A Child Abuse Prediction Model Fails Poor Families

wired.com/story/excerpt-from-automating-inequality/

Virginia Eubanks January 15, 2018

It's late November 2016, and I'm squeezed into the far corner of a long row of gray cubicles in the call screening center for the Allegheny County Office of Children, Youth and Families (CYF) child neglect and abuse hotline. I'm sharing a desk and a tiny purple footstool with intake screener Pat Gordon. We're both studying the Key Information and Demographics System (KIDS), a blue screen filled with case notes, demographic data, and program statistics. We are focused on the records of two families: both are poor, white, and living in the city of Pittsburgh, Pennsylvania. Both were referred to CYF by a mandated reporter, a professional who is legally required to report any suspicion that a child may be at risk of harm from their caregiver. Pat and I are competing to see if we can guess how a new predictive risk model the county is using to forecast child abuse and neglect, called the Allegheny Family Screening Tool (AFST), will score them.

The stakes are high. According to the US Centers for Disease Control and Prevention, approximately one in four children will experience some form of abuse or neglect in their lifetimes. The agency's Adverse Childhood Experience Study concluded that the experience of abuse or neglect has "tremendous, lifelong impact on our health and the quality of our lives," including increased occurrences of drug and alcohol abuse, suicide attempts, and depression.

Excerpted from Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, released this week by St. Martin's Press.

In the noisy glassed-in room, Pat hands me a double-sided piece of paper called the "Risk/Severity Continuum." It took her a minute to find it, protected by a clear plastic envelope and tucked in a stack of papers near the back of her desk. She's worked in call screening for five years, and, she says, "Most workers, you get this committed to memory. You just know." But I need the extra help. I am intimidated by the weight of this decision, even though I am only observing. From its cramped columns of tiny text, I learn that kids under five are at greatest risk of neglect and abuse, that substantiated prior reports increase the chance that a family will be investigated, and that parent hostility toward CYF investigators is considered high risk behavior. I take my time, cross-checking information in the county's databases against the risk/severity handout while Pat rolls her eyes at me, teasing, threatening to click the big blue button that runs the risk model.

The first child Pat and I are rating is a six-year-old boy I'll call Stephen. Stephen's mom, seeking mental health care for anxiety, disclosed to her county-funded therapist that someone—she didn't know who—put Stephen out on the porch of their home on an early November day. She found him crying outside and brought him in. That week he began to act out, and she was concerned that something bad had happened to him. She confessed to her therapist that she suspected he might have been abused. Her therapist reported her to the state child abuse hotline.

about the author

Virginia Eubanks is Associate Professor of Political Science at the University at Albany, SUNY, a founding member of the Our Data Bodies project, and a fellow at New America.

But leaving a crying child on a porch isn't abuse or neglect as the state of Pennsylvania defines it. So the intake worker screened out the call. Even though the report was unsubstantiated, a record of the call and the call screener's notes remain in the system. A week later, an employee of a homeless services agency reported Stephen to a hotline again: He was wearing dirty clothes, had poor hygiene, and there were rumors that his mother was abusing drugs. Other than these two reports, the family had no prior record with CYF.

The second child is a 14-year-old I'll call Krzysztof. On a community health home visit in early November, a case manager with a large nonprofit found a window and a door broken and the house cold. Krzysztof was wearing several layers of clothes. The caseworker reported that the house smelled like pet urine. The family sleeps in the living room, Krzysztof on the couch and his mom on the floor. The case manager found the room "cluttered." It is unclear whether these conditions actually meet the definition of child neglect in Pennsylvania, but the family has a long history with county programs.

An Issue of Definition

No one wants children to suffer, but the appropriate role of government in keeping kids safe is complicated. States derive their authority to prevent, investigate, and prosecute child abuse and neglect from the Child Abuse and Prevention and Treatment Act, signed into law by President Richard Nixon in 1974. The law defines child abuse and neglect as the "physical or mental injury, sexual abuse, negligent treatment, or maltreatment of a child ... by a person who is responsible for the child's welfare under circumstances which indicate that the child's health or welfare is harmed or threatened."

Even with recent clarifications that the harm must be "serious," there is considerable room for subjectivity in what exactly constitutes neglect or abuse. Is spanking abusive? Or is the line drawn at striking a child with a closed hand? Is letting your children walk to a park down the block alone neglectful? Even if you can see them from the window?

The first screen of the list of conditions classified as maltreatment in KIDS illustrates just how much latitude call screeners have to classify parenting behaviors as abusive or neglectful. It includes: abandoned infant; abandonment; adoption disruption or dissolution; caretaker's inability to cope; child sexually acting out; child substance abuse; conduct by parent that places child at risk; corporal punishment; delayed/denied healthcare; delinquent act by a child under 10 years of age; domestic violence; educational neglect; environmental toxic substance; exposure to hazards; expulsion from home; failure to protect; homelessness; inadequate clothing, hygiene, physical care or provision of food; inappropriate caregivers or discipline; injury caused by another person; and isolation. The list scrolls on for several more screens.

Three-quarters of child welfare investigations involve neglect rather than physical, sexual, or emotional abuse. Where the line is drawn between the routine conditions of poverty and child neglect is particularly vexing. Many struggles common among poor families are officially defined as child maltreatment, including not having enough food, having inadequate or unsafe housing, lacking medical care, or leaving a child alone while you work. Unhoused families face particularly difficult challenges holding on to their children, as the very condition of being homeless is judged neglectful.

In Pennsylvania, abuse and neglect are fairly narrowly defined. Abuse requires bodily injury resulting in impairment or substantial pain, sexual abuse or exploitation, causing mental injury, or imminent risk of any of these things. Neglect must be a "prolonged or repeated lack of supervision" serious enough that it "endangers a child's life or development or impairs the child's functioning." So, as Pat and I run down the risk/severity matrix, I think both Stephen and Krzysztof should score pretty low.

In neither case are there reported injuries, substantiated prior abuse, a record of serious emotional harm, or verified drug use. I'm concerned about the inadequate heat in teenaged Krzysztof's house, but I wouldn't say that he is in imminent danger. Pat is concerned that there have been two calls in two weeks on six-year-old Stephen. "We literally shut the door

behind us and then there was another call," she sighs. It might suggest a pattern of neglect or abuse developing—or that the family is in crisis. The call from a homeless service agency suggests that conditions at home deteriorated so quickly that Stephen and his mom found themselves on the street. But we agree that for both boys, there seems to be low risk of immediate harm and few threats to their physical safety.

On a scale of 1 to 20, with 1 being the lowest level of risk and 20 being the highest, I guess that Stephen will be a 4, and Krzysztof a 6. Gordon smirks and hits the button that runs the AFST. On her screen, a graphic that looks like a thermometer appears: It's green down at the bottom and progresses up through yellow shades to a vibrant red at the top. The numbers come up exactly as she predicted. Stephen, the six-year-old who may have suffered sexual abuse and is possibly homeless, gets a 5. Krzysztof, the teenager who sleeps on the couch in a cold apartment? He gets a 14.

Oversampling the Poor

Faith that big data, algorithmic decision-making, and predictive analytics can solve our thorniest social problems—poverty, homelessness, and violence—resonates deeply with our beliefs as a culture. But that faith is misplaced. On the surface, integrated data and artificial intelligence seem poised to produce revolutionary changes in the administration of public services. Computers apply rules to every case consistently and without prejudice, so proponents suggest that they can root out discrimination and unconscious bias. Number matching and statistical surveillance effortlessly track the spending, movements, and life choices of people accessing public assistance, so they can be deployed to ferret out fraud or suggest behavioral interventions. Predictive models promise more effective resource allocation by mining data to infer future actions of individuals based on behavior of "similar" people in the past.

These grand hopes rely on the premise that digital decision-making is inherently more transparent, accountable, and fair than human decision-making. But, as data scientist Cathy O'Neil has written, "models are opinions embedded in mathematics." Models are useful because they let us strip out extraneous information and focus only on what is most critical to the outcomes we are trying to achieve. But they are also abstractions. Choices about what goes into them reflect the priorities and preoccupations of their creators. The Allegheny Family Screening Tool is no exception.

The AFST is a statistical model designed by an international team of economists, computer scientists, and social scientists led by Rhema Vaithianathan, professor of Economics at the University of Auckland, and Emily Putnam-Hornstein, director of the Children's Data Network at the University of Southern California. The model mines Allegheny County's vast data warehouse to try and predict which children might be victims of abuse or neglect in the future. The warehouse contains more than a billion records—an average of 800 for every resident of the county—provided by regular data extracts from a variety of public agencies,

including child welfare, drug and alcohol services, Head Start, mental health services, the county housing authority, the county jail, the state's Department of Public Welfare, Medicaid, and the Pittsburgh public schools.

The job of intake screeners like Pat Gordon is to decide which of the 15,000 child maltreatment reports the county receives each year to refer to a caseworker for investigation. Intake screeners interview reporters, examine case notes, burrow through the county's data warehouse, and search publically-available data such as court records and social media to determine the nature of the allegation against the caregiver and to ascertain the immediate risk to the child. Then, they run the model.

A regression analysis performed by the Vaithianathan team suggested that there are 131 indicators available in the county data that are correlated with child maltreatment. The AFST produces its risk score—from 1 (low risk) to 20 (highest risk)—by weighing these "predictive variables." They include: receiving county health or mental health treatment; being reported for drug or alcohol abuse; accessing supplemental nutrition assistance program benefits, cash welfare assistance, or Supplemental Security Income; living in a poor neighborhood; or interacting with the juvenile probation system. If the screener's assessment and the model's score clash, the case is referred to a supervisor for further discussion and a final screening decision. If a family's AFST risk score is high enough, the system automatically triggers an investigation.

Human choices, biases, and discretion are built into the system in several ways. First, the AFST does not actually model child abuse or neglect. The number of child maltreatment–related fatalities and near fatalities in Allegheny County is thankfully very low. Because this means data on the actual abuse of children is too limited to produce a viable model, the AFST uses proxy variables to stand in for child maltreatment. One of the proxies is community re-referral, when a call to the hotline about a child was initially screened out but CYF receives another call on the same child within two years. The second proxy is child placement, when a call to the hotline about a child is screened in and results in the child being placed in foster care within two years. So, the AFST actually models decisions made by the community (which families will be reported to the hotline) and by CYF and the family courts (which children will be removed from their families), not which children will be harmed.

The AFST's designers and county administrators hope that the model will take the guesswork out of call screening and help to uncover patterns of bias in intake screener decision-making. But a 2010 study of racial disproportionality in Allegheny County CYF found that the great majority of disproportionality in the county's child welfare services actually arises from referral bias, not screening bias. Mandated reporters and other members of the community call child abuse and neglect hotlines about black and biracial families three and a half times more often as they call about white families. The AFST focuses all its predictive

power and computational might on call screening, the step it can experimentally control, rather than concentrating on referral, the step where racial disproportionality is actually entering the system.

More troubling, the activity that introduces the most racial bias into the system is the very way the model defines maltreatment. The AFST does not average the two proxies, which might use the professional judgment of CYF investigators and family court judges to mitigate some of the disproportionality coming from community referral. The model simply uses whichever number is higher.

Second, the system can only model outcomes based on the data it collects. This may seem like an obvious point, but it is crucial to understanding how Stephen and Krzysztof got such wildly disparate and counterintuitive scores. A quarter of the variables that the AFST uses to predict abuse and neglect are direct measures of poverty: they track use of means-tested programs such as TANF, Supplemental Security Income, SNAP, and county medical assistance. Another quarter measure interaction with juvenile probation and CYF itself, systems that are disproportionately focused on poor and working-class communities, especially communities of color. Though it has been billed as a crystal ball for predicting child harm, in reality the AFST mostly just reports how many public resources families have consumed.

Allegheny County has an extraordinary amount of information about the use of public programs. But the county has no access to data about people who do not use public services. Parents accessing private drug treatment, mental health counseling, or financial support are not represented in DHS data. Because variables describing their behavior have not been defined or included in the regression, crucial pieces of the child maltreatment puzzle are omitted from the AFST.

Geographical isolation might be an important factor in child maltreatment, for example, but it won’t be represented in the data set because most families accessing public services in Allegheny County live in dense urban neighborhoods. A family living in relative isolation in a well-off suburb is much less likely to be reported to a child abuse or neglect hotline than one living in crowded housing conditions. Wealthier caregivers use private insurance or pay out of pocket for mental health or addiction treatment, so they are not included in the county’s database.

Imagine the furor if Allegheny County proposed including monthly reports from nannies, babysitters, private therapists, Alcoholics Anonymous, and luxury rehabilitation centers to predict child abuse among middle-class families. “We really hope to get private insurance data. We’d love to have it,” says Erin Dalton, director of Allegheny County’s Office of Data Analysis, Research and Evaluation. But, as she herself admits, getting private data is likely impossible. The professional middle class would not stand for such intrusive data gathering.

The privations of poverty are incontrovertibly harmful to children. They are also harmful to their parents. But by relying on data that is only collected on families using public resources, the AFST unfairly targets low-income families for child welfare scrutiny. “We definitely oversample the poor,” says Dalton. “All of the data systems we have are biased. We still think this data can be helpful in protecting kids.”

We might call this poverty profiling. Like racial profiling, poverty profiling targets individuals for extra scrutiny based not on their behavior but rather on a personal characteristic: They live in poverty. Because the model confuses parenting while poor with poor parenting, the AFST views parents who reach out to public programs as risks to their children.

False Positives—and Negatives

The hazards of using inappropriate proxies and inadequate datasets may be inevitable in predictive modeling. And if a child abuse and neglect investigation was a benign act, it might not matter that the AFST is imperfectly predictive. But a child abuse and neglect investigation can be an intrusive, frightening event with lasting negative impacts.

The state of Pennsylvania’s goal for child safety—“Being free from immediate physical or emotional harm”—can be difficult to reach, even for well-resourced families. Each stage of a CYF investigation introduces the potential for subjectivity, bias, and the luck of the draw. “You never know exactly what’s going to happen,” says Catherine Volponi, director of the Juvenile Court Project, which provides pro bono legal support for parents facing CYF investigation or termination of their parental rights. “Let’s say there was a call because the kids were home alone. Then they’re doing their investigation with mom, and she admits marijuana use. Now you get in front of a judge who, perhaps, views marijuana as a gateway to hell. When the door opens, something that we would not have even been concerned about can just mushroom into this big problem.”

At the end of each child neglect or abuse investigation, a written safety plan is developed with the family, identifying immediate steps that must be followed and long-term goals. But each safety action is also a compliance requirement, and sometimes, factors outside parents’ control make it difficult for them to implement their plan. Contractors who provide services to CYF-involved families fail to follow through. Public transportation is unreliable. Overloaded caseworkers don’t always manage to arrange promised resources. Sometimes parents resist CYF’s dictates, resenting government intrusion into their private lives.

Failure to complete your plan—regardless of the reason—increases the likelihood that a child will be removed to foster care. “We don’t try to return CYF families to the level at which they were operating before,” concludes Volponi, “We raise the standard on their parenting, and then we don’t have enough resources to keep them up there. It results in epic failures too much of the time.”

Human bias has been a problem in child welfare since the field's inception. The designers of the model and DHS administrators hope that, by mining the wealth of data at their command, the AFST can help subjective intake screeners make more objective recommendations. But human bias is built in to the predictive risk model. Its outcome variables are proxies for child harm; they don't reflect actual neglect and abuse. The choice of proxy variables, even the choice to use proxies at all, reflects human discretion. The AFST's predictive variables are drawn from a limited universe of data that includes only information on public resources. The choice to accept such limited data reflects the human discretion embedded in the model—and an assumption that middle-class families deserve more privacy than poor families.

Once the big blue button is clicked and the AFST runs, it manifests a thousand invisible human choices under a cloak of evidence-based objectivity and infallibility. Proponents of the model insist that removing discretion from call screeners is a brave step forward for equity, transparency, and fairness in government decision-making. But the AFST doesn't remove human discretion; it simply moves it. In the past, the mostly working-class women in the call center exerted some control in agency decision-making. Today, Allegheny County is deploying a system built on the questionable premise that an international team of economists and data analysts is somehow less biased then the agency's own employees.

Back in the call center, I mention to Pat Gordon that I've been talking to CYF-involved parents about how the AFST might impact them. Most parents, I tell her, are concerned about false positives: the model rating their child at high risk of abuse or neglect when little risk actually exists. I see how Krzysztof 's mother might feel this way if she was given access to her family's risk score.

But Pat reminds me that Stephen's case poses equally troubling questions. I should also be concerned with false negatives—when the AFST scores a child at low risk though the allegation or immediate risk to the child might be severe. "Let's say they don't have a significant history. They're not active with us. But [the allegation] is something that's very egregious. [CYF] gives us leeway to think for ourselves. But I can't stop feeling concerned that ... say the child has a broken growth plate, which is very, very highly consistent with maltreatment ... there's only one or two ways that you can break it. And then [the score] comes in low!"

The screen that displays the AFST risk score states clearly that the system "is not intended to make investigative or other child welfare decisions." Rhema Vaithianathan told me in February 2017 that the model is designed in such a way that intake screeners are encouraged to question its predictive accuracy and defer to their own judgment. "It sounds contradictory, but I want the model to be slightly undermined by the call screeners," she said. "I want them to be able to say, this [screening score] is a 20, but this allegation is so minimal that [all] this model is telling me is that there's history."

The pairing of the human discretion of intake screeners like Pat Gordon with the ability to dive deep into historical data provided by the model is the most important fail-safe of the system. Toward the end of our time together in the call center, I asked Pat if the harm false negatives and false positives might cause Allegheny County families keeps her up at night. “Exactly,” she replied. “I wonder if people downtown really get that. We’re not looking for this to do our job. We’re really not. I hope they get that.” But like Uber’s human drivers, Allegheny County call screeners may be training the algorithm meant to replace them.