes reported in cell h
  - t_i are the dates said crimes were reported
- θ_h and ω_h are parameters expressing how quickly the shocks of recent crime events dissipate

Readers familiar with stochastic processes may recognize this as a Hawkes process, where μ_h is the rate of the underlying Poisson process and the summation is of an exponential kernel. Thus, $\mu_h, \theta_h, \omega_h$ are parameters of a distribution; however, maximum likelihood estimation is intractable. Instead, these are estimated via the expectation-maximization (EM) algorithm, an iterative algorithm that converges to local maximum likelihood estimators of $\mu_h, \theta_h, \omega_h$.

2.3.2 Positive Feedback Loops

As described previously, the PredPol algorithm uses only three variables to predict crime rate: a constant background crime rate, crime date, and crime location. The background crime rate is calculated using incident data, a combination of reported (e.g. calls for service) and discovered (e.g. arrest) data (Ensign et al., 2017). Thus, the algorithm is predicting where incidents, and not necessarily crime, will occur, and officers are dispatched to areas with highest predicted incident intensity. Furthermore, the summation in the algorithm is a self-exciting process, where the occurrence of new incidents in a cell temporarily boosts $r_h(t)$, the predicted crime rate for it. Two main sources of bias in PredPol's formulation arise:

1. Historical bias in μ_h: μ_h is based on past recorded crime data, which can be influenced by historically biased policing practices, such as over-policing of minority communities. A cell h with a high μ_h will be designated as a hot spot from the very beginning ($t = 0$), and, due to the addition of μ_h to the positive summation term, $r_h(t)$ for that cell can never fall below μ_h. Unless other cells' incident rates increase and are persistently high, it would be difficult to "undesignate" that cell as a hot spot. However, it can be problematic for a cell to experience an increase in incident rate, as explained below.

2. Positive feedback loops: Once a cell becomes a hot spot, it is likely to remain a hot spot. A crime that occurs at time t_i adds a positive term $\theta_h \omega_h \exp(-\omega(t - t_i))$ to future predictions, so $r_h(t)$ can only decay gradually as $t - t_i$ becomes larger. If $|t_i|$ denotes the number of crimes on day i in a cell, then a high $|t_i|$ contributes to a higher $r_h(t)$, and thus more officers will be dispatched to patrol that cell. But this just leads to more discovered incidents, as increased police presence is bound to increase the number of arrests, since officers may have incentives to catch crime (Lum & Isaac, 2016). This further increases $|t_i|$ and hence $r_h(t)$. However, the actual crime rate (i.e. the crime rate if the same number of officers were patrolling from day to day) may be stabilizing or decreasing.

In their study, (Lum & Isaac, 2016) demonstrate the positive feedback nature of PredPol using a synthetic population of Oakland, California. They estimated drug use by race using real survey data, and ran PredPol on this population. Despite the nearly equal proportions of Black and White drug users in the population, PredPol results showed the percentage of the Black population experiencing targeted policing for drug crimes to be nearly double that of targeted policing for White people. To better illustrate the feedback loop behind this result, they compared PredPol results for two datasets: the original data, and another dataset where they increased the observed crime in locations where PredPol assigned targeted policing by 20%, so these additional crimes were fed into the algorithm. For both datasets,

they plotted the ratio of $r_h(t)$ for targeted locations to $r_h(t)$ for non-targeted locations versus t. They found that the ratio became larger over time for the modified dataset, whereas it was relatively stable for the original dataset. This indicates the existence of a feedback loop, as PredPol became increasingly confident that the hot spots were the targeted locations.

2.3.3: Model Limitations and Misuse

As previously mentioned, PredPol's algorithm uses a mathematical method generally employed to model earthquake aftershocks. On its own, this is not inherently inappropriate. This is known as "off-the-shelf" machine learning, where an algorithm used to model one process is used to model a different, similarly-patterned process (Shah, 2024). It is that last criteria, that these processes have similar patterns, which is most important in determining the appropriateness of an "off-the-shelf" algorithm. When PredPol was first conceived, this was well accounted for. Gang violence and burglary tend to follow a pattern very similar to that of earthquake aftershocks, and so its use in predicting those crimes may be appropriate (Mohler et al., 2011). However, not all crime proliferation is patterned this way. It is well established, for example, that crimes directed against persons (such as sexual violence by a stranger) are patterned differently than crimes against property (Santos, 2017). PredPol is designed to predict the latter, and yet police departments seemingly use predictive policing algorithms without regard to the kind of crime being predicted. An investigative report by Gizmodo uncovered that many police departments were using PredPol to predict sex crimes (i.e, persons based crime), despite the fact that "the company advises clients against trying to predict [sex crimes]" (Lash, 2021).

So while PredPol's algorithm may be fit for purpose, there is no guarantee that clients of PredPol are using the algorithm appropriately. There is thus a potential gap to be filled by our project; by creating a general-purpose crime prediction algorithm which is able to use multiple models, we can obviate the risk of clients misusing the algorithm.

Chapter 3: Methodology

3.1: Building the GAHSP Algorithm

Historical crime data, dating back at least several years, was collected from police districts with public data reporting, namely Montgomery County, Maryland; Chicago, Illinois; and San Francisco, California (City of Chicago, 2024; dataMontgomery, 2024; DataSF, 2025). A square grid, with size chosen to match the approximately 70 people-per-cell density of Mohler et al.'s analysis of Los Angeles, was then overlaid onto these districts, and each crime labeled with the date and grid cell of occurrence (2015). Using 2020 census data, we obtained estimates for the White, Black, Hispanic, and Asian populations of each grid (U.S. Census Bureau, 2020). We then used state-of-the-art machine learning framework AutoGluon to forecast the future rates of crime in each grid, designating the top 100 cells with highest predicted rates as that day's hotspots (Erickson et al., 2020). AutoGluon trains multiple different models, ranging from simpler statistical models to complicated transformer-based models, and selects the best-performing one. To evaluate model performance, we used the same metrics Mohler et al. used in evaluating PredPol's performance. Below, we use the following notation in addition to the notation in Section 2.3.1:

- T is our window of observation, or how many days into the future we forecast
- H_t is the set of cells designated as hotspots for day t, which is a subset of C, the set of all cell labels
  - $H = |H_t|$ is the number of hotspots for day t, and is constant over time; for our analysis, we used $H = 100$
  - $H_T = \{H_t | 1 \leq t \leq T\}$ is each day's hotspots
- $n_t(h)$ is the number of crimes reported in cell h on day t
  - $N_t = \sum_{h \in C} n_t(h)$ is the total number of crimes reported that day
- $p_g(h)$ is the (estimated) population count of racial group g in cell h
  - $P_g(H_t) \neq \sum_{h \in H_t} p_g(h)$ is the population count of racial group g in all hotspot cells on day t
  - P is the total population
  - For our analysis, g is an element of $G \neq \{\text{White}, \text{Black}, \text{Hispanic}, \text{Asian}\}$

1. For accuracy, predictive accuracy index (PAI) is defined as

$$\text{PAI}(H_T) = \frac{N_t(H_t) \cdot |C|}{N_t \cdot H}, \text{ where:}$$

   - $N_t(H_t) = \sum_{h \in H_t} n_t(h)$ is the number of crimes reported in hotspots
   - $|C|$, the size of C, is the total number of cells (Chainey et al., 2008).

2. For demographic parity of policing, a so-called *fairness metric*, defined as

$$F(H_T) = \sum_{(g,g') \in G^2} (a_g - a_{g'})^2, \quad \text{where } a_g = \frac{\sum_{t=1}^{T} P_g(H_t)}{TP}$$

   (Mohler et al., 2018). Essentially, the fairness metric measures the variation in a_g, the police attention racial group g receives under hotspots.

Using AutoGluon's ability to receive custom loss functions, we created a loss function

$$L_\lambda(H_T) = \lambda F(H_T) - \text{PAI}(H_T),$$

a linear combination of PAI and fairness metrics, in an attempt for our model to maximize both hotspot accuracy and demographic parity. Here, λ is the weight we assign to fairness metric F, which varies as we examine the relationship between accuracy and parity. The signs of the functions come from the fact that the loss function should be minimized, so the lowest F and highest PAI is desirable.

To make a comparison with PredPol, we used Lum and Isaac's implementation of the PredPol model (2016). Both our model and PredPol were made to forecast the crime rates for the year to date starting from February 22, 2022, and the two aforementioned performance metrics were calculated and compared for both models. We used a 1-tailed paired T-Test to compare the performance of each model in both fairness and accuracy.

3.2: GAHSP Website

We have created a website[1] detailing the team's research and results, with interactivity for users. There are three main pages:

- **About:** Provides an overview of the team's inspiration, research questions, and background information on predictive policing

[1] https://team-gahsp.uk.r.appspot.com/

- **Map:** Features an interactive Google map displaying Montgomery County crimes using data from the county's Open Data Portal, with filtering options tailored to the dataset
- **Model:** Demonstrates our selected model's predicted crime hot spots in Montgomery County over a one-year period using a series of GIS-generated images

To enhance user experience, we implemented authentication to allow users to save their filtering preferences. In the future, we hope to expand functionality by:

- Displaying model results in real-time as new crime is recorded
- Allowing users to upload their own GeoJSON files for visualization and to run the model on
- Allowing interactive comparisons of historical data to predicted hot spots
- Integrating additional datasets from various U.S. cities
- Integrating related datasets, such as income or demographic data, to drive more insights

Chapter 4: Results and Discussion

4.1: Comparisons with PredPol

This section compares the predictive power and fairness of our algorithm to that of PredPol. We found that our model was able to handily outperform PredPol in both respects, but with an increased level of volatility.

4.1.1 Fairness and Accuracy Comparisons

Our comparison against PredPol went through several stages of development. At first, we tried to run our model with only a single training step, such that T was equal to 365. This appeared to give us better predictive power than PredPol, but it was less accurate in the top 100 hot spots. This is problematic because police resources are not infinitely divisible among each hot spot. In all likelihood, the top 100 hot spots are going to be the ones that departments pay most attention to, and thus it is more important for those to be correct.

To account for this, we switched to weekly model forecasting, such that T was equal to 7. This meant our model made predictions 7 days in advance, after which point the model could use the now-previous week's data as input in forecasting. This would still theoretically give PredPol an edge over our algorithm, as PredPol is designed to be retrained on a daily basis (Geolitica, 2023) while we simply roll a prediction window without retraining. PredPol is a much simpler model, and thus does not require as many computational resources for its predictions. Still, our model demonstrates PredPol's accuracy and parity can be jointly improved.

Our best-performing model was one where $\lambda \neq 5 \cdot 10^4$. Since AutoGluon trains multiple models and selects the best-performing one, the best model turned out to be an ensemble model with the following weights:

Component Model	Type	Weight (%)
DeepAR	(RNN) Encoder-Decoder	52
PatchTST	Transformer	23
TiDE	Transformer	16
TemporalFusionTransformer	Transformer	3
ChronosFineTuned[bolt_small]	(Pre-trained) Transformer	2
ChronosZeroShot[bolt_base]	(Pre-trained) Transformer	2
AutoETS	Statistical	1
DynamicOptimizedTheta	Statistical	1

To compare our model's performance against PredPol, we calculated each week's PAI and fairness metrics over the year of forecast. We found that our model (with $\lambda \neq 5 \cdot 10^4$) had a significantly lower average fairness metric compared with PredPol's fairness metric in all 3 cities (See Table 2), and our model either had a comparable (San Francisco) or significantly better (Montgomery County and Chicago) average PAI in all 3 cities (See Table 1) (Note that in this context, a lower score on the fairness metric means that an algorithm is more fair). In simple terms, our model was not only as or more accurate, but also more fair.

Figure 1: Weekly PAI Metric Improvement over PredPol

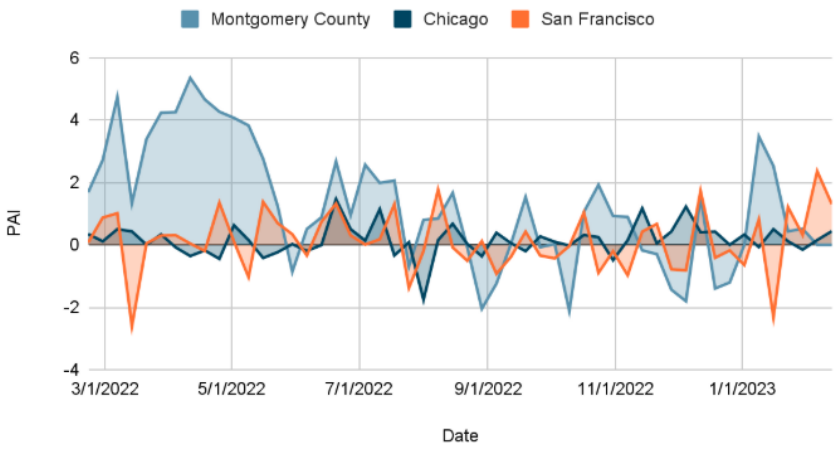


Table 1: Average PAI for Each Model in Each Location

Location	Model	Average PAI	P-value (on 1-tailed paired T-Test)
Montgomery County	GAHSP	19.50585343	>0.001**
	PredPol	18.19104892	
Chicago	GAHSP	9.177876486	0.013*
	PredPol	9.021024272	
San Francisco	GAHSP	16.26732116	0.16
	PredPol	1.625718051	

** Indicates significance at α of* 0.05
*** Indicates significance at α of* 0.01

However, there was a greater level of volatility in our model as compared to PredPol. For PAI in Montgomery County, the average absolute difference in score from week to week was 1.32, compared to 0.30 for PredPol. The same calculation done for the week to week fairness metric shows the GAHSP model to have a weekly absolute difference of $1.7 \cdot 10^{-5}$,

compared to PredPol's $7.7 \cdot 10^{-6}$. It is not entirely clear why this difference exists. One possible reason might be that PredPol is retrained more frequently, but further research is needed to determine the precise cause.

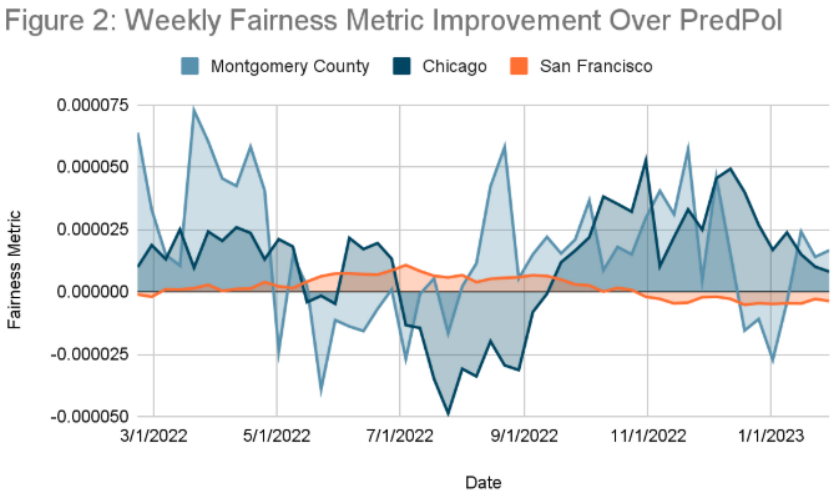


Table 2: Average Fairness Metric for Each Model in Each Location

Location	Model	Average Fairness Metric	P-value (on 1-tailed paired T-Test)
Montgomery County	GAHSP	1.49E-04	>0.001**
	PredPol	1.65E-04	
Chicago	GAHSP	1.59E-04	>0.001*
	PredPol	1.70E-04	
San Francisco	GAHSP	3.18E-05	>0.002**
	PredPol	3.37E-05	

** Indicates significance at α of* 0.05
*** Indicates significance at α of* 0.01

4.1.2: Discussion of Fairness and Accuracy Comparisons

Concern is understandable when using complex models, especially deep learning models, as their black-box nature makes them inexplicable. While AutoGluon indicated that the best-performing model was an ensemble of mostly black-box deep learning models, exponential smoothing (AutoETS) and theta (DynamicOptimizedTheta) statistical models delivered comparable performance with the benefit of explainability. Further research can be done into fitting an ensemble of only statistical models.

There is also a significant trade-off to be had in terms of the usage of computational resources. As previously stated, we did not do daily retraining to match PredPol because doing so would have been prohibitively costly. It's not clear that police departments have the necessary computational resources to run the GAHSP model daily given that we could not.

Nonetheless, our results show that there is not an absolute trade-off between fairness and accuracy when looking at differences between models. The fact that the GAHSP algorithm was both more accurate and less biased shows that it is not a fruitless endeavor to account for demographic fairness. Moreover, it speaks against the idea touted by PredPol's proponents that fewer inputs necessarily leads to a lesser degree of bias, and supports the use of demographic metrics to counter biases in crime data.

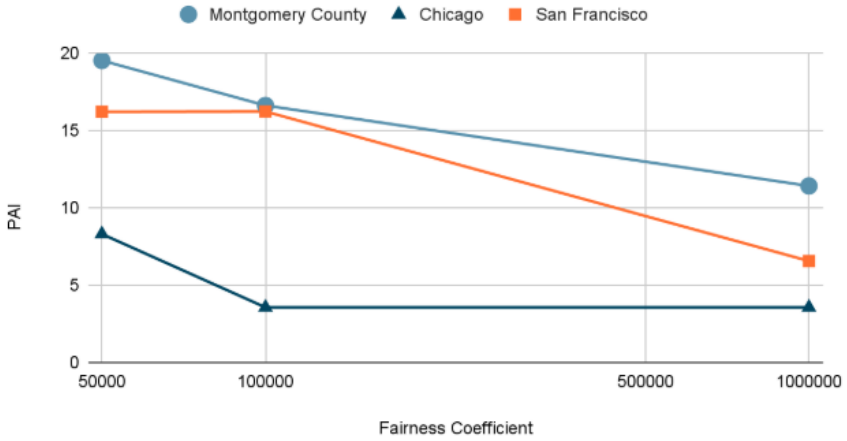
4.2: Impact of Fairness on Accuracy

Having demonstrated that our model is able to outperform PredPol, this section shows our attempt to estimate the relationship between fairness penalization and accuracy. Accuracy falloff is relatively slow, and it is not entirely clear what this would mean in practice.

4.2.1: Predictive Accuracy and Fairness as a Trade-Off

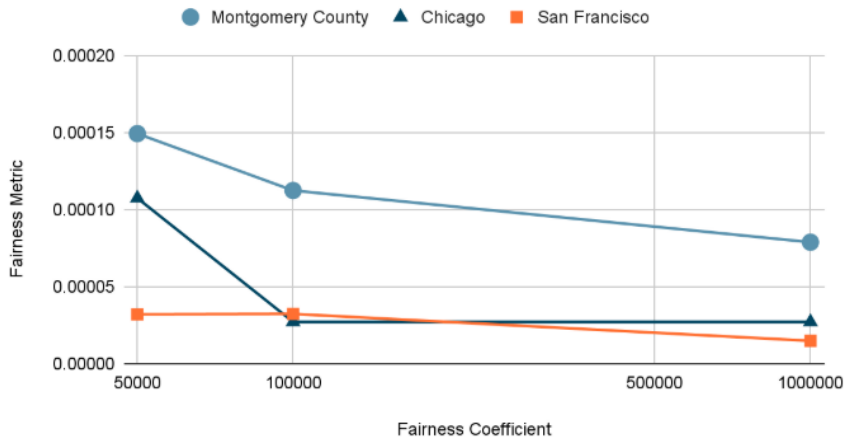
To test the relationship between increased fairness with accuracy falloff, we trained multiple models for each city, each with different weighting λ of the fairness metric. Due to computational limitations, we were only able to train two additional models. For the first model, we doubled our initial $\lambda \neq 5 \cdot 10^4$, and for the second we increased it by an order of magnitude. Figure 3 shows a plot (with x-axis logged) of the relationship between PAI and the fairness coefficient.

Figure 3: PAI vs. Fairness Coefficient



Unsurprisingly, PAI decreases as more weight is assigned to demographic parity. The fairness metric analogously decreases as the weight is assigned, as seen in Figure 4. In Montgomery County, doubling the coefficient led to an accuracy falloff of 15%, from 19.5 down to 16.6. The order of magnitude increase led to a falloff of 41%. While our model outperformed PredPol even with the fairness metric, it remains the case that within our own model, there may be a tradeoff between fairness and predictive power when using retrospective crime data. However, the relevance of this is unclear, as shall be discussed.

Figure 4: Fairness Metric vs. Fairness Coefficient



4.2.2: Discussion of Predictive Accuracy and Fairness as a Trade-Off

The results point to the idea that in retrospective studies of this type, we may see tradeoffs between predictive accuracy and fairness. This tradeoff is presumably why, despite being made for PredPol specifically, this metric was never implemented. In an interview with Gizmodo, then-CEO Brian MacDonald was quoted as saying that implementation of the fairness penalty "would reduce the protection provided to vulnerable neighborhoods with the highest victimization rates" (Lash, 2021). However, while it is tempting to read these results as proving that there is a necessary tradeoff after all, this is not necessarily justified from the data. Instead, it points to a fundamental limitation in our research, that being its nature as a retrospective analysis.

One of the most important issues with predictive policing, algorithmic or otherwise, is the possibility of feedback loops (Ensign et al., 2017). When police are sent to a specific area they will, of course, report more crime in that area, which is then fed back into the system. The upshot of this is that, with any dataset we use, the actual observed crime is a product of whatever methods the policing methods that department is using. The picture of crime that emerges after the implementation of an algorithm at the level of a department is much different than what we can observe from retrospective data. This much is clear

with PredPol itself; its implementation in Los Angeles resulted in a 7.4% reduction in crime compared to the control condition (Mohler et al., 2015). As such, it isn't entirely clear that the relationship we see in retrospective data would hold if this algorithm were to actually be used by police departments. This presents a rich opportunity for further research, but should also temper any conclusions drawn about the exact relationship between fairness penalties and decreased accuracy.

Chapter 5: Impact and Limitations

5.1: Equity-Impact Report

Our project is rooted in the pursuit of justice and equity within predictive policing (PP) systems. Recognizing that the existing deployment of PP algorithms often reinforces historical inequities – particularly along racial and socioeconomic lines – our work directly addresses the disproportionate impacts of such systems on marginalized communities.

Communities with larger Black and Hispanic populations are routinely over-policed due to entrenched feedback loops in crime data collection. These communities, often designated as Hot Spots, bear the brunt of heightened police surveillance and discretionary enforcement, which contributes to inflated arrest statistics and further justifies continued policing presence. This cyclical process entrenches racial disparities and erodes community trust (see Section 2.1.2). Our project aims to disrupt this dynamic by building an algorithm that actively corrects for historical and structural bias in crime data.

To this end, our algorithm includes an explicit fairness metric, balancing predictive accuracy with demographic parity. Unlike previous models, ours does not assume that inputs perfectly reflect the reality of crime. While far from a perfect solution, building fairness directly into the algorithm helps to mitigate the effect of those biased inputs, which may in turn send police to areas where they are needed more.

5.2: Risks and Limitations

We acknowledge that even well-intentioned algorithms can produce unintended harm. Over-correcting for bias may lead to the under-policing of areas with legitimate safety concerns or the failure to identify serious crimes, a tradeoff well-documented in fairness-accuracy research (Kleinberg et al., 2016; Lash, 2021). Furthermore, excessive fairness penalties may reduce predictive accuracy, potentially compromising public safety. While it remains inconclusive if this reduced accuracy would continue to manifest in field trials, it remains a hard sell for police departments that are concerned about their public image.

Conversely, algorithms – even those designed to mitigate bias – remain vulnerable to hidden proxies for race or socioeconomic status. These proxies may be statistically neutral yet still yield racially disparate outcomes if the training data reflect systemic inequities (Corbett-Davies et al., 2024; Yeom et al., 2018). This means that without deliberate detection and correction, historical injustices can be systemized and perpetuated through predictive outputs.

We recognize that there exists limitations in our own approach. Our algorithmic approach

focuses solely on police resource allocation, one small component to a broader policing system. While we have aimed to make this component more equitable, HSP – algorithmic or otherwise – operates in tandem with broader tactics and institutions that carry their own biases. As noted by Rosenbaum (2009), improving resource allocation alone does not reform the fundamental structures of policing, nor does it address deeper issues like adversarial police culture or structural racism. This result bears out empirically. Braga et al. (2019b) found that the crime reduction effects of hot spots policing was significantly moderated by the type of policing used in those hot spots. Problem-oriented policing, a policing method which emphasizes solving underlying causes of crime (RAND Corporation, n.d.), was significantly more effective in conjunction with hot spots policing than was increasing traditional police presence (Braga et al., 2019b). This implies that, no matter how fair our algorithm may be, it cannot be viewed as a panacea. Efforts to reduce systemic racism should not stop here. While fairness in predictive policing can serve as one small mitigating variable, the underlying structure of policing is still discriminatory and needs to be rethought. Pushback on other fronts will likely have more significant effects than what our algorithm is able to do.

References

[1] ACLU-DC & ACLU Analytics. (2020). RACIAL DISPARITIES IN STOPS BY THE D.C. METROPOLITAN POLICE DEPARTMENT: REVIEW OF FIVE MONTHS OF DATA. ACLU. https://www.acludc.org/sites/default/files/2020_06_15_aclu_stops_report_final.pdf

[2] Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016, May 23). Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Ansfield, B. (2020). The Broken Windows of the Bronx: Putting the Theory in Its Place. *American Quarterly, 72*(1), 103–127. https://doi.org/10.1353/aq.2020.0005

[4] Barnes, S. F. (2018). *Police-Community Relations: A Study of Racial Disparity and the Effects of Hot Spots Policing Leadership Strategies* [Ph.D., North Carolina Agricultural and Technical State University]. https://www.proquest.com/docview/2050744039/abstract/4ECE5095DEAE4481PQ/1

[5] Bennett Moses, L., & Chan, J. (2018). Algorithmic prediction in policing: Assumptions, evaluation, and accountability. *Policing and Society, 28*(7), 806–822. https://doi.org/10.1080/10439463.2016.1253695

[6] Braga, A. A., Turchan, B., Papachristos, A. V., & Hureau, D. M. (2019a). Hot spots policing of small geographic areas effects on crime. *Campbell Systematic Reviews, 15*(3), e1046. https://doi.org/10.1002/cl2.1046

[7] Braga, A. A., Turchan, B. S., Papachristos, A. V., & Hureau, D. M. (2019b). Hot spots policing and crime reduction: An update of an ongoing systematic review and meta-analysis. *Journal of Experimental Criminology, 15*(3), 289–311. https://doi.org/10.1007/s11292-019-09372-3

[8] Braga, A. A., Welsh, B. C., & Schnell, C. (2015). Can Policing Disorder Reduce Crime? A Systematic Review and Meta-analysis. *Journal of Research in Crime and Delinquency, 52*(4), 567–588. https://doi.org/10.1177/0022427815576576

[9] 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, 5*(1), 1–6. https://doi.org/10.1080/2330443X.2018.1438940

[10] Bratton, W. J. (1997, August 19). Opinion | New York's Police Should Not Retreat. *The New York Times*. https://www.nytimes.com/1997/08/19/opinion/new-york-s-police-should-not-retreat.html

[11] Chainey, S., Tompson, L., & Uhlig, S. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal, 21*(1), 4–28. https://doi.org/10.1057/palgrave.sj.8350066

[12] Chauhan, P., Fera, A., Walsh, M., Balazon, E., Misshula, E., Travis, J., Citizens Crime Commission of New York City, & John Jay College of Criminal Justice. (2014). *Trends in misdemeanor arrest rates in New York* (Reports from John Jay College). John Jay College. https://academicworks.cuny.edu/jj_arch_pubs/44

[13] City of Chicago. (2024). Crimes—2001 to Present | City of Chicago | Data Portal [Dataset]. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

[14] Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2024). The Measure and Mismeasure of Fairness. *J. Mach. Learn. Res., 24*(1), 312:14730-312:14846.

[15] dataMontgomery. (2024). Crime | Open Data Portal [Dataset]. https://data.montgomerycountymd.gov/Public-Safety/Crime/icn6-v9z3/data_preview

[16] DataSF. (2025). Police Department Incident Reports: 2018 to Present | DataSF [Dataset]. https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783/about_data

[17] Delua, J. (2021, March 12). Supervised vs. Unsupervised Learning: What's the Difference? Supervised vs. Unsupervised Learning: What's the Difference? https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

[18] Dike, H. U., Zhou, Y., Deveerasetty, K. K., & Wu, Q. (2018). Unsupervised Learning Based On Artificial Neural Network: A Review. *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 322–327. https://doi.org/10.1109/CBS.2018.8612259

[19] Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

[20] Eck, J. E., Chainey, S., Cameron, J. G., Leitner, M., & Wilson, R. E. (2005). *Mapping Crime: Understanding Hot Spots* (p. 79). U.S. Department of Justice.

[21] Elgart, A. S., Alcaraz, D. C., Jackson, A., Koshy, T., Martin-Walton, A., Micklethwaite, K., & Rick, A. (2022). *RACIAL AND IDENTITY PROFILING ADVISORY (RIPA) BOARD* (p. 279).

[22] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). *Runaway Feedback Loops in Predictive Policing* (No. arXiv:1706.09847). arXiv. https://doi.org/10.48550/arXiv.1706.09847

[23] Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). *AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data* (No. arXiv:2003.06505). arXiv. https://doi.org/10.48550/arXiv.2003.06505

[24] Gelman, A., Kiss, A., & Fagan, J. (2005). An Analysis of the NYPD's Stop-and-Frisk Policy in the Context of Claims of Racial Bias. *Columbia Public Law Research Paper No. 05-95*. https://scholarship.law.columbia.edu/faculty_scholarship/1390

[25] Greene, J. A. (1999). Zero Tolerance: A Case Study of Police Policies and Practices in New York City. *Crime & Delinquency, 45*(2), 171–187. https://doi.org/10.1177/0011128799045002001

[26] Justice Policy Institute. (2007). *The Vortex: The Concentrated Racial Impact of Drug Imprisonment and the Characteristics of Punitive Counties* (p. 36). Justice Policy Institute. https://justicepolicy.org/wp-content/uploads/justicepolicy/documents/vortex.pdf

[27] Kelling, G. L., & Wilson, J. Q. (1982, March 1). Broken Windows. *The Atlantic*, March 1982. https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/

[28] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores*. arXiv:1609.05807 [Cs, Stat]. http://arxiv.org/abs/1609.05807

[29] Lash, A. S., Dhruv Mehrotra, Surya Mattu, Dell Cameron, Annie Gilbertson, Daniel Lempres, and Josh. (2021, December 2). Crime Prediction Software Promised to Be Bias-Free. New Data Shows It Perpetuates It. *Gizmodo*. https://gizmodo.com/crime-prediction-software-promised-to-be-free-of-biases-1848138977

[30] Lee, J. M., Steinberg, L., & Piquero, A. R. (2010). Ethnic identity and attitudes toward the police among African American juvenile offenders. *Journal of Criminal Justice, 38*(4), 781–789. https://doi.org/10.1016/j.jcrimjus.2010.05.005

[31] Lum, K., & Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

[32] Mastrofski, S. D., & Fridell, L. (2011). *Police Departments' Adoption of Innovative Practices* (p. 5).

[33] Mohler, G., Raje, R., Carter, J., Valasik, M., & Brantingham, J. (2018). A Penalized Likelihood Method for Balancing Accuracy and Fairness in Predictive Policing. *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2454–2459. https://doi.org/10.1109/SMC.2018.00421

[34] Mohler, G., Short, M. B., Brantingham, P. 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. https://doi.org/10.1198/jasa.2011.ap09546

[35] Mohler, G., 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. https://doi.org/10.1080/01621459.2015.1077710

[36] Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., & Goel, S. (2020). A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour, 4*(7), 736–745. https://doi.org/10.1038/s41562-020-0858-1

[37] RAND Corporation. (n.d.). Problem-Oriented Policing. Retrieved March 31, 2025, from https://www.rand.org/pubs/tools/TL261/better-policing-toolkit/all-strategies/problem-oriented-policing.html

[38] Reaves, B. A. (2011). *Local Police Departments (2007)*. DIANE Publishing.

[39] Rosenbaum, D. P. (2009). The limits of hot spots policing. In *Police Innovation: Contrasting Perspectives* (p. 20). Cambridge University Press. https://www.researchgate.net/publication/293120048_Critic_The_limits_of_hot_spots_policing

[40] Santos, R. B. (2017). Identifying Meaningful and Useful Patterns. In *Crime analysis with crime mapping* (Fourth edition). SAGE Publications, Inc.

[41] Shah, K. (2024, December 6). What Is Off-the-Shelf AI? Bloomreach. https://www.bloomreach.com/en/blog/off-the-shelf-ai

[42] Swencionis, J. K., & Goff, P. A. (2017). The psychological science of racial bias and policing. *Psychology, Public Policy, and Law, 23*, 398–409. https://doi.org/10.1037/law0000130

[43] Taylor, R. B. (2019). *Breaking Away from Broken Windows: Baltimore Neighborhoods and the Nationwide Fight Against Crime, Grime, Fear, and Decline*. Routledge. https://doi.org/10.4324/9780429502019

[44] Thompson, J. P. (2015). Broken Policing: The Origins of the "Broken Windows" Policy. *New Labor Forum, 24*(2), 42–47. https://doi.org/10.1177/1095796015579993

[45] U.S. Census Bureau. (2020). HISPANIC OR LATINO, AND NOT HISPANIC OR LATINO BY RACE. Decennial Census, DEC 118th Congressional District Summary File, Table P9 [Dataset]. https://data.census.gov/table/DECENNIALCD1182020.P9?q=P9+race

[46] Weisburd, D., & Majmundar, M. K. (Eds.) (with Committee on Proactive Policing, Committee on Law and Justice, Division of Behavioral and Social Sciences and Education, & National Academies of Sciences, Engineering, and Medicine). (2018). *Proactive Policing: Effects on Crime and Communities* (p. 24928). National Academies Press. https://doi.org/10.17226/24928

[47] Wong, T. (2022). THE MATHEMATICS OF POLICING. https://www.semanticscholar.org/paper/THE-MATHEMATICS-OF-POLICING-Wong/c29724d3a3139d9d0caa1f30bdd606329ddc2023

[48] Yang, S.-M. (2010). Assessing the Spatial–Temporal Relationship Between Disorder and Violence. *Journal of Quantitative Criminology, 26*(1), 139–163. https://doi.org/10.1007/s10940-009-9085-7

[49] Yeom, S., Datta, A., & Fredrikson, M. (2018). *Hunting for Discriminatory Proxies in Linear Regression Models* (No. arXiv:1810.07155). arXiv. https://doi.org/10.48550/arXiv.1810.07155