Black Box Justice

Richard Berk designs computer algorithms that predict crime. As courts and cops increasingly use his and similar tools to shape everything from parole decisions to street policing, Berk has a warning: accuracy comes at the cost of fairness, and citizens must decide where justice lies.

BY TREY POPP

ON June 26, 2010, Michael Eric Ballard walked out of the Allentown Community Corrections Center in a blue Superman shirt and a vindictive mood. The ACCC was a halfway house for Pennsylvania parolees. Ballard, who was 36, had been placed there after a two-year imprisonment for failing to complete the anger-management therapy mandated by his first parole. He had originally been locked up for the murder of a 56-year-old man in 1991, to which he pled guilty. The victim, according to Ballard, had made unwanted sexual advances. Ballard stabbed him 13 times, then took the older man's car and credit cards on a spending spree stretching from Pennsylvania to Arkansas. Pennsylvania sentenced him to a minimum of 15 years, and paroled him when that term was up.

PHOTO BY DUSTIN FENSTERMACHER

Ballard left the halfway house carrying a radio. He was angry about a phone call the previous evening, which had fueled his suspicions that a woman he'd been dating, Denise Merhi, was seeing another man. According to court records, he took the radio to a couple of pawnshops and tried to exchange it for a knife. When neither dealer bit, he bought one instead. A public bus took him to Northampton, a small borough on the Lehigh River. After prowling around the back alley of the house where Merhi lived, Ballard went in and stabbed her to death, along with her father, grandfather, and a neighbor who rushed to their aid. He left covered in blood—some of it his own, from a self-inflicted wound to his knee—and fled in Merhi's Pontiac, which he crashed into trees lining a nearby highway.

When a police officer responding to the accident asked Ballard where he was coming from, Ballard replied, "I just killed everyone." He confessed again at a local hospital—to the doctor who treated his wounds and to another state trooper. "Make your report," he told the officer. "I'll plead guilty."

Then he added, "Blame the parole board."

There's little doubt that many residents of Northampton County did. They would have had plenty of company. The month of Ballard's deadly rampage, 29 people were murdered in Philadelphia. They ranged from a 16-month-old boy beaten to death by his mother to a 61-year-old man shot for his money and showing "disrespect." In those two cases, everyone but the toddler was a repeat offender. Odds are high that parolees and probationers were overrepresented in the rest, too, among both the killers and the dead. According to Lawrence Sherman, who chaired Penn's criminology department between 2003 and 2007 ["A Passion for Evidence," Mar|Apr 2000], some 55 of the 344 people murdered by gun in Philadelphia in 2006 were under supervision by the city's adult parole and probation department (APPD). Another 53 were arrested for homicides.

As Sherman wrote at the time, probationers in Philadelphia were murdered in 2006 at "20 times the national homicide rate." The tally of homicide arrests suggests that other probationers were a big reason why.

But that statistic runs up against another: the APPD was supervising 52,000 people that year. "Most Philadelphia probationers have very low risks of killing or being killed," Sherman emphasized. "Only about 1 or 2 percent of them drive the overall caseload risk to such a high level."

The problem with blaming the parole board is that for every Michael Eric Ballard, there are a hundred offenders who will never do anything like he did. Ballard represented the costliest kind of error the criminal justice system can make, a *false negative*—someone estimated to be less of a danger than he proved to be. Four people lost their lives because of it. But false negatives can exact an additional cost: they can pressure decision-makers to err in the opposite direction, driving up prison populations with *false positives*—people who wouldn't actually be a menace to public safety. Society pays a price for false positives, too. Not only is prison expensive, but some evidence indicates that time inside can increase the future criminality of inmates.

The APPD's leadership knew this all too well. In fact, they had approached Penn's criminology department for help. They hoped to develop a tool capable of distinguishing the high-risk cases from all the rest. That might permit them to concentrate limited supervisory resources where they were most needed—and maybe even loosen the leash on individuals deemed low-risk.

Due partly to timing and partly to a budding trend in the field, the project came to be dominated by a somewhat unlikely figure. Richard Berk, who joined Penn's faculty in 2006, represents a new species of criminologist: the kind who claims not to be one.

"I don't really know that much criminology," he said recently, sitting at a glass table in his book-lined office in the McNeil Building. "I've never had a criminology course, never taught a criminology course." Reminded that he currently chairs the department, he insisted that he is merely an administrator. "I have sort of a layperson's knowledge, after years of hanging around with criminologists. But, no, I'm not really a criminologist."

Berk prefers to be known as a statistician with a yen for machine learning. Machine learning is a type of artificial intelligence that enables computers to "learn without being explicitly programmed," as one of the field's pioneers put it. Your email junk filter probably uses some form of machine learning. So does Google's text-translation tool, and Apple's virtual personal assistant Siri.

Berk wanted to use it to predict crime.

Berk doesn't come across like someone who calculates p values for a living. He looks like he'd rather be tackling someone. He has a horseshoe jaw, a barrel chest, and probes questioners with a staring-contest gaze. He relishes sparring over the controversial aspects of his work—perhaps because rhetorical clashes are the closest he can get to the contact sports that shaped his younger years. Berk played football at Yale in the early 1960s, picked up rugby after college, and then got into judo and martial arts. Now he calls himself an "orthopedic disaster" with "a metal knee, a metal shoulder, and a fused spine." Berk is 74 years old, but a particular kind of 74. Kodiak bears can live a long time too, and even an old, scarred-up one can maul you.

Berk tackled Philadelphia's probation-and-parole challenge with a computer-modeling technique called "random forests." Here's how it works. Berk gathered a massive amount of data about 30,000 probationers and parolees who'd been free for at least two years. He fed it into an algorithm that randomly selected different combinations of variables, and fit the information to a known outcome: whether someone had been charged with homicide

or attempted homicide in that time frame. The algorithm repeated this process hundreds of times, producing a "forest" of individual regression trees that took arbitrary paths through the data. For instance, one tree might begin by considering parolees' ages, then the number of years that had passed since their last serious offense, then their current residential ZIP code, then their age at the time of their first juvenile offense, then ZIP code (again), then the total number of days they had been incarcerated, and so on, creating a sort of flow chart that sorts any given individual into a category: *homicide*, or *no homicide*. Another tree would follow the same procedure, but using different combinations of variables in a different order.

To test the predictive power of this forest, Berk then fed it data on 30,000 different cases—whose outcomes were also known, but which had not been used to build the model. Each was assessed by every tree in the forest, which cast a "vote" on the likelihood that the individual would try to kill again. Those votes were tabulated to generate a final forecast for each case. Importantly, the forest is a black box; there's no way to know how—let alone why—it arrives at any given prediction.

Assessing a prediction's value is tricky. Out of the 30,000 individuals in the test sample, 322 had actually been charged with homicide or attempted homicide within two years. So simply predicting that any given person would not kill again would make you right 99 percent of the time. But that would prevent no deaths. A standard logistic regression using the same data, by comparison, fingered two out of 30,000 subjects as likely to commit murder, and it was right about one of them. Not very impressive, but at least it might have saved one life.

Berk's algorithm was in a different universe. It forecasted that 27,914 individuals would not attempt murder within two years, and it was right about 99.3 percent of them. It identified 1,764 as at risk for killing, 137 of whom in fact faced homicide charges. Generating a prediction for any given individual, using data already available to criminal-justice decision-makers, took "just 10 or 15 seconds," according to a subsequent review.

Berk's crime prediction algorithms are black boxes. There's no way to know how—let alone why—they arrive at a prediction.

This was the kind of tool the APPD could use. Only not exactly the way Berk had designed it.

Berk believes in data—virtually any kind of data, no matter how tangentially it may relate to crime. "I'm not trying to explain criminal behavior, I'm trying to forecast it," he likes to say. "If shoe size or sunspots predict that someone's going to commit a homicide, I want to use that information—even if I have no idea why it works."

Berk didn't use sunspots in his algorithm, but he used more information than a judge is meant to consider when estimating an appellant's potential future dangerousness. In addition to information about a given individual's history—like age of first contact with the adult court system, prior gun-related convictions, and number of psychiatric conditions imposed in previous adjudications—he included race, the proportion of African Americans in an offender's ZIP code, the median income in the offender's ZIP code, and other features of reality over which the individual had little or no control.

This makes many people uneasy. "Our law punishes people for what they do, not who they are," wrote Supreme Court Chief Justice John Roberts earlier this year, in a majority opinion ruling that a Texas death-row inmate's lawyers had unconstitutionally presented expert testimony suggesting that their client was more likely to commit future acts of violence because he was black. "Dispensing punishment on the basis of an immutable characteristic flatly contravenes this guiding principle."

Berk says it's up to judges to decide what contravenes the law, but he has little patience with arguments for ignoring data that might improve the accuracy of a prediction. After all, the whole point of forecasting future dangerousness—which has long been a mandatory consideration in determinations about pre-trial detention, bail, and parole—is to protect the public.

"Say I have two identical people who have been arrested and convicted of a crime, let's say burglary," Berk says. "One happens to live in my neighborhood in Mount Airy; another one lives in Germantown … The data is crystal clear that the kid who's released in Germantown is more likely to reoffend."

But is it fair to deprive someone of liberty just because life dealt him a crummy address?

"How many more homicides are you prepared to tolerate for me to drop that variable?" Berk retorts. "A hundred? Fifty? Twenty-five? You make that choice. Remember, I'm predicting as well as I can predict. If you won't let me use that information, I'm going to predict less well."

Furthermore, the question can be pointed in the other direction: Is it fair to subject the law-abiding residents of Germantown to higher-risk parolees than Mount Airy has to deal with?

These were not the only questions raised by Berk's method, but in 2007 the APPD used it to conduct a real-world experiment. It took roughly 1,600 parolees and probationers forecast to be low-risk, and randomly assigned them to one of two supervisory approaches. Some got the standard treatment: monthly visits, on average, with parole officers who could order drug tests, intervention services, and conduct infrequent field visits. Others met their parole officers every six months, phoned in a couple times a year, and could only be drug-tested by court order or their own request. Standard-treatment supervisory officers had caseloads of 150. Experimental-treatment officers oversaw 400 offenders.

Over the next 12 months, there was no meaningful difference in recidivism between the two groups. With the confidence gained from Berk's algorithm, parole officers could effectively double their productivity when supervising low-risk individuals, which would free up manpower to focus on higher-risk cases.

In short order, the APPD began reorganizing its supervisory procedures, and implemented Berk's algorithm for all its incoming cases.

"We are always looking for better and smarter ways to deploy our limited resources," says APPD director Charles Hoyt. "Dr. Berk's model provided us with a reliable, data-driven method of assessing our probationers' and parolees' likelihood of reoffending that could be customized to fit our offender population and agency characteristics, while also being easy to implement and sustain."

The operational model—which has gone through three iterations since its 2009 debut, each using a different blend of variables—did away with explicitly race- and wealth-related data.

"This is a difficult decision for any criminal-justice agency using actuarial techniques to help it achieve its mission," Hoyt says. "While we understand that these indicators make additional statistical contributions in producing accurate forecasts, we decided not to include several demographic and contextual predictors in our risk tool. We found that relying on criminal-justice data commonly used in decision-making at other stages of the justice process was sufficient to achieve accurate predictions."

(There is a statistical technique that quantifies how much accuracy is lost if a specific factor is excluded. In his original model, Berk determined that excluding race led to a roughly 2 percent increase in forecasting error. By comparison, dropping age increased forecasting error by 12 percent. Age of first contact with the adult court system, prior firearm-related convictions, gender, and total prior violent offenses were all more important predictive variables than race. Less-important factors included the total number of prior incarcerations, prior drug convictions, psychiatric adjudications, and whether the specific case in question involved violence. But for any variable, the accuracy loss is "just an association," Berk says. "I still don't know why … I have no explanatory power.")

The APPD has used offenders' residential ZIP codes, however, in each iteration of the tool. Which means that demographic and contextual variables aren't gone completely.

"Because of the highly clustered and segregated nature of housing, knowing what your address is will tell me a lot about your wealth, your education, your race," says Charles Loeffler, the Jerry Lee Assistant Professor of Criminology, who studies the effects of criminal-justice processes on life-course outcomes. So even if those variables are dropped from the algorithm, "the information sneaks in."

Some critics of these kinds of predictive algorithms object to what they see as a more fundamental unfairness: the fact that they explicitly judge individuals on the basis of what unrelated people have done. "The technocratic framing of [these instruments] should not obscure an inescapable truth," writes Sonja Starr, a professor at the University of Michigan Law School and leading scholar on this issue. "[S]entencing based on such instruments amounts to overt discrimination based on demographics and socioeconomic status. The instruments' use of gender and socioeconomic variables, in particular, raises serious constitutional concerns."

"I get it all the time from lawyers," Berk says. "The perspective is that what we're supposed to do is sentence the individual. I don't know what that means. But they keep coming back to that: *It's the individual, not the group*.

"Well, if a person's standing in front of you, and they have three prior convictions, and they've had a drug problem and whatever, you look at them and say, as a judge, *I've seen people like you before. I kind of know what you're like. I'm going to sentence you like them*. How else can you do it? So the problem with the judges is no different," he says. They're black boxes, too, and just as prone to substituting group judgments for individualized ones. "Except they don't have the data," Berk emphasizes. "The judges just have their intuition and their experience. That's not transparent. You can't look into that."

There are other ways in which algorithms may reproduce, and possibly amplify, existing biases in the criminal-justice system. For example, the APPD's algorithm is heavily dependent on data about arrests and charges, rather than convictions. Berk acknowledges that neither variable is perfect. "Convictions," he says, "fold in all of the plea negotiations and all the other things that follow from arrest," and are thus substantially a product of prosecutorial priorities, resources, and legal gamesmanship—not just an offender's behavior. "Arrests have a downside," he says, "because there isn't much oversight, and police make arrests for a variety of reasons."

But for Berk, the drawbacks matter less than where the rubber meets the road: "We can run the two and see which predicts better. And arrests predict better."

Yet the road to conviction is lined with Constitutional protections, including the

right of the accused to a jury trial. Certain kinds of arrests, by contrast, may have as much to do with biased patterns of police deployment as with actual levels of crime.

Using arrests as a predictive factor in decisions about punishment is "problematic," says David Rudovsky, a civil-rights attorney and senior fellow at Penn Law.

"Black kids get arrested a lot more frequently for drugs—not because they use drugs more frequently, but because that's where the cops are," he says, referring to evidence that rates of drug use and selling are comparable across racial lines. "If you had as many cops in the Penn dorms as you have up in North Philadelphia, you'd have a different arrest pool … So if you count the arrest rate for a kid who's got four drug arrests, I'm not sure how fair that is."

And the stakes of Big Data criminal-justice algorithms extend beyond parole and probation decisions. Berk is working on another one to aid Pennsylvania judges at the sentencing stage. He designed one for the Occupational Safety and Health Administration to guide the federal agency's selection of dangerous workplaces to inspect, in order to more efficiently reduce injuries and deaths. (That one wasn't implemented, but as Penn Law professor Cary Coglianese and David Lehr C'16 detailed in a recent article in the *Georgetown Law Review*, titled "Regulating By Robot," machine-learning algorithms are beginning to slide into the role of democracy's enforcers. The IRS has used them to aid its auditing and collection functions, the Environmental Protection Agency has used them to help identify potentially toxic chemical compounds for further testing by traditional means, and agencies ranging from the Food and Drug Administration to the Securities and Exchange Commission have explored what appear to be an enormous range of applications. "Academic researchers," Coglianese and Lehr noted, "have demonstrated how machine-learning algorithms can be used to predict cases of financial-statement fraud, electoral fraud, and even illegal fishing practices.")

Federal regulation may be the next frontier, but criminal justice is the current one. Algorithmic risk-assessment tools are proliferating. States as varied as Louisiana, Colorado, Delaware, and Wisconsin use them during criminal sentencing. Cities like Chicago and Los Angeles have incorporated them into policing tactics.

Berk is currently working with researchers in Norway to create the granddaddy of all criminal-justice algorithms. Using that nation's famously comprehensive civilian records, he aims to predict, from the moment of birth—"or even before"—whether people will commit a crime by their 18th birthday.

"We're getting better and better at this," he proclaimed in a presentation at the Chicago Ideas Week festival in 2012. "We're not in the world of *Minority Report* yet," he said, referring to the sci-fi flick in which Tom Cruise arrests offenders in advance of their crimes, "but I think there's no question we're heading there. The only question is how fast, and what sort of oversight we're going to provide."

Berk seems at times to revel in such provocations, and projects a nonchalant immunity from the Luddite anxieties they predictably spark. "I think it's the same problem that people have with driverless cars," he muses. "They're just not prepared to acknowledge that the technology can perform that well. And even if they do, then they start worrying about, well, where are humans in all this?"

But the confidence he expresses in his digital crystal ball springs from a surprising source: deep uncertainty.

Berk's criminology background is broader than he lets on. As an undergraduate at Yale, he studied under Neal Miller, a pioneer of experimental psychology whose interests included the mechanisms of aggression. Violence was the original focus of Berk's doctoral studies in sociology at Johns Hopkins, during which he worked for three years as a social worker interacting with gangs in Baltimore. "The violence was real," he recalls, "and it was tragic—but it was also fascinating."

He has vivid memories of the riots that rocked Baltimore in 1968, after the assassination of Martin Luther King. "The city was burning," he says. "And they put a bunch of us on the streets to try to keep my white kids and the black kids, who were just a couple blocks away, from getting into it … The Black Panther party was really active, and they were trying to keep the peace, and I knew some of those people. So we were able to keep them separated. It would have been a terrible bloodbath."

But he felt like he was merely forestalling the inevitable. "His" kids were mainly poor whites from Appalachia. "Many of them were terribly disadvantaged economically, but they didn't have the excuse of discrimination," he says. "They were where they were often because their parents were just not very capable. So we had a lot of kids who really couldn't read, and stuff like that … It fed into a cynicism. I mean, you see these kids, and you know they have no future—13- or 14-year-old kid, you know they're doomed. You just hope you can get them some kind of menial job, even if it's at a car wash, so they can earn a living. But you know they're doomed. And nothing that was being proposed was going to make any difference."

Berk has published scores of papers on crime—exploring phenomena as varied as the relationship between race and crack-cocaine charging practices in Los Angeles, sexual harassment in the workplace, and the deterrent effects of arrest for domestic assault. But the longer he's been at it, the warier he has become about the quest to reduce crime by identifying its root causes.

"Since the 1930s," he says, "there's been studies about neighborhoods and crime. Very smart people have gone after it with

all kinds of data, and it only sort of reproduces the obvious. And we can't really pull apart the mechanisms.

"Peer pressure, opportunities for crime, availability of drugs, availability of guns, and all kinds of other things are part of a neighborhood. And we don't know—criminologists don't know—how those mechanisms play out."

At the outset of his career, Berk wrote books and articles bearing titles like *The Roots of Urban Discontent* and "Local Political Leadership and Popular Discontent in the Ghetto." There are reasons why his late career work tends toward papers like "Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment."

"One piece is that the problems are much, much harder than I thought when I was 30," he says. "The second thing is that the quality of the social science being done has not been strong. So you've got a very hard problem, with the science not being very good. And it's partly not good because we don't have the tools and the data, and partly because the quality bar was set too low. And then the third thing that's frustrating is, even when we had good answers—and we have some, from time to time—getting anything done in a practical sense is frustrating. So you put all those together, and, yeah, I feel frustrated."

So if there's a touch of hubris in his sweeping claims about computerized fortune-telling, its coexists with a measure of surrender.

"What it comes down to is a much more modest set of aspirations," Berk says. "I'm just trying to help you anticipate the future better. I think I can do that with these new tools."

AS any science-fiction fan can attest, the whole point of foreseeing the future is changing it. Berk's algorithms, and others like it, portend more changes and challenges than many people realize. Most importantly, they will force a deep reckoning with our ideas about fairness.

Earlier this year, Berk co-authored a paper with Hoda Heidari, Shahin Jabbari, Micheal Kearns, and Aaron Roth—doctoral students and faculty members, respectively, in the School of Engineering and Applied Science's computer and information-science department—investigating tradeoffs between accuracy and fairness in machine-learning risk assessments. They demonstrated that not only does one typically come at the expense of the other, but that there are different kinds of fairness, and with rare exceptions it is mathematically impossible to satisfy all of them simultaneously.

One way to grasp the underlying problem is to consider college admissions.

When he was a faculty member at University of California-Los Angeles, Berk worked with admissions officers trying to figure out how to achieve gender equity. "They told me that if they used the standard quantitative measures—SAT scores, GPA, and so forth—UCLA would be 75 percent female, because females are better on those measures." But the administrators wanted the student body to be representative of the state's population—a mandate codified in the 1868 Act that created the University of California. "To do that, what we have to do is admit men who have lower SAT scores, lower GPAs. on average, than women," Berk says. "So we solve one kind of unfairness by introducing another kind of unfairness."

The same problem rears up in criminal justice. Men and women commit violent crimes at vastly different rates. Should an algorithm be programmed to ensure that the proportion of women predicted to succeed on parole matches the proportion of men predicted to succeed? Doing so would likely require incarcerating women who pose a far lower threat to public safety, even as more dangerous men are released. Perhaps the algorithm should be designed to ensure that the predicted ratio of false negatives to false positives is equal for both groups? But that may clash with both previous measures of fairness, and additional ones.

"If men are more likely to commit violent crimes than women," says Berk, adding: "*Yes*. If blacks are more likely to commit violent crimes than whites— *yes*—if there are different base rates, you're going to have to do something to get the outcomes and equality that you want."

His work on this issue reflects how potentially fraught every aspect of algorithm design can be.

"Let's say we use prior record as a predictor," he explains. "Everybody says: *Yeah, that's good*. But then they go on to say, *Well, because of the historical disadvantage that blacks have had, and problems with the police, that record doesn't really reflect how nasty they are—it reflects how aggressively they've been policed*. So prior record actually folds in past injustice. What do you do about that? There are algorithms we're working on which will take that into account."

One approach is to "clean up the data before you begin." Another is to "post-process" the results, tweaking outcomes to accord with some fairness goal. He is also working on building a "fairness constraint" directly into an algorithm. "The algorithm as it proceeds normally tries to maximize accuracy, but you can maximize accuracy subject to a certain level of fairness. You can say: We're going to do this in such a way that the false-positive rate for blacks is the same as the false-positive rate for whites."

None of these approaches solves the intractable tension between different sorts of fairness, though, or between fairness and accuracy. And the repercussions are already mounting. Last year, ProPublica analyzed the use of a machine-learning algorithm called COMPAS in Broward County, Florida. Data on more than 10,000 people arrested for crimes there showed that black defendants were twice as likely to be incorrectly labeled as high-risk than white defendants. Furthermore, white defendants labeled low-risk were "more likely to end up being charged with new offenses than blacks with comparably low COMPAS risk scores." Northpointe,

the for-profit company that makes the algorithm, countered that it had been designed to achieve a different measure of fairness—forecasting recidivism for black and white defendants with roughly the same overall accuracy—which in its view was superior. (A critical difference between COMPAS and Berk's algorithms is that Northpointe conceals its code under the veil of proprietary secrecy, whereas Berk's is open for examination—by their users, or judges, or defendants.)

"[T]here are no easy answers," Berk and his co-authors concluded in their article about fairness and accuracy. "In the end, it will fall to stakeholders—not criminologists, not statisticians and not computer scientists—to determine the tradeoffs … These are matters of values and law, and ultimately, the political process. They are not matters of science."

The proliferation of machine-learning crime forecasts will force stakeholders to navigate these tradeoffs with an uncomfortable level of specificity.

"It's funny," says Coglianese, the Edward B. Shils Professor of Law and Professor of Political Science. "We tolerate a lot more ambiguity when we delegate to humans than we would when we delegate to robots.

"On tough questions, like fairness, what we do as a society is proceed in a manner along the lines of what Cass Sunstein called 'incompletely theorized agreements,'" he adds. "By that he simply means that we don't have a definitive answer to a question like, *What is the appropriate test of fairness?* And he even argues that in judicial decision-making, judges shouldn't try to answer these questions in a once-for-all manner that tries to lay out a commitment to one particular principle that will answer all cases.

"The thing that has become clear is that the computer scientists and statisticians who are designing these algorithms, they need some mathematical precision. They need a resolution: tell me what the ratio is."

And to a certain degree, they need to know what *every* ratio is. Remember Berk's original parole algorithm, which

"We're not in the world of *Minority Report* yet, but I think there's no question we're heading there. The only question is how fast, and what sort of oversight we're going to provide."

predicted that 1,764 people would be charged with homicide and was correct about 137—a little over 7 percent—of them? That was the result of a determination by APPD administrators that the costs of false negatives (failure to identify a future killer) were 10 times greater than the costs of false positives (erroneously classifying a low-risk individual as a high-risk one).

Where did that cost ratio come from? A gut-level ethical instinct, probably. But when the APPD actually implemented the algorithm, it quickly became apparent that a 10-to-1 cost ratio "placed far more offenders into the high-risk category—the vast majority of whom were actually moderate- or low-risk—than the department could ever hope to manage," according to a 2012 review by Geoffrey Barnes, an assistant research professor of criminology, and Jordan Hyatt, a senior research coordinator at Penn's Jerry Lee Center of Criminology. So a new cost ratio emerged—one conditioned more closely on the APPD's budget: 2.5-to-1.

Choosing one ratio instead of another will have wide-ranging repercussions on life and liberty—and may even feed back into determinations about how richly or poorly to fund the public-safety agencies charged with protecting both.

And of course any direct attempt to redress a systemic injustice rooted in unequal treatment of different demographic groups (or synchronize the algorithmic odds for men and women) requires that immutable characteristics like race and gender be plugged into the model, after all.

Whether such information is directly inputted or merely permitted to "sneak in," the black-box nature of the algorithm itself has some curious implications. To violate the Fifth Amendment, which guarantees individuals equal protection under the law, "the government has to actually have discriminatory *intent*, and not just discriminatory outcomes" when it acts, says Coglianese. "So to the extent that the algorithm is learning and picking up on variables that respond to race, but doing so without anybody telling it to pick up on those variables, and without anybody having any *intention* that it would pick up on those variables … you almost inherently get some immunity from any claims in that regard."

Yet jurisdiction matters, he adds. The Fifth Amendment applies to federal government actions. Under 14th Amendment jurisprudence, which governs states, discriminatory outcomes can form the basis of a legal challenge. The same goes for statutory civil-rights cases brought under Title VI of the 1964 Civil Rights Act.

Andrew Ferguson L'00 is interested in what law-enforcement algorithms will mean for the Fourth Amendment, whose protections against unreasonable searches and seizures have long been the main restraints on American policing tactics. Ferguson, a professor at the University

of the District of Columbia's David A. Clark School of Law, is the author of *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, which will be published by New York University Press in October.

In it, he examines the use of forecasting algorithms by police departments in several big American cities. Chicago, for instance, has used them to develop a "Heat List" of over a thousand young men ranked in order of their risk for inflicting or suffering from violence. It has been "tragically accurate," he writes. "On a violent Mother's Day weekend in 2016, 80 percent of the 51 people shot over two days had been correctly identified" by the algorithm. Los Angeles, by contrast, uses facial-recognition software, automated license-plate readers, police field reports, and other data to predict *where* crime is likeliest to occur, and deploys officers accordingly.

Both approaches hold promise and peril.

"In Chicago there's a knock on the door by a detective who is saying to the individuals, *Look, we think you are at risk. We have an opportunity to change your life around*," Ferguson explains. "And while that sort of public-health approach to policing has a lot to commend it—if it was done with resources and actual jobs and social services behind it—it is also a pretty stark vision of social control." And when the social-services element is missing or lacking, it can seem cynical. After all, what good is an algorithm that identifies 80 percent of shooting victims if they still get shot anyway?

"They didn't fund that part of it," Ferguson says. "So it's no wonder that the system didn't work that well." But New Orleans, he adds, "actually funded the social-services programs" along with the Big Data approach, "and violent crime dropped."

That's what Berk hopes to be able to do, on a grander scale, with his Norway at-birth predictions. "If I know my kid is at risk to commit a homicide as a teen, we know things that work," he says. "There's parent training, there's proper nutrition. I'll start at the moment of conception: you try to get the mother proper nutrition, get her off alcohol and drugs, provide visiting nurses to help her, and then help the baby. There's preschool. There's all kinds of positive things you can do. We should do it with all kids, but we can't afford to do it with all kids—or for political reasons we don't want to. So we could find the high-risk kids and saturate them with things that we believe will make a difference."

Place-centric approaches like the one used in Los Angeles raise different issues. "Communities of color bear the brunt of these tactics," Ferguson writes. "By and large, it is poor people from minority backgrounds who are stopped by police. By and large, it is people of color who are populating the growing police databases. If these racially skewed databases of past police contacts become the justification for future police contacts, then biased data collection will distort police suspicion."

On the other hand, "if you talk to officers and administrators who are doing it right," he says, "they say one of the benefits of predictive policing is reshaping officers' perspectives to not necessarily be there to arrest more people, or stop more people, or frisk more people, but to make your presence known so you deter crime … The LAPD claims that they've been doing that in certain ways. And if they are, that's a great benefit of predictive technologies."

But both place- and person-based approaches to predictive policing pose a fundamental challenge to Fourth Amendment protections, Ferguson contends.

"Policing came of age in an era of small data, limited to what police officers could see, what they knew. In big cities, they were playing in a world of imperfect and sometimes very poor information. And so our legal doctrines were created to give a little bit of leniency to police making snap decisions on the street."

That changes in a world of Big Data, he continues, when police have vast information about neighborhoods and social networks, and are served up predictive risk scores for an individual they encounter. "That information will distort how a police officer will deal with that person," Ferguson says. "Maybe in an intelligent way, maybe in a problematic way, but it *will* distort."

In effect, Big Data makes it far easier for an officer to justify a claim of "reasonable suspicion" for detaining someone—even if all they've done is unwittingly enter an area flagged by software as having an elevated risk of crime at a particular time of day. In *Illinois v. Wardlow* (2000), the Supreme Court ruled that a "high crime area" can be considered a factor in determining reasonable suspicion. The Court didn't define that term, but an algorithm will—for reasons that may or may not be knowable.

"With more information about individuals, the rather weak limitations of reasonable suspicion will fall away," Ferguson writes. "Innocent factors cobbled together (friendships, neighborhood, clothing, social media connections) will be crafted into suspicion as necessary. The result of a 'small data doctrine' confronting a big data world will be less constitutional protection."

He concludes: "If walking through a predicted red box changes my constitutional rights to be free from unreasonable searches and seizures, then a higher level of scrutiny might need to be brought to bear on the use of the technology."

There's also a risk that using person- and place-centric prediction in parallel could create feedback loops that undermine the accuracy of both. Imagine that someone is stopped largely for being in a predicted high-crime area, and that stop itself becomes data, both for future person-based targeting and place-centric forecasting. "There could be some double-counting," Berk says. Depending on the circumstances, it could exaggerate some risks while draining adequate attention from others.

In a more hopeful light, the coming tidal wave of data could make for better and

more sensitive policing—which is the overarching goal of the Black Lives Matter movement. Rudovsky has seen the glimmerings of this in the wake of legal challenges to the use of discriminatory stop-and-frisk tactics. Following a class-action suit against the Philadelphia Police Department, his firm gained the ability to monitor the department's implementation of reforms. He says interrogating police data has led to marked improvements.

"They were doing a lot of stops for *loitering*, whatever that means," Rudovsky says, as an example of a hazy pretext for stops that disproportionately targeted African-American residents—who actually possessed weapons or illegal drugs less frequently than whites who were stopped. "The city started retraining the police—you can't stop a loiterer" if that's the only pretext.

And data from police reports contained other revelations. "For many years, the Supreme Court has talked about common sense" as a reasonable justification for police stops," says Rudovsky, giving an example: "*Oh, if you see a bulge in his pocket, it might be a gun.* Well, it turns out, in Philadelphia, 99.2 percent of all the bulges are cellphones, not guns. They don't get a lot of guns from a bulge. So it gives us some insight. When you have a whole comprehensive data set, and I can look at 200,000 stops in a year—which is what Philadelphia has been doing—you can make a better analysis of what works and what doesn't work." Which, in turn, could influence what courts will sanction under the rubric of common sense.

This year, he says, there's been a 30 percent reduction of the number of stops by Philadelphia police, and an increase in the percentage of them that actually reveal illegal behavior. "There's still a way to go," Rudvosky allows. "But it's data, and internal accountability, [that] has allowed us to do it."

New York City, he adds, was stopping 750,000 people five years ago, and is now down to 15,000. "Now, even if you triple it—even if cops aren't reporting all of them—it's still a 90 percent drop," Rudovsky says. "And the crime rate is down. Arrests are down, citations are down, everything's down in New York.

"One of the problems with stop-and-frisk is that it alienates the community. So that you don't get cooperation, and there's *more* crime," he says. "I think there's really a phenomenon where people won't report crime, they won't cooperate with the police—*These guys hassle me every day, I know who that shooter was, I'm the last one that's going to go to the police and tell them.*"

"I'm not against good policing," he concludes. "What I'm against is policing that doesn't work." And the rise of predictive policing has the potential to root out ineffective or counterproductive practices, and illuminate better ones.

Ferguson concurs that a flood of data could help sweep away entrenched prejudice and bad practices. "You may objectively be able to see what is happening. It may be the case that judges are holding people for risk assessments based on, you know, their own experience, their own proxies, their own bias. And the data might actually be revealing something that would allow us to critique it in an open way."

The danger, he cautions, is that algorithmic decision tools might also make people comfortable and sloppy. "You can look at the score and not think behind it. You can think, *Well, I will defer to whatever this risk score is, even though I don't understand it. I'm not really sure where it came from, but it's the best I have.* And it can cause people to sort of stop thinking."

For all Berk's optimism about automated crime prediction, he is adamant that humans need to retain ultimate control—and think harder, in many ways, than we have in the past.

He opposes giving mandatory weight to an algorithm's determinations about parole, probation, or sentencing, for instance. "The only information that's available to the judge from the algorithm is what the algorithm had access to," he explains. "The algorithm didn't see the trial. The algorithm didn't hear any statements from victims, or whatever. So there's additional information that's not in the algorithm because it's not available. You lose that if you build that forecast into [mandatory] guidelines."

He also argues that criminal-justice algorithms should be open-source. This is a pressing concern. The Supreme Court of Wisconsin recently ruled that the use of Northpointe's COMPAS algorithm did not violate a defendant's due process rights—even though neither the defendant nor the judge was able to evaluate the tool's methodology.

"I think it's got to be open," Rudovsky concurs. "If courts are using it, if judges are using it—and saying you're going to get bail or not get bail, or be released today or in two years—it's got to be transparent."

"I'm prepared to sit down with IT people, and sometimes I do," Berk says of his code. "I'll sit down and take them line by line, if that's what they want to do." He also stresses that every actor in the criminal-justice system should have its own custom-designed algorithm, and that it should draw from data that's relevant to the jurisdiction. "Even within Philadelphia, I'll have one algorithm for the police, another algorithm for probation. Because the decision that's being made by each criminal-justice actor is different. And they have to consider different factors. And the consequences of their mistakes are different. So you need a different algorithm for each application."

Most fundamentally, he maintains that, as indispensable as criminal-justice algorithms may be, they are no substitute for democratic decision-making about what constitutes true justice.

"Sometimes articles come out that basically say, *Well, these smart guys will solve the problem.* Unh-unh," he says. "These are very hard problems. And the message has to be: We have to get together and decide about these tradeoffs. And they're not going to be easy."