An Ethical and Technical Evaluation of the Use of Machine Learning Models in Health and Human Services: A Case Study of the Allegheny Family Screening Tool

An Ethical and Technical Evaluation of the Use of Machine Learning Models in Health and Human Services: A Case Study of the Allegheny Family Screening Tool

A thesis presented by
Samuel Oh

to
The Department of Computer Science and the Department of Philosophy
in partial fulfillment of the requirements
for the joint degree of Bachelor of Arts

Harvard College
Cambridge, Massachusetts
December 7, 2019

Acknowledgments

The completion of this work would not have been possible without the support and guidance of family and friends. I would like to thank David Gray-Grant for hours of editing sessions, providing extensive feedback on every part of this paper in addition to emotional support in the writing process. I would like to thank Elena Glassman for adopting me into her lab and providing suggestions for related readings. I am thankful to Jim Waldo for encouragement throughout the process and help getting through the initial ideation phase. All of the support and feedback I have received was critical in finishing this work.

I am also thankful to friends for helping read and edit the paper. Thank you to Bilal Nadeem for help with citations and editing, Caleb Oh, Amany Belay, and Allen Lai for reading and providing edits on drafts of the paper.

Abstract

In this paper, I will conduct a case study on the Allegheny Family Screening Tool (AFST), a risk assessment tool used in child protection services in Allegheny County, Pennsylvania. First, I will review the implementation, use, and impact of the AFST on screening decisions from both a technical and ethical standpoint. Then, I will consider two points of contention in the public debate surrounding the tool: a charge of discrimination against the tool and a confusion about the purpose of the tool. I argue that the tool is not wrongfully discriminatory and that the underlying design of the tool is flawed and must be changed in order to increase accuracy and consistency of screening decisions.

Table of Contents

Chapter 1: Background

1. Predictive Risk Models in Health and Human Services

The adoption of new machine learning techniques has promised to revolutionize a diverse set of fields, from healthcare to finance to the distribution of welfare. Advocates of the adoption of machine learning techniques in different industries point to its ability to offer accurate results while reducing human bias in making decisions.[1] The original thought is reasonable: if a machine is given the ability to make a decision based on a set of training data, then human bias would be eliminated in the decision-making process. In practice, however, human biases can still enter the algorithm at many different points, including the setup and training of the algorithm and the human-computer interaction involved with the use of the program.

The machine learning model is created by first taking in a set of training data, finding statistical relationships between the training data and a chosen target variable, and then using these relationships to guess the likely outcome of future, unknown cases. The basis of machine learning is the process of creating a model that “learns” from previously known data and predicts results for future data based on statistical relationships. Machine learning models can assist in any situation in which a human must make a decision about a future outcome.

Given the promise of improved decision-making by harnessing large amounts of available data, the use of machine learning is becoming more popular in healthcare. In

[1] Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* 542.7639 (2017): 115.

healthcare, a machine learning model that is used to generate a risk score for the occurence of an adverse event is called a Predictive Risk Model (PRM).[2]

In many ways, healthcare is a good candidate for the early adoption of these systems. Many healthcare centers have access to vast amounts of data without a way to use all of the data to make the best possible choice about a patient's health. Moreover, the decisions that healthcare professionals must make have potentially fatal consequences, and any improvement in their ability to make these decisions that rely on their prediction of the occurrence of an adverse event is valuable. In New Zealand and Australia, machine learning models are already in use to assist in predicting the likelihood of the occurrence of a disease.[3]

Accurate information about the risk of certain diseases can help national health organizations better allocate their limited resources. At Kaiser Permanente in the United States and the National Health Service in the United Kingdom, these models are used to determine an individual's risk of future emergency hospitalization.[4] In each case, the PRM promises to improve the decision making skills of a human by allowing the human to analyze more data in a timely manner than would be possible without the PRM.

Despite the promise of PRMs, healthcare organizations must consider many of the ethical implications tied to the adoption of PRMs, including the repercussions of mistakes. Organizations must consider the quality of professional training and proper understanding of the PRM, the impact of the PRM's adoption on disparate impact along racial or socioeconomic lines, and the difficulties of obtaining meaningful consent, among other concerns. Beyond the accuracy of the PRM, a central ethical consideration involved with its adoption in healthcare

[2] Vaithianathan et al. "Developing predictive models to support child maltreatment hotline screening decisions." Center for Social Data Analytics (2017).

[3] Panattoni et al. "Predictive risk modelling in health: options for New Zealand and Australia." *Australian Health Review* 35.1 (2011): 45-51.

[4] Ibid, 46.

involves the proper user understanding of the limits and uses of the risk score generated by the PRMs. Unrealistic user expectations of the risk scores or complete aversion of the risk assessment system both might lead to fluctuations in accuracy and consistency of the system.

In this paper, I will conduct a case study on the Allegheny Family Screening Tool (AFST), a risk assessment tool used in child protection services in Allegheny County, Pennsylvania. In the first chapter, I will introduce the context of the tool by reviewing the context of call screening decisions, the implementation of the AFST, and the impact that the AFST has had on screening decisions from both a technical and ethical standpoint. In the second chapter I construct a precise philosophical charge of discrimination against the AFST and consider whether the AFST is wrongfully discriminatory. In the final chapter, I focus on the technical design of the tool and offer suggestions for the AFST to improve call screening decisions.

2. Allegheny County DHS

The Allegheny County Department of Human Services (DHS) houses the Office of Children, Youth, and Families (CYF). One of the roles of the CYF is to operate a call center for their local child neglect and abuse hotline. Call screeners in the center receive a call from an individual, either from the community or a mandated reporter, about potential maltreatment occurring in a family. The call screener must then search their internal database for more information about the family. After examining the available information, the call screener assigns a safety rating to the family before they come to a decision about whether to screen the family *in* for a formal investigation to determine if maltreatment has occurred and there is potential future harm for a child, or to screen the family *out* without any further evaluation or assessment.[5]

[5] https://www.alleghenycounty.us/human-services/index.aspx

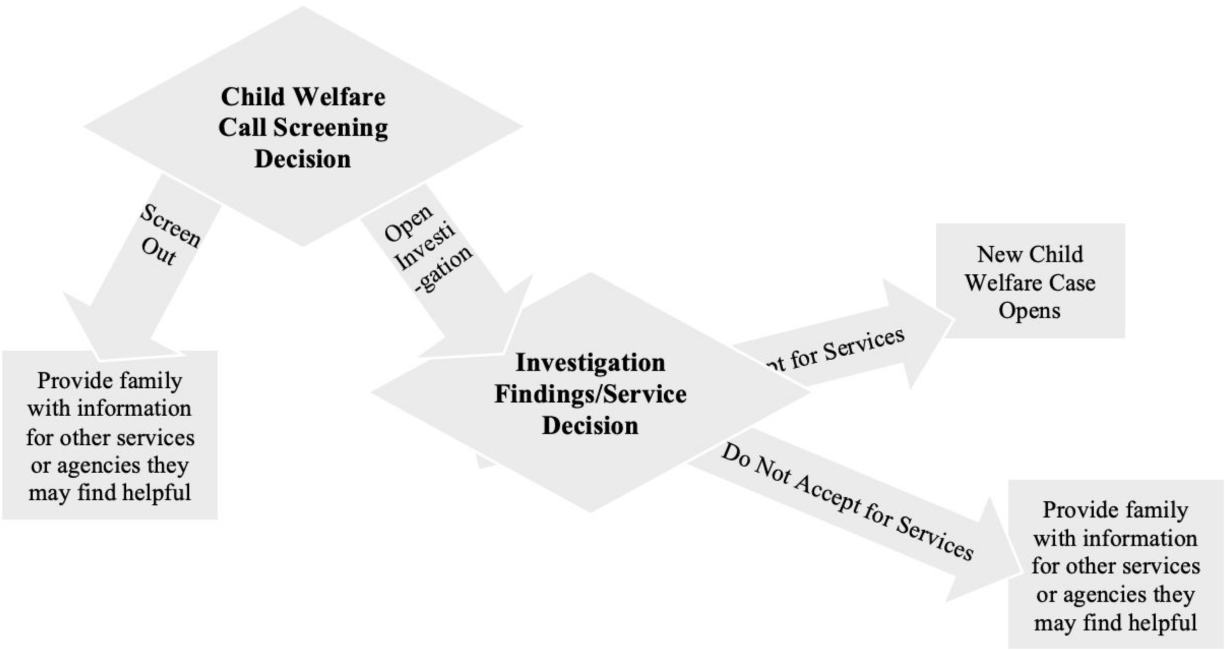

Child Welfare Call Screening Decision
Screen Out
Open Investi -gation
New Child Welfare Case Opens
nt for Services
Provide family with information for other services or agencies they may find helpful
Investigation Findings/Service Decision
Do Not Accept for Services
Provide family with information for other services or agencies they may find helpful

Figure 1: Screening Decisions [6]

In 1999 the Allegheny County DHS created its own data center to keep public records. Since then, they have collected over one billion electronic records from different public services. [7] These records include historical and cross-sector administrative data over twenty years from child protective services, mental health services, drug and alcohol services, homeless services, and many other related public services. Call screeners have access to a vast amount of data when they search the internal database for information about a family.[8]

The availability of this large amount of data leads to two potential issues: (1) call screeners cannot efficiently access and review all available records in a meaningful and timely manner, and (2) there may be a lack of consistency in the ways that a call screener weighs certain variables. In the first case, call screeners may not have the time to search through all of

[6] Retrieved from Vaithianathan et al. "Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation."

[7] Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press, 2018, 108-9.

[8] Ibid.

the relevant records. In the second case, consider a variable such as the recent involvement of a parent in a criminal case. This variable might be weighted heavily by one call screener but totally missed by another. These two potential issues suggest that the power of the data is not being effectively harnessed to make the most informed decision about the future.

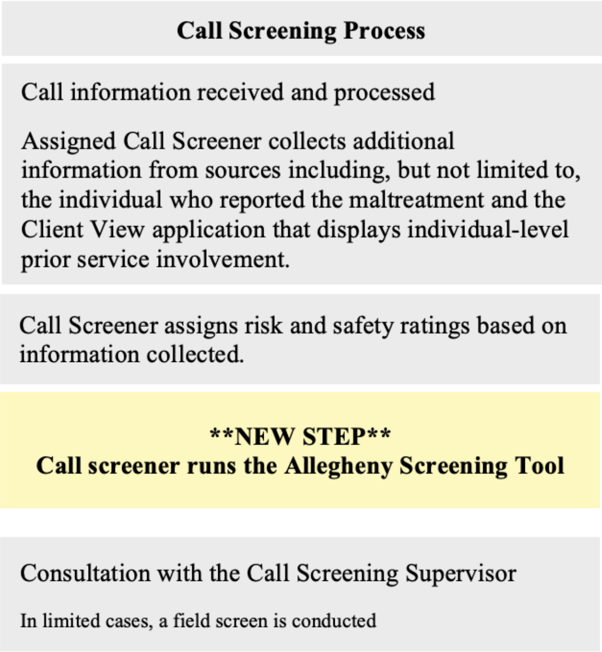
With these issues in mind, the DHS offered a contract to a team of researchers from New Zealand and California to create a PRM to help analyze the vast amount of available data to help call screeners make better choices. Economist Rhema Vaithianathan of the Auckland University of Technology partnered with Emily Putnam-Hornstein, the director of the Children's Data network at the University of Southern California, to implement the Allegheny Family Screening Tool (AFST) which uses machine learning techniques to predict the likelihood that a child will experience abuse or neglect in the future.

The purpose of the AFST is not to replace the call screener, but rather to augment the screener's ability to make the most informed decision by analyzing the data and offering a summarized risk score of findings. As such, the AFST is a tool that must be used in conjunction with the human process to assist the human in making the most informed screening decision.

The DHS began use of the AFST in August 2016 to assist the decisions made by child welfare call screeners in the call center for the child neglect and abuse hotline. The call screening process was kept mostly the same, with an additional step in the information gathering phase where the call screener uses the AFST to calculate a risk score for the family. After call screeners receive the call from the community, they still gather information from the caller as well as search the internal database for prior involvement with the CYF.

The call screener then assigns their own safety rating on the family, as they did before. At this point, however, there is an additional step in which the call screener runs the AFST to see the risk score calculated based on the predictive model. This latter risk score is only intended to be used as an additional source of information. Finally, the call screener makes a final decision about whether to screen-in the family or not, then consults with the Call Screening Supervisor before making the final decision.

Figure 2: Referral Progression Process[9]



Call Screening Process

Call information received and processed

Assigned Call Screener collects additional information from sources including, but not limited to, the individual who reported the maltreatment and the Client View application that displays individual-level prior service involvement.

Call Screener assigns risk and safety ratings based on information collected.

****NEW STEP****
Call screener runs the Allegheny Screening Tool

Consultation with the Call Screening Supervisor

In limited cases, a field screen is conducted

Figure 2 captures the updated workflow for the call screener, with the placement of the additional step of calculating the AFST score highlighted in the yellow box. The creators of the tool explicitly place the AFST after a call screener has come up with her own independent risk score so that the AFST does not unduly influence the call screener’s own judgement. The idea behind this design choice is to retain the call screener’s ability to independently come up with their own judgments.

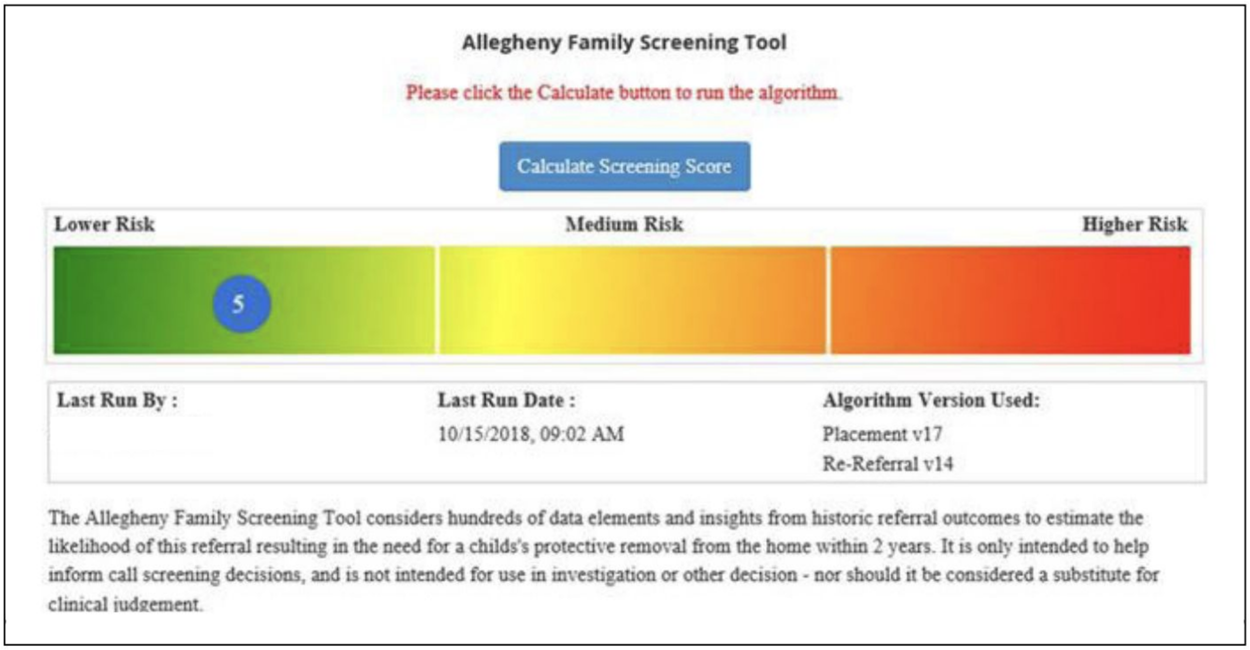
3. The AFST Risk Score

The AFST is a predictive tool that assesses a child’s risk of needing services. The score itself is a number between 1 and 20 where 1 is the lowest risk and 20 is the highest risk. The

[9] Retrieved from Vaithianathan et al. "Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation."

score for a family is determined by the maximum risk score given across all of the children in the family. So, if there are four children, each with a potentially unique score, the ultimate score shown to call screeners for the family will be the maximum score among the four children. For each family, a single risk score is shown in the manner of Figure 3, in the form of a thermometer figure that includes a colored scale.

Figure 3: AFST Presentation[10]



Allegheny Family Screening Tool

Please click the Calculate button to run the algorithm.

Last Run By :	**Last Run Date :** 10/15/2018, 09:02 AM	**Algorithm Version Used:** Placement v17 Re-Referral v14

The Allegheny Family Screening Tool considers hundreds of data elements and insights from historic referral outcomes to estimate the likelihood of this referral resulting in the need for a childs's protective removal from the home within 2 years. It is only intended to help inform call screening decisions, and is not intended for use in investigation or other decision - nor should it be considered a substitute for clinical judgement.

Figure 3 depicts the physical manifestation of the tool that the call screeners see. After assigning their own independent risk and safety ratings, call screeners click the blue “Calculate Screening Score” button in the center of the figure. Once clicked, the image renders a blue dot with a score that is situated between the two ends of the thermometer gauge, with green indicating lowest risk and red indicating highest risk. Below the thermometer itself, there is additional information that keeps track of the call screener who last ran the tool and a timestamp for when it was run to help screeners keep track of when they ran the tool. In addition, there is an indication of the versions of

[10] Retrieved from Vaithianathan et al. "Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation."

the underlying algorithms used, helping the call screener understand whether the tool is using the most up-to-date models in its calculation of the score.

Finally, there is a paragraph of static text at the bottom of the screen offering a reminder of the indicated purpose of the machine as an information aid rather than as a substitute for independent clinical judgement. This reminder is meant to help screeners understand the role of the machine in the context of the overall decision to investigate. The text also includes a high level description of how the AFST calculates its score, alluding to “hundreds of data elements and insights from historic referral outcomes to estimate the likelihood of this referral resulting in the need for a child’s protective removal from the home within 2 years.”

This text reminds call screeners of two important aspects of the AFST. First, the tool is built up of statistical models that gather data from historical records. That is, the scores of the tool itself are derived from an underlying statistical model. Second, and relatedly, the score itself estimates the likelihood that a specific referral will result in action on the family if an investigation is carried out. That is, the score is not directly measuring risk of child abuse, but rather measuring the likelihood that a decision to investigate based on a specific referral will lead to action. Both of these aspects of the reminder raise an important question about what the AFST is actually measuring and what the score means.

The AFST itself is a tool that constitutes a risk score. Although the risk score is meant to measure the likelihood of a child experiencing abuse or extreme neglect in the future, these events cannot be measured directly because of a lack of cases of recorded abuse coupled with an ambiguity between extreme neglect and neglect.[11] As a result, the risk score is an indirect

[11] Ibid., 9.

measurement of child risk determined through an underlying proxy statistical model, the placement model.[12]

The placement model measures the likelihood that a child experiences a safety issue that requires action on the child's behalf, such as removing the child from their home into a safer setting within two years following a referral.[13] In essence, the placement model is predicting the likely behavior of the investigators themselves. That is, if a call screener decides to screen a family in for investigation, what is the likely outcome of the investigation? By providing a prediction of the outcome of an investigation, the placement model offers a proxy for child mistreatment.

4. Independent Evaluations

The DHS has publicly released a set of three independent evaluations of the tool, in addition to a technical design methodology written by the creators, as part of their commitment to proceed with the adoption of the AFST in a transparent manner. The three evaluations consist of a process evaluation, an evaluation of the impact of the tool in use, and an ethical review of the tool.[14] I will address the main findings of each of these evaluations starting with the process evaluation, then moving to the impact evaluation before finally considering the ethical analysis.

[12] The AFST used to consist of two underlying models. But, with the publication of the methodology for Version 2 of the AFST in April 2019, one of the models, the re-referral model, was removed leaving only the placement model.

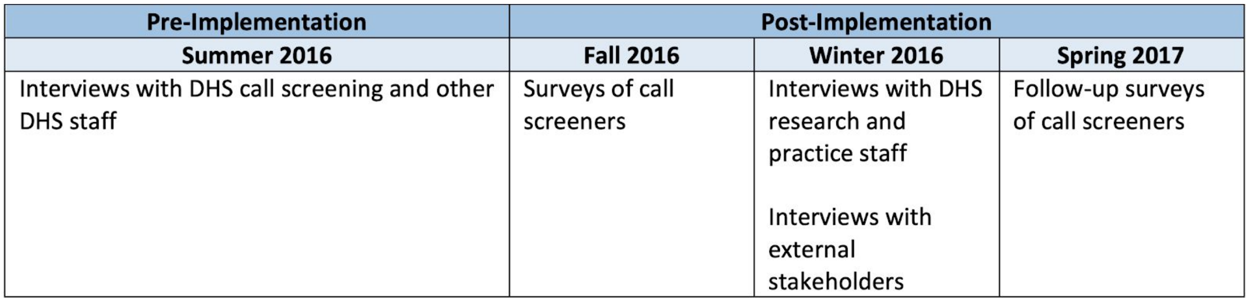
[13] Vaithianathan et al., "Developing predictive models to support child maltreatment hotline screening decisions" 8.

[14] These publications were all published within the last three years https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx

4.1 Allegheny County Predictive Risk Modeling Tool Implementation: Process Evaluation

Hornby Zeller Associates, a public-sector consulting firm, examined the process by which the tool was implemented, analyzing the tool's impact on the call screening experience, the practice and policy implications of its use, and perceptions and reactions to the tool. The study was carried out through interviews of call screening staff before and after evaluation following the schedule shown in the following table.

Figure 4: Schedule of Methodology[15]



Pre-Implementation	Post-Implementation		
Summer 2016	Fall 2016	Winter 2016	Spring 2017
Interviews with DHS call screening and other DHS staff	Surveys of call screeners	Interviews with DHS research and practice staff Interviews with external stakeholders	Follow-up surveys of call screeners

With regards to the call screeners, the process evaluation came to two major conclusions. First, the study concluded that call screening staff report that they have a good understanding of the AFST and find the tool easy to use. The majority of call screeners claim that they understand how the score works, with 100% of participants agreeing that they are "adequately prepared to use the tool."[16] The results point to an overwhelming confidence in call screeners' perception of their own ability to use the tool.

[15] Retrieved from Hornby Zeller Associates. "Allegheny County Predictive Risk Modeling Tool Implementation: Process Evaluation." 2018.

[16] Ibid. 11.

A second major conclusion of the study is that despite the apparent understanding, call screening staff vary greatly in their actual use of the AFST scores. The survey of call screeners two months after implementation of the AFST found a wide distribution of the use of the AFST, with around 40% of call screeners using the tool to inform their decisions on a consistent basis and around 31% of call screeners reporting they rarely use the tool, if at all.[17] Despite widespread understanding of the use of the tool, call screeners seem to differ greatly in their actual use of the tool in practice. I will explore these findings further in chapter three.

4.2 Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office

In April 2019, Jeremy D. Goldhaber-Fiebert and Lea Prince, both faculty at Stanford University, published a report on the impact of the implementation of the AFST on call screening decisions. The study focused on the AFST's impact on the accuracy of call screening decisions as well as the consistency of screening decisions, both of which are the stated goals of the AFST. [18] The data was collected in a period spanning August 1, 2013 through May 31, 2018, comparing the years prior to implementation of the tool with the two years following its adoption.

In terms of accuracy, the study concludes that the implementation of the AFST increased the accuracy of families correctly screened-in for an investigation and slightly decreased the accuracy of families screened-out for an investigation. In the first case, the researchers found that a higher proportion of children screened-in for investigation result in an investigation that had further action taken. On the other hand, the researchers also found a slight decrease in the proportion of children who are screened-out and had no re-referrals within 60 days.[19] In other

[17] Ibid., 15.
[18] Goldhaber-Fiebert and Lea. "Impact Evaluation." Pittsburgh: Allegheny County. (2019), 4-5.
[19] Ibid., 2.

words, the accuracy of decisions to screen-in families increased while the accuracy of decisions to screen-out families decreased.

In terms of consistency, the study found no significant change in the outcomes across call screeners. That is, the researchers reported no increase or decrease in the consistency of the decisions made by call screeners between the pre-AFST and post-AFST implementation periods. However, the researchers qualify this finding with the note that there was likely an insufficient sample size to detect any minor changes.[20]

4.3 Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County

In addition to the impact and process evaluations, the DHS commissioned a third party to independently conduct an ethical review of the adoption of the tool. The researchers considered potential ethical issues that could arise from the adoption of the tool, including issues of consent for the use of information, stigmatization of families, racial disparity, professional competency and training, and the importance of ongoing monitoring.[21] For each of these issues, the researchers reasoned through the importance of certain uses and understandings of the AFST, arguing that the AFST can be used in an ethical manner.

By the end of the report, the researchers concluded that it would actually be unethical to refrain from using the AFST because of the potential for the AFST to increase the accuracy of decisions made. Given the potential for an increase in accuracy and a decrease in stigmatization that might come with the adoption of the AFST, the researchers argued that the adoption of the tool is actually morally necessary. To justify this claim, the ethicists pointed to

[20] Ibid., 2.

[21] Dare and Gambrill. “Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County.” Apr. 2017, 2-7.

the necessity of making child screening decisions and the central importance of minimizing errors, both of which can be harmful to families.[22]

5. Controversy with AFST

Despite the DHS's efforts at full transparency with the creation and adoption of the AFST, the tool has still been the subject of public criticism. One vocal critic, Virginia Eubanks, published a book called *Automating Inequality* which devotes a chapter to the various issues with the AFST. Eubanks offers a sociological overview of the complex issues, providing insightful anecdotes from interviews with call screeners at the county hotline, staff at the DHS in charge of the AFST, and families that have interacted with the Allegheny County CYF.[23]

Eubanks is wide-reaching in her critisisms, offering a general overview of the main issues in public discourse about the AFST including questions of discrimination, accuracy, and biases. Her book was generally well-received by the community as seen by its reception of numerous book awards and honorable mentions like the McGannon Center Book Prize and the Goddard Riverside Stephan Russo Book Prize for Social Justice. Moreover, her criticisms elicited a direct response from the DHS, which offers even greater insight into the nature of the debate.

Given the general reception of and reaction to her book, the issues raised in the book offer a valuable starting point for an ethical analysis of the AFST based on public controversy. By clarifying, structuring, and building upon the issues raised by Virginia Eubanks and mediating these with responses from the DHS, we can better understand the underlying questions that

[22] Ibid.

[23] Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press, 2018, 108-9.

drive the debate. My task will be to test the strength of Eubanks' claims, arguing that Eubanks' claims are not as convincing as they seem on further inspection.

The two charges I will focus on are (1) a charge of discrimination against the AFST, and (2) a charge of improper usage of the tool. In the first case, I consider Eubanks' general argument that the AFST unfairly discriminates against poor families before constructing a philosophically rigorous account of the conditions for wrongful discrimination. Then, I consider what it means for a machine like the AFST to discriminate before evaluating whether or not the AFST fulfills the conditions of wrongful discrimination. Ultimately, I argue that the AFST does not wrongfully discriminate against poor families.

In the second case, I take a more technical approach, exploring the related computer science literature to help clarify an underlying flaw in the design of the AFST as an information aid. Again, I start by constructing Eubanks' general argument that call screeners are not using the tool as intended to set the context of the debate. Then, I argue that one of the reasons for the variation in call screeners' use of the AFST is an underlying confusion between the actual design of the AFST and its intended purpose as an information aid. Ultimately, I argue that the AFST should be completely redesigned to serve its intended function as an information aid.

Chapter 2: Discrimination and the AFST

In this chapter, I will consider Virginia Eubanks' argument that the AFST wrongfully discriminates against certain groups of people, namely people of low socioeconomic status. Eubanks raises concerns that human biases are built into the AFST such that the resulting scores potentially discriminate against certain groups. Specifically, Eubanks suggests that the AFST targets poor families through the inclusion of the use of public services in its predictive data. Part of the predictive power of the AFST comes from data about a family's use of public services, which might initially suggest that the AFST discriminates against poor people.[24]

Despite this intuition, it is not immediately clear what it would mean for a predictive risk algorithm to be discriminatory, and when this sort of discrimination is morally impermissble.

In order to consider the strength of this ethical claim against the AFST, I first need to establish an understanding of what discrimination is and when cases of discrimination are wrong. Once I establish the conditions under which an example of discrimination is wrongful, I will reframe Eubanks' ethical claims of discrimination against the AFST according to this more rigorous definition, and evaluate the strength of these specific claims by considering whether the AFST fulfills these conditions. Ultimately, I will argue that the AFST does not fulfill the conditions of wrongful discrimination and therefore does not wrongfully discriminate against poor families.

1. What is Discrimination?

The term "discrimination" is used in general discourse in two main ways: (1) as a non-moralized, descriptive term or (2) as a necessarily normative assertion of disadvantage.[25] In

[24] Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*.

[25] Altman, "Discrimination", *The Stanford Encyclopedia of Philosophy*; makes a similar distinction in section 1.2: The Moralized Concept.

the first case, there is no prima facie moral judgement attached to the term, such that an act can be discriminatory, but morally permissible. In the second case, a negative moral weight is necessarily attached to the term, such that all cases of discrimination are wrong according to this definition. This second use is common in public discourse, where an assertion that an act is discriminatory implies that the speaker believes the act is necessarily morally problematic.

To spell out this distinction further, consider the following example of a morally nebulous charge of discrimination: A men's clothing company is accused of employing only male models, explicitly using gender as a requirement for the role. One may ask two different questions about this accusation. First, whether or not the men's company actually adopted a policy that differentiates between workers based on sex. And second, whether or not this differentiation is morally wrongful. The non-moralized term allows us to distinguish between these two questions. For instance, one could ask if this case is discriminatory and also if it is wrongfully discriminatory. The moralized term, on the other hand, is not nearly as precise because it captures both the act of differentiation along with the moral claim of wrongfulness.

For the purpose of clarity and precision in this essay, I will adopt the first term, with the task of qualifying whether the discrimination in question is morally problematic. Thus, discrimination will be defined in this paper as the differential treatment of an individual based on their membership in a group, defined by a specific trait. My next task is to identify the conditions under which discrimination is morally wrong.

2. When is Discrimination Wrong?

The aim in this section is to provide a general understanding of wrongful discrimination by identifying the necessary conditions that make discrimination morally wrongful. I will build up a definition of wrongful discrimination, starting with the general non-moralized definition of discrimination as the differentiation of an individual based on their membership in a group as defined by a specific trait. Ultimately, I will argue that discrimination is wrongful if it satisfies the following three conditions.[26]

> (1) Salient Group Membership Condition: The differential treatment of an individual based on their membership in a group defined by a socially salient trait.
>
> (2) the Differentially Favorable Treatment Condition: The differential treatment disadvantages one group over another.
>
> (3) the Explanatory Moral Justification Condition: There is no reasonable moral justification for the discriminatory treatment.

2.1 The Salient Group Membership Condition

Consider the following two examples of discrimination that result in different moral conclusions. In the first example, a school teacher decides to split students into two lines for lunch based on the first letter of the student's last name with A-M in one group and N-Z in another group.[27] In this case, the teacher's differentiation between students does not seem morally problematic. Now contrast this example with another school teacher who splits his class into two lunch lines based on the student's race, placing black and white children in different lines. The main difference between these two examples is simply the type of trait used for differentiation, which suggests that the type of trait chosen has moral weight in the wrongfulness of a discriminatory act.

The difference between the traits in the examples is that one trait refers to a socially salient group whereas the other does not. By socially salient, I mean a group that carries important social and moral weight because the group has been historically, unjustly

[26] Conditions based on Hellman, Deborah. *When is discrimination wrong?*. Harvard University Press, 2008.

[27] Adapted from the first chapter of Hellman (2008).

disadvantaged in the past.[28] These groups generally map well onto the legal category of protected classes such as race, religion, sex, age, and disabilities among others. Hellman describes the traits that define such groups as traits that have a history of being singled out for mistreatment or social disadvantage.[29]

In the first example, the trait used to create groups for the activity is not socially salient – the first letter of a student's last name does not carry historical moral weight, nor is it a group that is systematically disadvantaged in society. On the other hand, a student's race is a trait that carries significant moral and historical societal weight, and therefore seems problematic as a use for differentiating groups. Thus, we have reason to believe that one of the conditions for wrongful discrimination is the salience of the group membership. We can add this first necessary condition of wrongful discrimination, the Salient Group Membership condition, to our working definition. So far, wrongful discrimination is the differential treatment of an individual based on their membership in a group defined by a socially salient trait.

2.2 The Differentially Favorable Treatment Condition

However, this condition seems insufficient as the sole indicator of wrongful discrimination when we look at additional examples of discrimination across socially salient groups that do not seem morally impermissible. In the first example, a doctor routinely prescribes a regular mammogram for a female patient because she is a woman. The doctor treats her patients differently because of their sex. Contrast this first example from a second in which an employer

[28] Philosophers like Deborah Hellman, Owen Fiss, and John Hart Ely agree that the type of trait chosen is important to an understanding of the wrongfulness of discrimination (Hellman 2008, 15). Specifically, they agree that traits that point to historically disadvantaged groups carry moral weight in making discrimination wrongful. However, much of the dialogue between these philosophers is around the question of why this condition is important to discrimination, with the assumption that it is important. For the purpose of this paper, we can leave the question of why this condition is important aside, as we only need to know what the conditions are for wrongful discrimination.

[29] Hellman, *When is Discrimination Wrong*, 21-22.

decides not to employ an applicant because the applicant is a woman. In both examples, the discriminatory agent treats an individual differently based on a socially salient group, namely sex. However, the first example does not seem like an example of wrongful discrimination, whereas the second does.

The contrasting moral intuitions in these two examples suggest that the consequences of the differentiation seem important to our analysis of the moral standing of the differentiation. Specifically, it seems that the differential treatment of a discriminatory act must be differentially favorable to be considered wrongful discrimination. In the case of the doctor, there is no clear differentially favorable treatment to either group even though the differentiation happens across a socially salient group. The doctor does not disadvantage one group over another, but rather tries to offer equally favorable treatment to both through the differentiation. In the example of discriminatory hiring policies, there is a clear disadvantage to a historically disadvantaged group, where an applicant is denied employment because of the fact that she is a woman.

Based on this explanation, we can add the Differentially Favorable Treatment condition to our working definition of wrongful discrimination. This condition measures whether the act of discrimination disadvantages one group over another. Thus, our current definition is that an act of discrimination is wrongful if it disadvantages an individual based on membership in a socially salient group.

One might object to this addition by pointing to cases of wrongful discrimination where there is no explicit disadvantage in treatment. A class of examples of this type are the Jim Crow laws passed to differentiate in such a way that is separate but equal. The defense of such discriminatory laws were that there was no explicit disadvantage between the groups. This ended up not being true, such that the resources for one group were significantly worse than that of the other. However, even if it were true that resources were the same, the fact of separation itself creates a disadvantage because of the stigma and harm caused by that separation in itself. Thus, the qualification of the existence of a disadvantage in the discrimination applies to any sort of harm or stigma as well as explicit disadvantages.

2.3 The No Independent Moral Justification Condition

The current account of wrongful discrimination seems insufficient when we consider discrimination where a clear disadvantage exists across a socially salient group, but is considered morally permissible. For example, suppose the state denies a twelve-year-old a driver's license.[30] In this example, the state disadvantages a person through the withholding of the privilege to drive because of the person's age, a protected class. According to our current definition, this should be a case of morally impermissible discrimination, yet moral intuition seems to say that this case is permissible. This case is permissible because the disability directly explains the differential treatment in this case. That is, the socially salient trait of the disability offers a direct, morally plausible justification for the disadvantage through the showing of a clear causal link between the trait and the differentiating act.

The purpose of this first example is to establish the basis for a plausible, explanatory moral justification. That is, it is possible to offer a reasonable justification for a clear disadvantage to a socially salient group that would render the discrimination morally permissible. The standard of reasonableness for a given justification requires a context-specific analysis of the explanatory strength of the justification. Thus, we can add the Explanatory Moral Justification condition to our working definition, but will also consider more examples to understand the limits of reasonableness for these justifications.

In order to understand the limits of a justifiable explanation, I will consider additional examples of discrimination where multiple justifications are offered. For example, consider a

[30] Hellman (2008) has a similar example.

construction site that decides not to hire an applicant based on the existence of a criminal history of a non-violent misdemeanor, a shoplifting charge, that took place over a decade ago. Contrast this example from one in which a daycare decides not to hire an applicant based on a criminal history that involves a similar category of gross misdemeanor as the first, except the specific charge is one of child abuse.[31] The first seems like wrongful discrimination whereas the second seems morally permissible.

Both of these examples are cases of discrimination where a person is disadvantaged due to membership in a socially salient group. However, the justifications given are not equally reasonable. Both of these discriminatory employers might offer the justification that the past crime makes the applicant unable to properly carry out the job well. However, it is unclear that the past history of the first applicant has any bearing on her ability or inability to work in construction. There is no clear explanatory link between the disadvantage and the reason given. Thus, the lack of a proper justification suggests that this first case is one of wrongful discrimination.

On the other hand, the justification given by the employer at the daycare seems reasonable. That is, the ability to work well in a daycare requires that the applicant is able to work well around and with children. The existence of a substantiated charge of child abuse suggests that the applicant is unable to properly carry out this job. The justification given is a reasonable explanation for the disadvantage such that this act of discrimination is not wrongful. The lack of a reasonable, explanatory moral justification for a disadvantage across a socially salient group indicates wrongful discrimination.

[31] The least serious cases of child neglect are considered gross misdemeanors in the same category as petty theft or driving under the influence of alcohol: https://criminal.findlaw.com/criminal-charges/child-abuse-penalties-and-sentencing.html

Thus, the final account of wrongful discrimination is a form of differentiation that disadvantages based on a socially salient trait without a justifiable explanation for the disadvantage. An act of discrimination is wrongful when it satisfies the following three conditions: (1) the Salient Group Membership Condition, (2) the Differentially Favorable Treatment Condition, and (3) the Explanatory Moral Justification Condition. Based on these conditions for wrongful discrimination, we will proceed to consider whether or not the AFST wrongfully discriminates. The next section will focus on objections to the claim that the AFST can be discriminatory, which will motivate a distinction between two types of discrimination, direct and indirect.

3. Types of Discrimination: Direct vs. Indirect

Despite establishing the conditions under which discrimination is wrongful, it is not immediately clear what it would mean for a predictive risk tool like the AFST to discriminate. One might object to the general claim of discrimination against the AFST, arguing that only people can discriminate, not algorithms or software systems. Each of the examples of discrimination offered earlier in the paper concern people discriminating against other people without mentioning machines. Moreover, one can claim that predictive risk assessment systems are objective in a way that humans generally are not.

There are two ways in which a predictive risk assessment system like the AFST can be discriminatory. First, the humans who created the AFST could use the tool as a means to discriminate against others, either explicitly or while hiding behind the veneer of technological objectivity. The AFST, as with all other risk assessment systems, was created by humans who potentially have biases or intentions to disadvantage a group. The creators and users of the AFST could use the tool to discriminate against certain people.

Second, the predictive risk assessment tool itself could make societal decisions that result in the differentially favorable treatment of a socially salient group, even if there is no intent by the creators to discriminate. Understanding the AFST as a decision-making agent that acts through its release of predictive scores for individuals and families, we can see that the decisions made by the AFST could systematically disadvantage one group. Even though these software systems can be more objective in their analysis, they often still capture the underlying biases that exist in the environment even if the creators did not intend for the tool to do so. This happens in many of the creative steps required in the construction of these systems, such as data collection, data training, and the choice of predictive variables among others.[32]

Based on these observations, one way to determine whether or not the AFST is actually discriminatory is to focus on the intentions and biases of the creators and users. If the creators and users have a clear intent to discriminate or a bias against a certain socially salient group, then there is reason to believe that the AFST is discriminatory. If there is no clear bias or intent to disadvantage, the AFST can still be found to be discriminatory through the existence of an unintended disadvantage to a socially salient group. These two types of discrimination fall along the philosophical distinction of direct and indirect discrimination.[33] In this section, I will distinguish between these two types of discrimination in order to ultimately argue that Eubanks' claim of discrimination is likely one of indirect discrimination.

In the direct case, the discriminating agent intentionally treats a group of people worse because of a bias against that group. This is the more paradigmatic example of discrimination, in which the differentially favorable treatment is motivated by animus. Consider the example of a restaurant owner who denies service to black guests because the owner does not like black

[32] I will leave the discussion of how biases enter into machines aside for now, and return to specific ways in which biases might enter into the AFST later in the paper.

[33] Altman, "Discrimination", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), section 2.

people. The owner intentionally and explicitly treats black guests differently and worse than white guests because of a specific trait, their race, resulting in disparate treatment of black guests.

I argue that the AFST is not directly discriminatory. Both the DHS and Vaithinathan's team are clear in their concern about discrimination and their desire to mitigate it, suggesting that they have no intent to discriminate nor any desire to promote differentially favorable treatment. Moreover, the creators of the system approached the development of this tool with a high level of transparency, indicating that they are not trying to conceal any discriminatory intentions. Based on these observations, it seems that the AFST is not discriminatory with regards to an existing intent to discriminate.

Absent any intent or bias against a group, an agent can still wrongfully discriminate against a group when the agent's actions result in a disproportionate negative effect on a group of people without a reasonable justification. Consider the paradigmatic example of the U.S. Supreme Court decision in *Griggs v. Duke Power* (1971), in which a utilities company in North Carolina decided to use written tests to determine promotions, resulting in an almost full exclusion of African American employees from promotion. The company was not accused of directly discriminating against African Americans, but given the disproportionate disadvantage to this socially salient group coupled with a lack of clear reason to use the written test, the U.S. Supreme Court considered this a case of wrongful indirect discrimination.[34] Without a reasonable justification for the existence of this disadvantage, the unfavorable differentiation is considered wrongful.

Unlike cases of direct discrimination, cases of indirect discrimination do not look for any intent or bias that motivated the act of discrimination. This type of discrimination depends on the

[34] Altman, "Discrimination", *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), section 2.2.

existence of differentially favorable treatment. Tools like the AFST can be indirectly discriminatory in a number of ways. For example, the risk scores given by a tool could vary on average across a socially salient group. Or, it could turn out that the risk scores given to one group are systematically less accurate than those given to another group. Both of these seem like ways in which a tool like the AFST can be indirectly discriminatory.

In the following parts of this chapter, I will consider whether Eubanks' claim of indirect discrimination fulfills the conditions of wrongful discrimination developed earlier. This will entail developing a more precise version of Eubanks' argument based on the direct and indirect distinction followed by an analysis of the wrongfulness of her charge of indirect discrimination. This analysis will proceed through a progression of the three conditions of wrongful discrimination considered earlier: (1) the Salient Group Membership condition, (2) the Differentially Favorable Treatment condition, and (3) the Explanatory Moral Justification condition.

4. Eubanks' Charge as Indirect Discrimination

Based on the earlier discussion, we can formulate Eubanks' charge as one of indirect discrimination in which she argues that the AFST systematically fails the poor. She claims that "the AFST interprets the use of public resources as a sign of weakness, deficiency, even villainy."[35] Her claim rests on a belief that a family's AFST score is higher with the inclusion of the use of public services as a feature than it would be without its inclusion as a feature. This initially seems problematic given that public services are predominantly used by those with a lower socioeconomic status. This implies that there is potentially differentially favorable treatment through the scores for poor families. Eubanks claims that the AFST scores given for a

[35] Eubanks, *Automating inequality*, 108-9.

poor family are higher than those for a middle-class or wealthy family because of the socially salient feature of socioeconomic status.

Eubanks establishes her claim by pointing to the high number of predictive variables in the AFST that are measures of poverty. Of the 131 predictive variables in the AFST published by the DHS in its efforts at transparency, almost half are either direct or indirect measures of poverty. For example, a quarter of the variables track direct measures of poverty like the "use of means-tested programs such as TANF [Temporary Assistance for Needy Families], Supplemental Security Income, SNAP [Supplemental Nutrition Assistance Program], and county medical assistance." Another quarter of variables track indirect measures of poverty like "interaction with juvenile probation."[36] Based on this evidence, Eubanks constructs a charge of wrongful discrimination against the AFST.

Prima facie, it seems that Eubanks' charge fulfills the conditions of indirect wrongful discrimination. The Salient Group Membership condition is fulfilled as the differential treatment occurs for poor families, a socially salient group with a history of disadvantage. Moreover, an initial pass suggests that the differential treatment disadvantages the poor. According to Eubanks, we have reason to believe that the risk scores for poor families are higher on average, which means that poor families are investigated more often. Eubanks draws on anecdotal evidence from families that have experienced the fear, loss of privacy, and intrusion sometimes caused by these investigative visits from government officials. Based on Eubanks' findings, it seems that increased investigations can constitute real harm. Therefore, it seems that the Differentially Favorable Treatment Condition can be satisfied.

Finally, we must consider the Explanatory Moral Justification Condition. One reasonable moral justification for the existence of a differentially favorable treatment of poor families through

[36] Ibid., 125

higher risk scores is simply that the base rates of child abuse for poor families are higher, suggesting that they are actually higher risk than for rich families. This justification is plausible and is empirically supported.[37]

However, Eubanks' claim is not only that poor families receive higher risk scores, but that the reason for the increase in risk scores is because a poor family uses public services. Even though it is likely the case that the base rates of child maltreatment are higher for poor families, this fact alone does not explain all of the differentially favorable treatment. Her point is that the inclusion of the use of public services as a predictive feature directly raises a family's score. It seems an unreasonable justification to claim that an increase in a family's percieved risk of child abuse is explained by a family's decision to use public services.

In the following section, I will outline the progression of the dialogue between the DHS and Eubanks, looking at the DHS's initial response to Eubanks followed by Eubanks' follow up response to the DHS. Then I will evaluate the strength of these responses, and ultimately argue that the DHS's response to Eubanks' charge of indirect discrimination is insufficient because it does not directly address Eubanks' claim.

5. The Public Debate: Eubanks vs. DHS

In response to Eubanks' claim, the DHS pushes back by arguing that the AFST does not actually disadvantage poor families, and is therefore not wrongfully discriminatory. The DHS point out that a family's use of public services in Allegheny County has a positive correlation with the likelihood of child abuse. That is, the use of public services actually decreases the scores for families that access the public services. In their executive statement, the DHS cites

[37] Researchers Paxson and Waldfogel find that the general socioeconomic trends of poverty in an area are positively correlated with a greater number of cases of child maltreatment. Francis. "Poverty and Mistreatment of Children Go Hand in Hand." Poverty and Mistreatment of Children Go Hand in Hand, https://www.nber.org/digest/jan00/w7343.html.

the statistic that “in reality, for 45% of families, receipt of the services is protective, that is, their receipt lowers the AFST score.”[38] This means that the use of public services actually has a negative correlation with the AFST score for a little under half of the families. A family’s AFST score is often lower if they use public services than if they do not.

Eubanks responds to the DHS by disputing the claim that the use of public services is protective. She focuses on the same statistic and suggests that a score that is protective for only 45% of families is not actually protective for the majority. On her website, Eubanks responds to the DHS’s executive statement by arguing that “the County’s January 31 statement suggests that for 55% of families – the majority – receipt of public services does in fact raise their AFST score, leaving them disproportionately vulnerable to child welfare investigation.” Eubanks attempts to use the complement of the same statistic given by the DHS to argue that the use of public services in the AFST score leads to higher scores for the majority of families.

Eubanks’ response is unsuccessful because her analysis of the data overlooks other important possibilities. We cannot assume that the use of public services raises the AFST score for 55% of families based on the fact that it is protective for 45% of families. From this statistic, we only know that the feature is protective for 45% of families, which means that the feature has either no effect or a negative effect on the remainder. Of the remaining 55% of families, it could be the case that the use of public services as a feature has no effect or little effect on the AFST scores. Even though Eubanks’ follow-up response is unsuccessful, her charge of wrongful discrimination against the AFST is not affected by the DHS’s response.

The response by the DHS is insufficient in addressing Eubanks’ claim of discrimination because the response subtly answers a related, but distinct question from the one that Eubanks’ claim poses. Eubanks argues that the inclusion of the use of public services as a feature in the

[38] Executive Statement from County of Allegheny, County Executive Rich Fitzgerald

AFST leads to higher scores than if the features were left out. However, she is not arguing that the physical use of public services itself has no benefit for families who need it. In fact, Eubanks would likely agree that a family's use of public services is protective for the family because the family receives important and helpful services. However, the use of this fact by the AFST might actually increase a family's score when compared with families who have accessed services but for whom this fact is not recorded.

The DHS however, responds to the former point, claiming that a family's use of public services leads to lower scores on the AFST. The DHS does not address the primary question of how the actual inclusion of the fact of the use of public services affects a family's score. This is an empirical question that would require additional research in order to reach a conclusion, but Eubanks' claim seems plausible.

For now, I will assume that the inclusion of the predictive variable does in fact increase scores for poor families. That is, I accept the claim of differential treatment across a socially salient group for the sake of argument. Even with this assumption, I argue that a higher AFST score does not necessarily mean a disadvantage to the family with the higher score, and there is therefore no clear differentially favorable treatment across a socially salient group. I will construct this argument by first analyzing the moral effects of a high AFST score. This moral analysis will consider the effects of investigations on the different moral stakeholders.

6. Moral Analysis of the Effects of Investigations

Contrary to Eubanks' assumption, a higher AFST risk score and a subsequently higher rate of investigations, on its own, may not necessarily be more harmful than a lower score and a lower rate of investigations. Eubanks' reasoning is as follows – higher AFST scores lead to a higher number of investigations. Investigations are generally harmful to families because of their intrusive nature. Thus, higher AFST scores are generally more harmful to families than lower AFST scores. Based on this reasoning, the moral weight of the risk scores depends on the harmfulness of investigations.

If it turns out that investigations are not more harmful, then higher AFST scores for one group would not constitute differentially favorable treatment. Thus, the moral analysis in this section will focus on the moral effects of government investigations.

The moral weight of these investigations depend on the outcome of the investigation and the moral stakeholders affected by the decisions. Eubanks correctly points out that these investigations can be harmful to families through an intrusive investigation, a loss of privacy, and potential repercussions for parents. However, investigations can also lead to important resources for parents, increased stability in the community, and most importantly, safety to a child who was in an abusive situation. Thus, the moral effects of an investigation seem to depend on the ultimate outcome of the investigation, whether or not the investigation leads to significant findings. In other words, the moral effects of an investigation depends on the accuracy of the investigation.

An analysis of the moral stakes of accuracy depends on an understanding of the benefits of a correct investigation and the harms of an incorrect investigation. The moral analysis in this section will focus on the benefits of true positives and true negatives as well as the harms of false positives and false negatives for different stakeholders with partially aligned interests. This will entail an initial discussion of the composition of the main moral stakeholders in an investigative case, followed by the effects of a true positive, true negative, false positive and false negative on these different stakeholders. After this moral analysis, I will consider the question of whether Eubanks' charge of a higher risk score for poor families disadvantages poor families.

Eubanks' claims focus on the disadvantage of scores to families as a singular unit. However, families are constructed of individuals that could have different moral claims and needs. For example, the parents might have a claim to custody of their child that conflicts with a child's claim to a non-abusive living environment. With regards to an investigation, the three main moral stakeholder groups within a family are the parental guardians, the children suspected to be at risk of abuse, and other children or extended family.[39]

Each of these groups have potentially competing claims on an investigation of child abuse directed at the family. The children have a moral claim to safe living conditions that are free of abusive or extreme neglect. The parents have claims to a degree of privacy and autonomy in their own child rearing practices as well as a claim to custody of their child. The remaining family members have claims to privacy in the generally private spheres of life within a home. Each of these claims work together to build some of the moral landscape that a decision to investigate a family for potential child maltreatment must consider.

6.1 True Positive and Negative

First, a true positive in the AFST means that a high risk score is correctly given, meaning that an investigation was substantiated through action. These actions fall into two general categories. In the first case, an investigation can lead to the placement of a child into foster care. This happens when the Child Protective Service (CPS) caseworker who conducts the investigation assesses that the living situation constitutes child abuse according to the state's legal definition of child abuse.[40] Another actionable outcome of an investigation is a provision of

[39] There are other moral stakeholders affected by the decision to investigate a family, including the surrounding community, the investigators themselves, and the child protection services. But, this analysis focuses only on the family unit, as this is the basis of the claim of harm proposed by Eubanks.

[40] According to PA family support alliance, child abuse is "when an individual acts or fails to prevent something that causes serious harm to a child under the age of 18. This harm can take many forms, such as serious physical injury, serious mental injury, or sexual abuse or exploitation." https://www.pa-fsa.org/Mandated-Reporters/Recognizing-Child-Abuse-Neglect/Abuse-Neglect-Definition

identified necessary resources for the family to reduce risk of harm to the child. The investigation itself also involves help in identifying goals and action steps for family members as well as a safety plan.[41]

Based on these actions, it seems that an investigation has important benefits when substantiated. At its best, an investigation can lead to stability in a community, important and necessary resources to parents, and most importantly, safety to a child who was otherwise in an abusive situation. When an investigation leads to action, the benefits of the investigation can be significant.

On the other end, a true negative in the AFST means that a low risk score is correctly assigned to a family, meaning that families who do not require investigation are left alone. Through powerful anecdotes, Eubanks reveals that investigations can be harmful to families, and avoiding these investigations when they are not necessary is an important benefit. Investigations can take several weeks and include unannounced home visits, can include interviews with different family members and members of the community, and can carry negative stigma within a community. As such, investigations can be intrusive and the avoidance of unnecessary investigations is an important benefit.

Both the true negatives and true positives provide important benefits to the individuals involved, such that having fewer of these benefits can constitute a harm. If the AFST is less predictive for one group than another, then that group misses out on important benefits that come with a properly running child welfare system.[42] The benefits of a true positives and negatives are tied to the harms associated with an investigation for individual families.

[41] From https://www.alleghenycounty.us/Human-Services/Programs-Services/Children-Families/Protective-Services.aspx

[42] Some might dispute this indication of harm by arguing against the child welfare system and the government's role in ensuring the safety of children in general. This is another lively debate, and one that I will set aside.

Therefore, an investigation that turns out unsubstantiated harms a family as much as a true negative benefits them, and a lack of an investigation when conditions of abuse exist harms the family as much as a true positive benefits a family.

6.2 False Positive and Negative

A false positive indicates an incorrectly high AFST score, meaning that it is more likely that an investigation takes place in a family where no conditions of child abuse exist. A false investigation is generally harmful to all of the relevant stakeholders within an individual family.[43] For parents, an investigation can carry with it an invasion of privacy, an increased negative stigma from the community, the potential loss of custody over a child, fear of future retribution, and feelings of being targeted tied to a loss of security. For the children, the potential harms of an investigation include the fear of future loss of parents or the actual loss of parents and the subsequent harms of being taken away from parents.[44]

One might object that the collective pursuit of investigations can justify these mistakes in a way that the group of investigations as a whole are beneficial because of the importance of keeping children in the community safe. However, the moral analysis here is only focused on the consequences of a single investigation on the direct moral stakeholders affected, rather than the overall moral permissibility of having investigations in general. The aim here is not to argue that investigations are harmful in general, but rather that a wrong investigation is harmful to the family that received the false positive score.

[43] The investigation might be beneficial to other actors, like the state or the community through the knowledge that there is no child abuse. But, for the purpose of this section, we are only looking at the benefits and harms to the specific family affected.

[44] There are vibrant debates in related literature about the nature and extent of the harm of taking a child away from a parent, and whether this is justified at all, let alone which conditions it is justified in. These discussions focus on the ethical challenges of investigators, but for this paper, we can lay this debate to the side.

As with false positives, false negatives also lead to harmful consequences for the stakeholders involved. A false negative indicates a mistakenly low score for a family, meaning that families that have unsafe, potentially abusive living conditions for the children are passed over for investigation. The clearest harm in this case is to the child who is left to live in an abusive living situation without mediation. This is a clear harm that takes precedence over some of the harms of having the investigation itself. Moreover, there is an additional harm to parents who may not receive the support they need from CYF if the family is mistakenly passed over for investigation.

7. Evaluation of Disadvantage

Based on the moral analysis above, I will argue that there is not a clear disadvantage to poor families based on the differential treatment of scores. That is, higher average AFST scores for poor families do not constitute differentially favorable treatment. A higher average AFST score means that a family is more likely to be investigated. A higher likelihood of investigation, however, do not constitute a clear disadvantage. The preceding moral analysis uncovered the nuanced moral weight of investigations, where true positives and true negatives generally benefit a family and false positives and false negatives generally harm a family. If it is the case that the greater number of investigations of poor families are substantiated at equal or greater rates, then these additional investigations might actually benefit rather than harm the families.

Thus, it seems that Eubanks' initial charge of indirect discrimination based on a disparate impact of generally higher scores for poor families does not necessarily disadvantage the poor, and is therefore not necessarily discriminatory. Whether or not the investigation disadvantages the family depends on the accuracy of the investigation. As a result, we can modify the original charge of discrimination to consider a differential treatment in the accuracy of the score across socioeconomic lines rather than in the scores themselves. Then, we can evaluate whether this new charge constitutes differentially favorable treatment of poor families.

I will first consider whether there could be differential treatment across accuracy in the AFST. For this, we will look at another step in the development of a predictive model that is susceptible to capturing human bias, namely the training and validation data. The training data is the initial data from which correlations are drawn to the desired outcome variable. The validation set is the data used to test the predictive correlations drawn from the training data to make sure that the model is predictive.

If the data is collected from only a subset of the population, the training data represents a skewed portion of the whole population for which it is making predictions. For example, consider COMPAS, a predictive risk model used by judges that provides risk scores on the likelihood that a criminal might recidivate. If the training data for this model is disproportionately gathered from a subpopulation with a socially salient trait, then the classification might only work differently for that subgroup than the other. In this example, the distribution of errors was different for African Americans than for other groups, all else equal, in part because the training data captured underlying biases.[45]

In the case of the AFST, Eubanks' presents a real concern that the data is misrepresentative, gathering additional data on families from a lower socioeconomic status. This concern stems from the fact that some of the data points gathered include direct and indirect measures of poverty, as discussed earlier.[46] It does seem that there is simply more data on poor and working-class families than middle-class families. Moreover, even if the training data properly represents the population, there might still be additional biases in the underlying environment, such that correct predictions have a disparate impact across a protected class.

[45] This example was adapted from the existing COMPAS predictive risk model.

[46] Eubanks, *Automating inequality*, 125.

As in the previous section, an empirical question remains to show that there is a true difference in the accuracy of the tool across a socially salient trait, like socioeconomic status or race. However, Eubanks' account gives us reason to believe that this might be the case. Given that there is a difference in accuracy between these two groups, does it constitute a disadvantage for poor families?

Based on Eubanks' claims about disproportionate data, it actually seems like the tool could be more accurate for poor families than for wealthier families. Given that there is more predictive data on poor families, it gives us reason to believe that the accuracy for poor families might be greater.

Based on our earlier moral analysis, we determined that the accuracy of the machine is morally salient, with a greater accuracy constituting more benefits, and a lower accuracy indicated more harm. Thus, there does not seem to be a disadvantage to poor families, but rather a potential advantage through a higher proportion of true positives and true negatives for this socially salient group.

Conclusion

I have argued that the AFST is not wrongfully discriminatory. The AFST is likely not directly discriminatory given the lack of evidence of any malintention from the creators of the machine. Moreover, the two charges of indirect discrimination seem not to fulfill the second condition of wrongful discrimination, namely, the Differentially Favorable Treatment Condition. In conclusion, it seems that Eubanks' claims of wrongful discrimination against the AFST are unsubstantial, and that the tool is not wrongfully discriminating against poor families.

Chapter 3: HCI Design Considerations for the AFST

The exact role of the AFST in the decision-making process is unclear. The creators of the tool claim that the tool is intended to be used as an information aide, simply augmenting the decision-making process while retaining full human agency. However, the technical design of the tool as a parallel decision-maker to call screeners implies that the tool should be used in a way that gives it agency in the decision-making process. This contrast between the stated use of the tool and the implicit use of the tool based on its fundamental design leads to a lack of clarity in the way that call screeners should use the AFST. As a result, it is unclear how a call screener should incorporate the decision of the AFST into their final decision.

This lack of clarity contributes to a lack of understanding, inconsistent usage, and a lack of calibrated trust in the tool. This lack of clarity fuels the public debate over the use of the AFST where critics argue that call screeners are relying too heavily on the AFST score while proponents argue that call screeners are not influenced greatly by the scores.

In this chapter, I will first analyze the public debate around the usage of the AFST by mediating Virginia Eubanks' charge of the misuse of the tool with responses from the DHS. This debate, along with further statistics from the process evaluation, indicate a wide disparity in call screeners' use of the AFST. I argue that this variation in usage is due not only to a need for better training of call screeners, but primarily from a confusion that lies with the designers themselves. That is, it is not only the case that the call screeners are confused about how to use the tool, but also that the designers themselves lack this clarity.

I then consider three distinct categories of the use of predictive risk models using a framework centered around the distribution of agency: (1) the machine as the sole decision-maker, (2) the machine and human as joint decision makers, and (3) the human as the sole decision-maker with the machine as a source of additional information. For each of these three categories, I provide a technical literature review of the ways that a predictive risk model can be used in a complex decision-making process along with an example of one in practice. Ultimately, I argue that the AFST is designed to be used according to either of the first two categories rather than the third.

Despite the current technical design of the AFST, I argue that it should be designed as an augmenting aide, leaving full agency with the call screener. Moreover, it is clear that the AFST is meant to be used as a decision support tool that assists call screeners in the decision making process rather than a mere failsafe. I then argue that the AFST is poorly designed to function as a decision support tool, ultimately offering suggestions for how the AFST should be redesigned to perform its intended function.

1. The Public Debate: How is the Tool Used?

One of the main points of contention in the public debate over the AFST is about the role of the tool in the decision-making process. Both sides agree that the tool is not meant to take over human decision-making, but rather augment it. However, they disagree in their perceptions of how the tool is currently being used in practice, with one side arguing that the AFST reduces human agency whereas the other arguing that the AFST maintains full human agency.

1.1 Eubanks’ Criticism

Critics like Eubanks argue that the tool is affecting call screeners to a great enough degree that it is starting to replace human decision making. Rather than acting as a supplement to human decision making, the AFST scores are training humans to comply with the scores and are therefore taking agency away from the call screeners. She claims that the call screeners’ deference leads them to be trained by the assessments of the system to the point that they are starting to predict what the AFST score will be rather than predicting child risk itself.

As evidence, Eubanks points to the call screeners' expressed desires to change an aspect of the current call screening workflow. In the current workflow, the AFST score is shown at the end of the call screening process, after the call screener has done her preliminary due diligence. This design decision is meant to limit the AFST's influence on the call screeners' ability to come up with independent judgements by giving the call screener the space to come up with a score without any interference from the AFST. The call screener is not allowed to change her initial assessment after seeing the score given by the AFST. Again, this decision is meant to protect a call screener's ability to make decisions independently of the AFST score without pressure to conform to the score.

The call screeners, however, have expressed a desire to go back and change their risk assessments after they see the AFST score. The DHS has resisted this change. Even though the DHS has not accommodated this request, Eubanks argues that the desire itself suggests that call screeners believe that "the model is less fallible than human screeners."[47] As such, it seems that call screeners believe that they should learn from the risk score in the cases where the risk score conflicts with a screener's judgment. She believes that this belief in the model is troubling due to the limited scope of the AFST, and that this belief leads to an inability for call screener's to make decisions independent from the AFST score.

1.2 DHS Response

In an executive statement published in response to the publication of Eubanks' book, the DHS mounts a defence of the AFST. The DHS argues that the AFST is not being used in such a way that it is influencing a call screener's ability to have confidence in their own decisions. That

[47] Eubanks, *Automating inequality,* 230.

is, call screeners often have confidence in their own decision making capabilities over the AFST assessment even when the scores conflict.

As evidence, the DHS offers empirical AFST usage statistics highlighting the seemingly low concurrency rates between call screeners and the AFST score. The statistics measures the percent of times that a call screener actively disagrees with the recommendation given by the AFST score. The argument is as follows: if the call screeners often do not act in concurrence with the AFST score, then it would seem that call screeners feel they have autonomy over the final decision. This evidence seems to address Eubanks' claim that the call screeners defer to the AFST score when there is a result that conflicts with their own result. A low concurrency rate would suggest that call screeners go with their own evaluations when a conflicting score appears.

The Allegheny County DHS's summary of the independently conducted Stanford *Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office* includes a table that measures the percent of children screened-in for investigation according to the AFST risk scores. These statistics measure screen-in rates between August 2016 through November 2018.[48]

Table 1: Percent of Children Screened-in for Investigation According to the AFST Risk Scores[49]

AFST Risk Score (1-20)	% Screened-In for Investigation
Mandatory (18-20)	61%
High (15-18)	47%

[48] Allegheny County Department of Human Services. Frequently Asked Questions, https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/FAQs-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-8.pdf, 12

[49] Data retrieved from Goldhaber-Fiebert, and Prince. "Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office."

Medium (10-15)	42%
Low (1-9)	31%
Total	41.4%

The AFST risk score is a single score between 1 and 20, with 20 being the highest risk. In the table above, a mandatory score is one that is 18 or higher, indicating very high likelihood of risk. The remaining three categories map onto distributions of scores between 1 and 20.[50] This table indicates that, for a significant portion of even the highest scores, call screeners are taking the opposite action from the one recommended by the AFST. Only half of all high risk scores result in a screen-in. Moreover, a significant portion of low scores also result in the opposite recommended action, with almost a third of low scores resulting in a screen-in for investigation.

These statistics tell us that call screeners are not always following the discretion of the AFST when the scores conflict with intuitions. Proponents, like the DHS, see the lack of concurrence as evidence that the call screeners have the autonomy to make their own decisions without deferring to the scores of the model.

However, the DHS' criticism does not sufficiently respond to Eubanks' criticism of the current use of the AFST. These statistics may show that that call screeners as a group are not always following the discretion of the AFST when scores conflict, but it might still be the case that some call screeners' are placing too much trust in the AFST. In other words, the DHS's

[50] The exact score threshold for each category is not included in published reports, but graphical representations of the categories give a general sense of the following distribution: High score (15 or greater), Medium (10 to 15), and Low (9 or less).

response does not rule out Eubanks' criticism that some call screeners place too much trust in the AFST score.

These statistics coupled with Eubanks' findings that some users trust the AFST score to a high degree suggest that there is likely a disparity in the way that call screeners use the tool. That is, some call screeners trust the tool to a fault whereas others completely reject the tool. Together, both sides of the debate help to build a more nuanced understanding of the actual use of the AFST, where some call screeners trust the score too much and others trust the score too little. This picture of a disparity in the amount of trust that call screeners have for the system is further corroborated by additional statistics about the usage of the tool.

1.2 Disparity in Usage

The independent process evaluation conducted by Hornby Zeller corroborates this observation of the disparity in the levels of trust of the AFST. In one of their post implementation surveys administered to call screeners, they found that "over 40% of the call screeners use the tool to inform their recommendation on a consistent basis" whereas "a little less than a third (31%) reported they rarely use it, if at all."[51] These numbers indicate a real disparity in the perceptions that call screeners have of the usefulness of the tool. A significant proportion of call screeners use the score always, or almost always, whereas a large number either never use it or rarely use it.

These survey results provide clarity to the public debate considered earlier. The critics argued that the call screeners tend to defer to the score whenever it conflicts with the call screener's assessment. The proponents argued that the call screener tends to disregard the score when it conflicts with the call screener's own assessment. The survey results from the

[51] Hornby Zeller Associates. "Allegheny County Predictive Risk Modeling Tool Implementation: Process Evaluation, 17.

process evaluation suggest that both sides are correct. That is, there is a disparity in the usage of the tool where some users rely heavily on the score whereas other uses simply ignore the score.

This disparity in use indicates a lack of consistency across users which could impact the accuracy of screening decisions negatively. One of the primary goals of the AFST is to increase accuracy and consistency with call screeners.[52] However, the disparity in usage suggests that the AFST is likely not increasing the consistency of outcomes related to accuracy. This suggestion is further supported by findings in the independent Impact Evaluation conducted by researchers at Stanford that found that the AFST "did not significantly alter the consistency of outcomes relating to accuracy or workload across call screeners."[53]

Based on these findings, it seems clear that there is a disparity in the ways that call screeners use the AFST to inform their decision making and that this disparity potentially mitigates the effectiveness of the AFST to fulfill its goals. These findings, however, lead us to another natural follow up question. Namely, why does this discrepancy exist?

1.3 The Source of the Issue

One reasonable response is that there must simply be better training for call screeners. According to this view, the tool is designed clearly such that the creators understand how to use the AFST well, but it is simply a lack of proper communication of this understanding that results in varied usage practices. Thus, in order to increase consistency, the DHS must simply improve their current training program to better inform call screeners of the proper way to use the tool.

It seems unlikely that a lack of proper communication in the training is the reason for varied usage. According to a set of interviews of call screening staff conducted in a process

[52] Goldhaber-Fiebert and Prince. "Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office," 1.

[53] Ibid., 2

evaluation of the AFST, 100% of 18 call screeners interviewed either agreed or strongly agreed that they are “adequately prepared to use the tool.”[54] Call screeners perceive that they are prepared to use the tool following the extensive training offered by the DHS, yet use the tool in such varied ways. This perception coupled with an understanding of the quality of the training itself suggests that the communication of the existing ideas is not to blame, but rather an omission of more specific guidance on how to use the tool.[55]

The source of the problem seems to lie instead in a lack of clarity in the way that the score should be used to augment decision-making. It is not a lack of proper training of call screeners, but rather a more fundamental problem with the design of the AFST that leads to the disparity in usage. In other words, it is not surprising that the call screeners are confused about how to use the tool – the designers of the tool are themselves are confused about its use.

The creators of the system explain only that the score should augment call screening decisions with an emphasis on call screener autonomy. They do not explain exactly how to use scores to modify decisions when the scores diverge with independent judgments by the call screeners. That is, how should a call screener choose an action if his independent evaluation conflicts greatly with the score given by the AFST?

2. The Intended Role of the AFST

The creators of the AFST distinguish between two potential uses of predictive screening tools in health and human services, one “is to replace clinical decisions (e.g., through automatically screening in children based on their score) and the other is to augment and

[54] Hornby Zeller Associates. “Allegheny County Predictive Risk Modeling Tool Implementation: Process Evaluation,” 13.

[55] The training includes a standardized three-hour session for staff members at the DHS including all full time and occasional call screening staff. The training focuses on teaching call screeners general knowledge about predictive risk models and risk modeling, along with how the AFST risk model was built, and how predictive the AFST is. Vaithianathan, et al., 2017, 32.

standardize clinical decisions (e.g., through a “risk score” or a summary statistic weighting information from the administrative data).” The creators claim that the “Allegheny County was interested in developing the latter type of tool – one in which an empirically derived score could be used in conjunction with clinical judgement … to generate a hotline screening decision (screen in or out).”[56]

The distinction made by the creators is one that focuses on the distribution of decision making power. In the first case, the model is given full agency over the decision, whereas humans have full agency in the second. They do not include a paradigm where machines share any agency with the human, but rather focus on the two extremes where there is only a single decision maker. As such, the model must not influence screening decisions too much as to take away the ability for independent judgement, but it must also influence screening decisions to a sizeable enough degree so as to render the model useful. This is where the confusion exists.

This underlying confusion also surfaces in the debate between the DHS and Eubanks. The DHS’ reply to Eubanks implies that the DHS does not want the AFST to affect a call screener’s ability to make their own decision. The DHS provides evidence that they believe suggests that the call screeners have autonomy. But surely, the DHS expects that the AFST score does influence a call screeners decision to some degree, otherwise the tool would be useless as an information aide. How exactly do they intend for the tool to influence call screening decisions, especially when the score conflicts with the call screeners’ independent judgements?

When the risk score converges with the call screening score, then a call screener’s decision will likely be the same. The likely next action is clear. Simply follow the previous judgement. However, when the score diverges, it is unclear how a call screener should proceed.

[56] Vaithianathan et al. "Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation,” 15.

In order for the score to standardize decisions across call screeners, the screening decisions must change along the lines of scores given by the AFST. However, it seems that screeners are influenced by the AFST to varying degrees, such that some might revise their ultimate decision whereas others might stick to their own interpretations.

Thus, the question we must answer is how the AFST should be used in the decision making process. This question has two distinct subquestions, both of which I will address separately. First, given the technical implementation of the AFST, how should the tool be used? This question is concerned primarily with the way that the technical design of the AFST implies a certain usage.

Second, regardless of the technical implementation, how should the tool be used given the context of call screening decisions at the DHS? This question is primarily focused on an analysis of the characteristics of the context in which a call screening decision must be made at the DHS.

Similar to the creator's distinction between two different ways that PRMs could be used, I will look at three different ways that a PRM can be used based on the different distributions of decision making abilities. Then, I will offer examples of good use cases of each of these three different types of decision-making paradigms to provide the conditions under which these use cases are good. Finally, I will look specifically at the AFST within the current context to determine how it should be used.

2.1 Categories of Predictive Risk Models

Predictive risk models can be used in importantly different ways. The importance of this question about the proper distribution of agency between the machine agent and the human agent often gets overlooked in algorithmic development.[57] As a result, the migration of these

[57] Madras et al. "Predict responsibly." *Advances in Neural Information Processing Systems*. 2018.

algorithmic tools from the lab into practice often fails.[58] Much of the current literature focuses on the creation of accurate and fair systems themselves, but neglects to touch on the important design considerations of the practical interactions between the system and the existing human-led decision making workflows.[59]

Moreover, the literature that does consider the design of these interactions focuses mainly on the use of a system as a co-decision maker with humans without considering the important characteristics of a decision making context that suggest how the machine should be used.[60] That is, even this literature focuses on optimizing for human use of the system itself without first considering the preceding question of when certain decision-making paradigms are actually better to use than others.

In response to the current confusion, I will categorize the uses of predictive risk models into three types based on the distribution of decision-making power between the human and the model. The three categories I will consider are the following: (1) the machine as the sole decision maker, fully replacing human decision making, (2) the machine as a co-decision maker, replacing human judgement where confident, and deferring where there is a lack of confidence, and (3) the machine as a supplement to human decision making with no real decision making power.

Each of these paradigms results in a different understanding of the system, its responsibilities, and its proper usage. If users are unclear about the role of the machine in the current context, then there will be a subsequent lack of understanding of the system. This distinction is central to a proper usage and adoption of predictive risk models in practice. Moreover, each category compels a different technical design of the system such that

[58] Yang et al. "Unremarkable AI." Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 2019

[59] Papers on fairness in ML.

[60] Madras et al. "Predict responsibly."

developers of algorithms should choose which category best suits a given context prior to the technical implementation of the tool.

For each of these categories, I will provide a description of the category coupled with an example of its usage. Then, I will consider the characteristics of a decision-making context that support a specific distribution of agency. These characteristics include different levels of the simplicity of the task, the gravity of the consequences, and the desired outcomes of the adoption of the model.

Table 2:

	Full Machine Agency	Joint Human-Machine Agency	Full Human Agency (Fail Safe)	Full Human Agency (Information Aide)
Level of Generalizability across contexts	High	Medium/Low	Low	Low
Gravity of Consequences	Low	Medium/High	High	High
Desired Outcome	Save resources, increase accuracy and consistency	Save resources, increase accuracy and consistency	Increase accuracy and consistency	Increase accuracy and consistency
Design of System (Focus)	What to decide	What to decide	What to Decide	How to Decide

2.1.1 Machine as Sole Decision Maker

In this first category of usage, the models are used as the sole decision makers, entirely replacing the judgment of a human. The model makes a decision and takes the subsequent action. There may be human oversight of the decisions made by the machine, but the ultimate decisions are made by the machine rather than the human. This use of models is uncommon in health and human services, but is often used in other domains such as spam filters, loan approvals, and credit scoring.[61] In this section, I will consider an example of a paradigmatic case of a model used in this manner and then explain the characteristics of a decision making workflow that make for a good use case.

Consider the simple example of spam filtering for emails. Almost 45% of all emails are considered spam, amounting to 14.5 billion messages globally each day.[62] As a result, it is practically impossible and undesirable for people to sift through and filter spam emails manually. As such, most email services, like Gmail, use a statistical model to predict the likelihood that an email is spam and filter emails above a certain level of confidence without the need for human oversight. That is, the algorithm itself makes the final decision to filter an email or not, without reliance on a human agent.

This decision making workflow is ideal for using a machine to replace a human decision maker. The characteristics that make this situation ideal for the usage of a machine learning algorithm as the decision making agent can be seen by looking at the considerations necessary to make a correct decision, the repercussions of a mistaken decision, and the desired outcome of the adoption of a machine learning algorithm.

The main consideration necessary to make a decision to filter an email out or not is how likely the email is spam. If an email is very likely spam, then the decision making agent, whether human or machine, should filter the email out. There are no additional considerations of fairness, equity, or competing interests that factor into the final decision given the simplicity of the task. The task of filtering an email is generalizable across different emails and environments because of the closed context involved with the simplicity of the task.

[61] Burrell. "How the machine ‘thinks’." *Big Data & Society 3.1* (2016): 2053951715622512
[62] “Spam Statistics and Facts.” Spam, https://www.spamlaws.com/spam-stats.html.

Moreover, the consequences of a mistaken decision are low. In the case of a mistakenly unfiltered spam email, the user can simply manually flag the email as spam or delete the message. These mistakes alone do not carry high stakes. The case of a mistakenly filtered email might be higher if the email filtered was an important one to the user. However, the decision making agent can simply err on the side of safety in order to avoid errors of this nature in exchange for a higher rate of false negative errors. The low stakes of a mistake by a machine make this problem an ideal one for a machine agent.

Finally, the primary purpose of adopting a machine in this case is to save resources in time and human effort rather than just to increase accuracy of decisions made. Given the size of the problem and the difficulty of filtering without an automated decision maker, it seems likely that the desired outcome of the adoption of the machine is to enable the filtering of spam emails to save human capital of time and effort. While accuracy and consistency are important considerations, these are not the primary goals for adopting the machine model. The desire to save human resources signals that this problem is one in which a machine replacement is ideal.

The use of a machine as a replacement to human decision making is ideal in situations where the task at hand is simple and easily generalizable, the consequences of a mistake are not significant, and the primary desired outcome is to save resources rather than increase consistency or accuracy. Given that decisions in healthcare are highly context-specific, have significant consequences, and are primarily concerned with accuracy, it is understandable why machines are not generally used in this manner in healthcare.

2.1.2 Machine as Co-Decision Maker

The second category of the use of a predictive machine learning model is the use of the machine as a co-decision maker. In this paradigm, both the human and the machine have the ability to make decisions, but one defers to the other when one has a higher confidence of accuracy. The two decision making agents work as one joint unit, switching off decisions made based on the level of confidence that each has in their decision.

The most simple joint decision making scheme is one called rejection learning. With rejection learning, the model is allowed to reject or conceal a score if it is below a predefined confidence threshold.[63] For example, consider a model, like the COMPAS algorithm, that predicts the likelihood that a criminal will recidivate, helping judges make decisions on questions of bail.[64] In a joint decision making paradigm, the algorithm will “pass” without offering a judgment when its confidence is lower than a specific threshold, say 90%. This allows the judge to make an independent judgment when the score is low, but the decision is made by the machine when the confidence is high. This type of learning, however, is not ideal.

Rejection learning is a nonadaptive procedure, meaning that each of the systems is working independently of the other, when in reality the two should depend on the strengths and weaknesses of the other. When a model is working as a joint decision maker with a human, then the ultimate decision to reject a model’s recommendation should depend on not only the model’s confidence, but also the human’s confidence and weaknesses. For example, consider the same recidivism risk algorithm from the earlier example. If it turns out that a judge is very uncertain about a certain group, or is often inaccurate or biased towards that group, then it may be better to follow the decision of the model despite an uncertainty in the model that falls below some established threshold. That is, the ultimate decision made should depend on both the strengths and weaknesses of the machine and the human.

This new sort of procedure is called adaptive rejection learning.[65] With this method, both agents must have a clear sense of calibrated trust of the other– the human must understand the

[63] Madras et al. "Predict responsibly."
[64] Angwin, et al. “Machine Bias.” ProPublica, 9 Mar. 2019,
[65] Madras et al. "Predict responsibly."

situations in which a machine is less accurate, and the machine must know when the human is less accurate than itself. Madras et al. offer a formalization of this adaptive learning procedure based on the intuition that a machine should only pass when it is less confident than a human in relation to the external human decision maker’s level of confidence as well.

Consider an example of joint decision making with a model that is trained to detect melanoma.[66] Using the adaptive rejection learning procedure, the model only passes if it is not confident about its prediction and the doctor has a more informed, nuanced opinion than the model. On the other hand, the model will not pass if its confidence level is high and the doctor is less accurate at detecting a specific type of melanoma. The two work together to increase the capacity to make better joint decisions than either one would make on its own.

Unlike the decision to filter an email as spam, the decision to test for melanoma is complex with multiple considerations. In addition to considerations of likelihood of the existence of melanoma, one must consider questions of fairness, discrimination, and cost of testing. Are the decisions systematically less favorable to individuals of a certain socially salient group? Should the doctor choose to test a patient with a low likelihood of melanoma despite a potentially crippling cost for this patient? The decision making context is fairly straightforward in the sense that the task is directly measurable and easily verifiable, yet the decision must be weighed across multiple considerations.

In addition to the contextual difficulty of the task, the consequences of a mistaken decision to test for melanoma are much more significant than that of a mistaken filtering of spam. When a doctor mistakenly tests for melanoma, the consequences are not as high as a case when a doctor mistakenly passes by an opportunity to test for the disease. In the first case, consequences include potential economic costs of the test and some emotional costs from fear

[66] Ibid.

or anxiety. On the other hand, the consequences for missing a test when melanoma does exist can be grave, with serious physical, economic, and emotional costs as the cancer is allowed to spread undetected for longer. The skewed distribution of potentially high stakes of a mistake is important when considering whether or not to incorporate a machine agent.

Finally the intended outcome of the adoption of a machine as a joint decision maker in this case is primarily to increase the accuracy and consistency of decisions made. Although the creators might want to save resources like time and money, human capital is still necessary in this scheme. Beyond the resource considerations, it seems that the main consideration in the adoption of a machine in this case is to increase the accuracy and consistency. The primary concern is to help physicians make better decisions about the diagnosis without necessarily trying to reduce resources to make the decisions.

The use of a machine as a joint decision maker with a human agent is ideal in situations where the task at hand is moderately context-dependent, the consequences of a mistake are moderately grave, and the primary desired outcome is to increase consistency and accuracy in decision making with some savings of resources. Given that decisions in healthcare are often context-specific, have significant consequences, and are primarily concerned with accuracy, it seems that this sort of joint decision-making scheme could be a good fit for certain contexts.

2.1.3 Machine as Supplement to Decision Maker

The third way to assist decision making using a machine learning algorithm is to use the algorithm as a supplement to the decisions made by a human. In other words, the model has no real agency, but rather works to augment the human’s decision by acting as an additional source of information. The human retains the ability to make all final decisions, but the machine is simply left to help the human in making those decisions. The machine output augments the human decision making without taking any agency away from the human.

With this category of usage, it is important to specify the exact manner in which the machine should augment the human and how the human is supposed to use the machine to make decisions. Without clarity in this regard, even a highly accurate machine output becomes less useful as a lack of specification leads to inappropriate uses of the underlying technology. Recent studies of the practical implementation of predictive risk scores in this manner, without proper specification of usage, found that usage of the system fails to achieve the desired goal partly because of the users' varied discretion of its use.[67] Specifically, a study on the impacts of adopting algorithmic predictions of future offenses to help judicial decisions found that, while the algorithm itself is more predictive than the judge, the combination of the two failed to reduce recidivism rates.[68]

The researchers, an economist and a law professor, posit that this failure is due to a lack of clear thinking about how the human and machine are supposed to interact. They found that the judges are influenced by the risk scores, leading to greater sentences for those with higher scores and shorter ones for those with lower scores. However, they also found an inconsistency in the judges' use of the machines through a selective use of the score to make a final decision. [69] That is, the judges do not use the machine in a consistent manner, similar to the way that call screeners do not use the AFST in a consistent manner.

Method #1: Fail Safe

This research suggests that any system in which a human has the final decision must specify how a machine should inform or augment decisions in order to work effectively. I will distinguish between two different ways that a machine can augment human decision making: as

[67] Stevenson et al. "Algorithmic Risk Assessment in the Hands of Humans." Available at SSRN (2019).
[68] Ibid.
[69] Ibid.

a fail-safe mechanism to increase consistency and accuracy, or as an informational aide that highlights certain aspects of the problem.

The use of a predictive risk model as a mere fail-safe mechanism within the decision making workflow has precedence in healthcare. For example, researchers have begun to use a predictive risk model to support the decision to implant a VAD (ventricular assist device), an artificial heart. For many heart failure patients who are ineligible for a heart transplant, a VAD is the only chance to extend their lives. However, the post-implant fatality rate is high, so decisions must be made with great care.[70]

In this example, a predictive risk model is used to predict the likely outcome of an implant, and is developed to be unremarkable – only pausing the decision-making workflow when the scores given differ from the scores of the physicians.[71] When the scores align with the physicians' decision, then the score does not affect the decision-making workflow. In essence, the tool is used as a fail-safe, stopping physicians when the score differs so that the physicians can go back and make sure that nothing was missed in their initial analysis. This use of a predictive risk model can help increase consistency as well as accuracy in high stakes decisions while leaving full agency to the human.

Method #2: Information Aide

In addition to a fail-safe mechanism, a predictive risk model can be used as an information aide. Unlike a mere fail-safe mechanism, an information augmenting aide should be designed to focus its advice on *how* to decide rather than *what* to decide.[72] Rather than offering a single risk score parallel to the risk score produced by the human agent, the tool should

[70] Benza, et al. "An evaluation of long-term survival from time of diagnosis in pulmonary arterial hypertension from the REVEAL Registry." Chest 142.2 (2012): 448-456.

[71] Yang et al. "Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes."

[72] Hilgard et al. "Learning Representations by Humans, for Humans." arXiv preprint arXiv:1905.12686 (2019).

highlight specific aspects of the problem, provide additional information through interactive graphics, present tradeoffs in risks and returns, or outline possible courses of action.[73]

There is further evidence that suggests that informative advice that acknowledges the central role of decision makers can enhance performance while retaining agency.[74] As such, machine learning tools like deep neural nets can learn representations without explicitly presenting a recommended action, allowing users to reason through decisions themselves. Results show that this type of framework that optimizes for both accuracy and human agency can help augment human intelligence.[75]

For example, consider the task of approving loans given the details of a loan application. The standard algorithmic solution would be to offer advice through a prediction or risk score. However, this solution reduces rich data about the application into a single score, losing much information along the way. The researchers offer and test an alternative solution in which the algorithmic advice is offered in the form of an ‘avatar’ that conveys information through a facial expression. This allows the researchers to successfully offer high-dimensional advice by representing multivariate data through the mapping of features to facial components. The results were promising, showing an increase in accuracy as well as an ability for users to reason through decisions when compared to either no advice or arbitrary facial expressions.[76]

Both methods of augmenting human decision making while retaining human agency are ideal for highly complex decision-making contexts. When the context is complex, there are

[73] Ibid.

[74] Zafar et al. "Fairness beyond disparate treatment & disparate impact." Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017.

[75] Hilgard et al. "Learning Representations by Humans, for Humans."

[76] The results showed higher accuracy for the machine itself as well as the predictive advice. However, the researchers hypothesise that the population of people in the study, mTurk workers, were more likely to follow the predictive advice than investors because mTurkers have nothing at stake. Either way, the tradeoff with accuracy of the machine is necessary in examples with high stakes so that the human can retain agency.

multiple, disparate considerations that must be taken into account to make a final decision. A human decision is able to weigh these different considerations well, in a way that a model would not be able to do. A decision augmenting scheme provides a way to enhance the decisions made by humans without projecting a specific outcome along a single dimension, allowing the human to reason through multiple considerations.

Moreover, this interaction scheme works well when the consequences of a mistake are high. In high stakes cases where serious harm can result from mistakes, a system is already in place to take care of questions of responsibility and liability for mistakes with a human decision maker. With the addition of a machine agent, questions of responsibility for potentially fatal mistakes must be addressed by the many different stakeholders involved. From this perspective, the distribution of agency to a model is not ideal in decisions that involve potentially high stakes.

Finally, this use case of machine algorithms is particularly suited to assist humans in increasing accuracy and consistency of decision making. Using the model as an augmenting tool does not necessarily save resources in the way that replacing a human decision maker might. As such, the use of machines as an augmenting tool is not primarily suited as a solution to cut costs, but rather to help with decision making alone.

This use of predictive risk models that retains full human agency is appropriate when the context is highly specific and not easily generalizable, the consequences of mistakes is high on both ends, and the main concern is not to save resources but rather to increase accuracy and consistency of decision making. If the system is designed in such a way that the risk score is simply a risk score offering a recommendation on what decision seems right, then the best use of the system is as a fail-safe mechanism. On the other hand, if the information provided focuses instead on assisting users on how to make a decision, then it would make sense to use this machine model as an informational aide.

3. Discussion of AFST Design

Predictive risk models can be developed in the lab, but in order to be used in practice, they must take into account design principles that fit into existing decision-making pipelines in highly context-specific cases. The design of the granular interaction between the human and the machine should be inherent in the design of the system itself. Too often, predictive risk models that are developed in research settings end up failing in practice.[77] Part of this failure is due to the lack of consideration of how the machine should be used in the decision making in practice.

Central to this discussion should be a clear understanding of the distribution of agency between the human and the machine. Researchers must determine how to use machine models prior to designing and implementing a machine model to increase the likelihood of success. The distribution of agency in given decision-making context should be based on a careful consideration of the generalizability of the context, the gravity of consequences, and the desired outcome of the adoption of a machine model.

3.1 Design of the AFST

The context of a child abuse call screening decision lends itself well to a use of a predictive model that leaves the power of agency to the human call screeners. The ultimate decision made depends on multiple considerations. When call screeners make a decision to screen a family in for investigation, they must weigh considerations of the likely accuracy of an investigation with considerations of fairness, parent privacy, and limited resources amongst other factors.

[77] Yang et al. "Unremarkable AI."

Moreover, the consequences of a mistaken call screen are high. A false positive screen represents a situation in which a screener mistakenly screens a family in for investigation. This situation subjects a family to an intrusive investigation that turns out to be unnecessary for that family. On the other hand, a false negative screen represents a situation in which a call screener mistakenly screens a family out for investigation. This situation is harmful to the child, who lives in a situation where there is evidence of child abuse or extreme neglect but is not receiving the necessary resources to keep them safe. Given the gravity of the consequences, it seems that a paradigm that gives the machine full agency is not desirable.

Finally, the stated goals of the adoption of the AFST are threefold: to increase accuracy, equity, and consistency of call screening decisions.[78] Each of these three goals are concerned with increasing the quality of screening decisions made. It does not seem that a primary aim of the adoption of the tool is to save resources for the department, but rather to increase the effectiveness of the decisions made indicating that a paradigm that leaves agency with human decision-makers is suitable.

Moreover, the tool offers a recommendation of what the call screener should do rather than focusing on how a call screener should make a decision. The tool is created as if it would be a separate agent, however, the tool is given no real agency. There are two recommendations by two separate decision makers, the call screener and the model. At times, these two recommendations conflict with one another.

When the model's recommendation conflicts with the humans', there is a lack of clarity about how competing decisions should be resolved. The human should not defer completely to the model given the decision making context. The model should not defer completely to the human, or else the model adds no value to the decision making process.

[78] Vaithianathan et al. "Developing predictive models," 1.

Based on the context under which call screeners are making decisions, the AFST should be used as either a joint decision maker with humans or as an augmenting tool. However, the creators of the AFST along with the DHS operate under the understanding that the human should retain full agency throughout the decision making process. They repeatedly affirm that the "new tool augments; it does not supplant call screening discretion and decision processes." [79] As such, it seems that the DHS would not pursue a joint decision making scheme where part of the decision making power lies with the machine.

Thus, it seems that an ideal use of the AFST would be as a human augmenting tool, leaving full agency to the human. However, the current design of the AFST as a single risk score that offers a competing recommendation to the call screener suggests that the tool should be used as a fail safe rather than as an information aide. Currently the design of the AFST as a model that outputs a single risk score implies that it should be used as either a full agent or a joint agent. This, however, is in contrast to the ideal human-machine decision-making scheme required by the context of the decision.

3.2 Design Suggestions for the AFST

As is, the AFST can be used as a mere failsafe, offering a check on call screening decisions when the call screening decisions differ from the AFST recommendations. This use case is valuable and can help to increase accuracy and consistency given that the creators of the system are clear that this is the primary use case. It is specific enough that call screeners are aware of how they are supposed to use the score when it differs with their own judgments – simply double check their decisions to see if they missed any important information. For the majority of cases, the score should simply allow the call screener to pass. In the cases of a

[79] Executive Statement from County of Allegheny, County Executive Rich Fitzgerald

conflict, the call screener must simply slow down and review all the details again to see if she missed any facts along the way.

However, it seems like the DHS wants to use the AFST as more than a failsafe. The DHS claims that the goal of the tool is to act as an information aide to human call screeners.[80] Moreover, the use of the AFST as a failsafe raises a few potential issues. First, call screeners might still experience unconscious human biases, such as an anchoring bias or an automation bias.[81] A screener may not be able to make judgements independent of a parallel judgement that comes from the AFST without unconsciously molding their score to the score of the AFST over time.

Moreover, the current design of the AFST is not transparent. The tool itself consists of a risk score, but does not offer a causal explanation as to how the AFST reached a decision, leaving the user without important information that could help in their final decision. A more transparent design would help call screeners assess the trustworthiness of the prediction, thereby increasing calibrated trust of the system. It would also help call screeners reconsider their own judgements in light of the evidence the tool is basing its prediction on, making the failsafe a more effective tool.

Based on these conclusions, I suggest that the tool ought to be redesigned according to its role as an augmenting aide. Even if the tool is to be used as a mere failsafe, the designers should increase the transparency of the machine to allow call screeners to understand how the tool came to its decision. An increase in interpretability of the risk score itself would move in the direction of an information aide which focuses on how to make a decision rather than what decision to make.

[80] Vaithianathan et al. "Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation," 5.
[81] Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

The AFST should be redesigned to focus on the question of how to make a decision rather than what decision to make. This can be done either through an increase in interpretability of the current risk score, or a redesign of the tool that does not offer any risk score, but rather highlights the important features of how to make a screening decision.

A more exact suggestion of how to design the interface of the AFST is out of the scope of this thesis, but should be pursued with the knowledge of the role of the AFST as an augmenting aide. A redesign according to this principle would require additional research into the ways that call screeners currently make decisions as well as the factors that are most predictive with the AFST.

4. Conclusion

Through a case study of the AFST and the public debate surrounding the use of the tool, we can understand better understand the considerations involved in the design of a tool. This chapter offers a new framework for understanding different types of models with an analysis of what characteristics make a context good for one of these uses, separating categories based on different distributions of agency between the human and the model. Ultimately, I argue that the AFST is designed to be used according to either of the first two categories rather than the third, even though the AFST ought to be designed as according to the third.

Despite the current technical design of the AFST, I argue that it should be designed as an augmenting aide, leaving full agency with the call screener. Moreover, I argue that the AFST is meant to be used as a decision support tool that assists call screeners in the decision making process rather than as a mere failsafe. In conclusion, I argue that the AFST is poorly designed to function as a decision support tool, ultimately offering suggestions for how the AFST should be redesigned to perform its intended function.

Works Cited

Allegheny County Department of Human Services. Frequently Asked Questions, https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/FAQs-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-8.pdf

Altman, Andrew, "Discrimination", *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2016/entries/discrimination/.

Angwin, Julia, et al. “Machine Bias.” ProPublica, 9 Mar. 2019, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Benza, Raymond L., et al. "An evaluation of long-term survival from time of diagnosis in pulmonary arterial hypertension from the REVEAL Registry." Chest 142.2 (2012): 448-456.

Burrell, Jenna. "How the machine ‘thinks’: Understanding opacity in machine learning algorithms." Big Data & Society 3.1 (2016): 2053951715622512.

Dare, Tim, and Eileen Gambrill. “Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County.” Apr. 2017, https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf.

Eubanks, Virginia. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018, 108-9.

Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* 542.7639 (2017): 115.

Francis, David. “Poverty and Mistreatment of Children Go Hand in Hand.” Poverty and Mistreatment of Children Go Hand in Hand, https://www.nber.org/digest/jan00/w7343.html

Goldhaber-Fiebert, Jeremy, and Lea Prince. "Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County’s Child Welfare Office." Pittsburgh: Allegheny County.[Google Scholar] (2019).

Hellman, Deborah. *When is discrimination wrong?*. Harvard University Press, 2008.

Hilgard, Sophie, et al. "Learning Representations by Humans, for Humans." arXiv preprint arXiv:1905.12686 (2019).

Hornby Zeller Associates. “Allegheny County Predictive Risk Modeling Tool Implementation: Process Evaluation.” 2018.

Kahneman, Daniel. *Thinking, fast and slow.* Macmillan, 2011.

Madras, David, Toni Pitassi, and Richard Zemel. "Predict responsibly: improving fairness and accuracy by learning to defer." *Advances in Neural Information Processing Systems*. 2018.

Panattoni, Laura E., et al. "Predictive risk modelling in health: options for New Zealand and Australia." *Australian Health Review* 35.1 (2011): 45-51.

Stevenson, Megan T., and Jennifer L. Doleac. "Algorithmic Risk Assessment in the Hands of Humans." Available at SSRN (2019).

Vaithianathan, Rhema, et al. "Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation." Center for Social data Analytics (2017).

Yang, Qian, Aaron Steinfeld, and John Zimmerman. "Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019.

Zafar, Muhammad Bilal, et al. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment." *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017.