ANNUAL REVIEWS

Annual Review of Criminology

Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement

Richard A. Berk

Departments of Statistics and Criminology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; email: berkr@sas.upenn.edu

Annu. Rev. Criminol. 2021. 4:209–37

First published as a Review in Advance on November 13, 2020

The *Annual Review of Criminology* is online at criminol.annualreviews.org

https://doi.org/10.1146/annurev-criminol-051520-012342

Keywords

artificial intelligence, algorithms, machine learning, predictive policing, risk assessment, criminal justice, fairness, transparency

Abstract

There are widespread concerns about the use of artificial intelligence in law enforcement. Predictive policing and risk assessment are salient examples. Worries include the accuracy of forecasts that guide both activities, the prospect of bias, and an apparent lack of operational transparency. Nearly breathless media coverage of artificial intelligence helps shape the narrative. In this review, we address these issues by first unpacking depictions of artificial intelligence. Its use in predictive policing to forecast crimes in time and space is largely an exercise in spatial statistics that in principle can make policing more effective and more surgical. Its use in criminal justice risk assessment to forecast who will commit crimes is largely an exercise in adaptive, nonparametric regression. It can in principle allow law enforcement agencies to better provide for public safety with the least restrictive means necessary, which can mean far less use of incarceration. None of this is mysterious. Nevertheless, concerns about accuracy, fairness, and transparency are real, and there are tradeoffs between them for which there can be no technical fix. You can't have it all. Solutions will be found through political and legislative processes achieving an acceptable balance between competing priorities.

INTRODUCTION

The use of artificial intelligence in the public and private sectors has of late generated considerable controversy. There can be concerns about technical matters (Domingos 2015), jurisprudence (Coglianese & Lehr 2017, Kroll et al. 2017), ethics (Berk et al. 2018, Dwork & Roth 2014), and even survival of the human species (Cellen-Jones 2014, Scharre 2018). Some treatments are cross-cutting integrations (Kearns & Roth 2019, Mitchell 2019a). Although the issues are very real and often judiciously addressed, many of the debates can be compromised by misunderstandings. Key concepts too often are undefined or defined differently, assertions commonly are made that are baseless or misrepresentations, and focus can be lost to a variety of other agendas that may be political, economic, or narrowly self-serving. Mass-media treatments that otherwise are very thoughtful (Mullainathan 2019, Pande 2018, Smith 2019) can at times get carried away (Chinoy 2019, Starr 2014).

Many of the same issues arise in criminal justice settings where the discussions can be no less contentious or poorly informed. Perhaps the best example is the controversy surrounding the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk instrument protocol. COMPAS is a widely used risk assessment tool deployed to inform criminal justice decisions such as sentencing. Angwin and her colleagues published an analysis claiming that the instrument was racially biased (Angwin et al. 2016). Their assertions were widely disseminated and, in many circles, accepted. However, the critique was far too narrowly formulated and poorly informed about the nature of racial bias and necessary tradeoffs between different kinds of unfairness (Berk et al. 2018, Spielkamp 2017). Furthermore, COMPAS became the poster child for all statistical forms of risk assessment. At the very least, it was guilt by association. A second equally controversial example is software used for facial recognition. *The Guardian* found in an internet search more than 130 articles on facial recognition over the past several years (**https://www.theguardian.com/technology/facial-recognition**). Among the many issues raised were privacy, politically motivated surveillance, and differential accuracy depending on race and gender.

Within this larger context, the pages below address the use of data by law enforcement to make predictions of various kinds. In some forms, the enterprise is called predictive policing. In other forms, it is called risk assessment. In either case, the predictions are often said to be the product of artificial intelligence, and controversy necessarily follows. There are, again, concerns about technical matters, jurisprudence, and ethics, with a special emphasis on criminal justice fairness. There will be no resolution to these issues in this review because the main disagreements are largely unresolved. Rather, the goal is to clarify the discussion and convey empirical findings when they can be found.

THE MUDDLE OF ARTIFICIAL INTELLIGENCE

Before examining law enforcement applications, confusion about the nature of artificial intelligence must be addressed. In part because there seems to be nothing more human than thinking, virtually any collection of computer code becomes fair game if it produces results like those that could arise from human mental processes. In an early formulation, Alan Turing (1950) argued that if a machine could exhibit behavior that was indistinguishable from that of a human, the machine would be executing intelligent behavior. A recent application of this perspective had a computer make a telephone reservation at a restaurant through such a human-like conversation that the person taking the reservation did not know he was speaking to a computer. "It had mastered the cadence of a human conversation, replete with *uh*'*s* and *hhhm*'*s* that we sprinkle into conversations while we're thinking" (Krohn 2019, pp. 35–36). However, the Turing Test is as controversial as it

is fascinating and has been met with a number of incisive critiques (Grahm & Dowe 2019). Among the simplest is to question whether imitation by itself is enough. Should one not care about how the imitation is assembled?

A rather different definition of artificial intelligence boils down to any device that is able to sense its environment and then take actions increasing the likelihood that its objectives are achieved (Poole et al. 1998). However, this formulation represents a very low bar because even the common household thermostat qualifies. An even lower bar only requires that an intelligent device compete favorably with humans on the same cognitive tasks. For example, if my goal is to travel by car to a supermarket to which I have never been, the GPS-based application *Waze* qualifies if its route and mine get me to my destination in effectively the same elapsed time. It would not matter if we took the same path or different paths. A very recent and spectacular illustration is the victory of Google's *AlphaGoZero* over the best Go player in the world (Krohn 2019). In both these examples, there is no requirement that the machine processes and the human processes be the same. All that matters is comparable output.

A related conception works backward from particular tasks whose execution exemplifies artificial intelligence (Goodfellow et al. 2016, p. 1).

> In the early days of artificial intelligence, the field rapidly tackled and solved problems that were intellectually difficult for human beings but relatively straightforward for computers—problems that can be described by a list of formal, mathematical rules. The true challenge to artificial intelligence proved to be solving tasks that are easy for people to perform but hard for people to describe formally—problems that we solve intuitively, that feel automatic, like recognizing spoken words or faces in images.

Artificial intelligence becomes nothing more than a means to tackle certain kinds of challenging empirical exercises that are easy for humans and represent activities that seem to be distinctly human such as speech.

In short, there is no commonly accepted definition of artificial intelligence and, so far at least, all the candidates have substantial flaws. Part of the confusion is that the activities to which artificial intelligence is applied are very heterogeneous and have become more diverse over time. In this context, there is merit in Jon Krohn's three kinds of artificial intelligence (Krohn 2019), which distinguish between real-world applications and applications that currently are highly speculative.

- Artificial narrow intelligence performs specific tasks such as translating languages, recognizing faces, providing medical diagnoses, guiding self-driving cars, and playing games such as chess or Go. This form of artificial intelligence currently is available.
- Artificial general intelligence implements a single set of code able to perform effectively across a range of tasks formerly undertaken separately under artificial narrow intelligence. There is at present no such code, and experts predict that it is several decades away, if it can be achieved at all. For example, the same code that translated spoken words into written text would also run a self-driving car, search video for specified images (e.g., CCTV video of a man in a blue sweater and tan sneakers who witnessed an industrial accident), and win at games of Settlers of Catan (**https://www.catan.com/game/catan**).
- Artificial superintelligence, which is little more than a placeholder for machine intelligence, is, at best, 50 years in the future and would surpass human intelligence. Current thinking has centered on the many obstacles that may be insurmountable, an important one being the enormous computing capability required that even quantum computing may not be able to provide (Arute et al. 2019, Metz 2019). Equally challenging are inherent connections between different kinds of thinking, feeling, and reasoning that all contribute to human intelligence (Mitchell 2019a). "Human intelligence is a strongly integrated system,

one whose many attributes—including emotions, desires, and a strong sense of selfhood and autonomy—can't easily be separated" (Mitchell 2019b, p. A27).

There is more to intelligence than information processing. Furthermore, as machines acquire increasing cognitive skill and neuroscience continues to make progress, conceptions of intelligence will evolve.

To summarize, the artificial intelligence of today is computer code written to accomplish a specific empirical task. Restated, it is just a computational procedure. In law enforcement settings, these procedures are by and large enhancements of activities that humans are already performing. Arguably, computers can do them faster and more accurately, but the internal operations are far removed from the way humans actually think. It is not even clear that the term intelligence applies.

MODELS, ALGORITHMS, AND MACHINE LEARNING

An essential feature of artificial narrow intelligence is that it depends on algorithms, not models. Algorithms are methods to compute things. When you balance your checkbook, you are using an algorithm. When you compute a regression coefficient or undertake a singular value decomposition, you are using an algorithm. In contrast, a model is an explicit theory about how the world works. Force = Mass × Acceleration is a model. The goal is explanation, perhaps involving cause and effect. Algorithms should not be confused with models because a range of errors can easily follow (Breiman 2001b). For example, model misspecification is a pervasive problem in criminology and the social sciences more generally (Freedman 2009), but model misspecification is not even defined for algorithms (Berk 2020b); it makes no sense to say an algorithm is misspecified.

Consider the pair of equations for the ubiquitous linear model with a numeric response variable.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{p_i} + \varepsilon_i, \quad 1.$$

where

$$\varepsilon_i \sim \text{NIID}(0, \sigma^2). \quad 2.$$

Equations 1 and 2 are a theory of how each case i came to be. Each response value is a linear combination of regressor variables weighted by their regression coefficients, a constant serving as the y-intercept, and a perturbation drawn independently and at random from some probability distribution, sometimes assumed to be normal.

Equations 1 and 2 are meant to capture exactly how the response values are generated; nature literally adopts the conventional linear model (Berk et al. 2014). For Equation 1, this means that the first-order conditions must be met; the mean function is correct. It then follows that one can obtain unbiased estimates of coefficients in Equation 1. To obtain the unbiased standard errors required for valid statistical tests and confidence intervals, Equation 2 must be correct as well. These are the second-order conditions. All uncertainty is produced precisely as Equation 2 states. Furthermore, if the first-order conditions are not met, the second-order conditions, as a formal matter, cannot be met. Even when the most apparent inferential consequences of a misspecified mean function are asymptotically addressed with robust standard errors such as the sandwich, violations of the first-order conditions invalidate all statistical tests and confidence. Bias in the regression coefficients systematically makes the coefficient estimates too large or too small, which offsets the test statistics and systematically makes the p-values too large or too small.

Although in practice, Equations 1 and 2 can be used for forecasting, interest typically centers on estimates of the regression coefficients. The primary goal is explanation, i.e., why things work the

way they do. These issues are revisited below when predictive policing and criminal justice risk assessment are discussed. For more background, didactic treatments are readily available (Berk et al. 2019, Freedman 2009).

In conventional matrix notation, a common mathematical expression for the calculations by which coefficient estimates from Equations 1 and 2 are obtained is

$$\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \quad 3.$$

Equation 3 is a computational procedure that can serve as an algorithm. It makes absolutely no claims about how the world works and has no aspirations to explain anything. Because there are no first- and second-order conditions to meet, concepts like model misspecification are inapplicable. Its sole purpose is to provide values for the regression coefficients and from them to derive fitted values. As addressed shortly, unbiased estimates are no longer the Holy Grail. They ideally are balanced against the estimation variance through the bias–variance tradeoff (Berk 2020b).

Artificial narrow intelligence is consistent with this perspective, although the algorithms can be far less transparent. The usual aim is to produce fitted values that can be used to inform decisions. In facial recognition, for instance, a properly trained algorithm can be tasked with identifying particular individuals from CCTV images. Explanation is irrelevant, and no claims are made beyond whether a match is found.

Because artificial narrow intelligence does not depend on models, a wide variety of algorithms can be used. These in general fit the data better than models (Berk & Bleich 2013). Many are classified as machine learning (ML). Here too, there are no commonly accepted definitions, but discussions usually center on data-fitting procedures that are adaptive (Berk 2020b). For example, the size of the window when a moving average is computed is determined on the fly by some measure of fit (e.g., mean squared error). Window size is not determined before the data are analyzed. The lasso (Tibshirani 1996), when used to determine the included predictors in a regression analysis, is in the same spirit because the model specification is not known before the analysis begins. Adaptive fitting causes fundamental problems for statistical inference, although in some situations, good solutions exist (Berk 2020b).

Immediately below is an illustration of an algorithmic procedure that has many similarities to an ML approach called boosting (Berk 2020b). It emphasizes the adaptive nature of the fitting process. These are the kinds of tools that are central when artificial narrow intelligence is deployed for criminal justice applications. Rarely is there anything remotely like a model.

1. Specify a mean function as in Equation 1 and apply least squares as usual.
2. Compute the fitted values.
3. Compute the residuals.
4. Compute a measure of fit.
5. Apply Equation 1 again, but weight the data so that observations with larger absolute values of the residuals are given more weight.
6. Update the fitted values obtained from the immediately preceding regression with the new fitted values weighted by their measure of fit.
7. Repeat steps 2–6 until the quality of the fit does not improve (e.g., 1,000 times).
8. Output the fitted values.

The output of interest is the linear combination of weighted fitted values computed over many passes through the data. The regression coefficients have no interpretative value because there can be thousands of them calculated from differently weighted data sets. Equation 1 is not being used as a model but, via Equation 3, is a component of an algorithm.

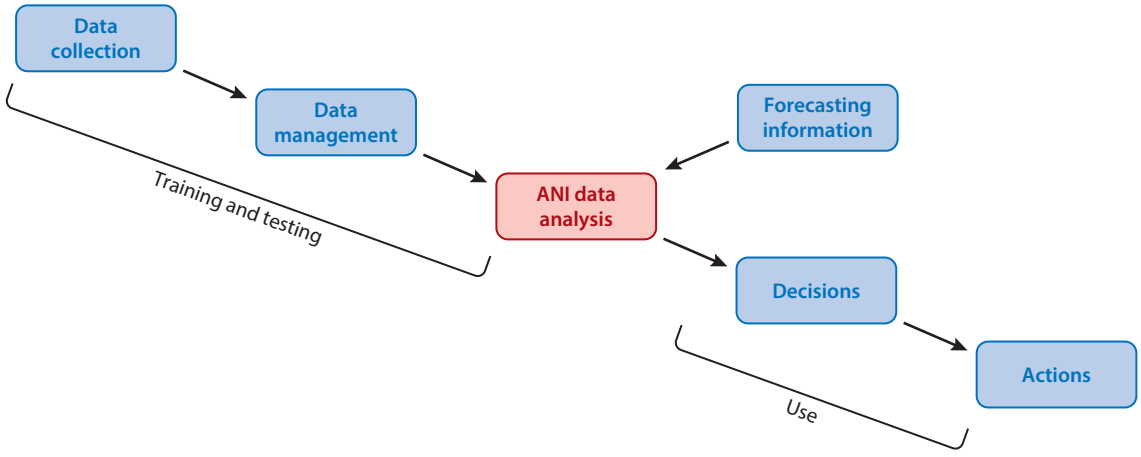

Figure 1

Artificial narrow intelligence (ANI) data analysis embedded in other activities and used to inform several criminal justice functions: data collection, data management, decisions, and actions.

ARTIFICIAL NARROW INTELLIGENCE IN THE FIELD

Another important source of confusion is a failure to separate artificial narrow intelligence from the other activities that surround it. **Figure 1** underscores that there are essential steps before the data analysis begins and after the data analysis ends. An ML algorithm is often blamed for problems introduced by the data and data management or for decisions and actions taken after the data analysis ends.

Typically, algorithms used in criminal justice settings must be trained. This requires training data that include predictor variables, which are essentially regressors in the linear regression setting, and one or more response variables to be fitted. Usually, there is a single response variable that may be numeric or categorical. From the training, the algorithm learns how the predictor variables are related to the response variable(s). Once these associations are established, they can be used to construct imputed or forecasted values of the response variable(s) when those values are unknown. There should also be test data, generated in the same manner as the training data but not used in the training. Because of the adaptive nature of learning algorithms, there are complications for statistical inference that can sometimes be properly addressed with test data. The test data are uncontaminated by adaptive procedures or "data snooping" in the training process.

From **Figure 1**, an initial step is collecting and organizing relevant training and test data, and at this point, a variety of difficulties may be introduced. For example, a prior record for misdemeanors can contain arrests motivated by racial animus. Perhaps more fundamentally, the arresting charge may not comport well with the actual behavior triggering the arrest. Regardless of an offender's race, police might in some cases overcharge to give prosecutors more leverage for plea bargaining. There can also be recording or data processing errors so that, for instance, some birthdates are incorrect. This can happen if, rather than obtaining the date of birth from a driver's license, the offender is asked for his or her date of birth. All errors, whatever their cause, are carried forward in the training data and can affect results from the algorithm. There is growing evidence that this is the major source of inaccuracies and unfairness in algorithms used by criminal justice decision-makers (Berk 2020a). The algorithm is faultless and, indeed, there is ongoing work to design algorithms that reduce the impact of problematic data (Berk et al. 2018).

Once the trained algorithm reaches satisfactory performance in test data, it is ready for use. Information used for imputation or prediction is presented to the trained algorithm. The output is intended to inform criminal justice decisions that, in turn, affect actions by criminal justice officials. Just as with the data, a variety of problems can be introduced. The algorithmic results for particular cases might be inaccurate. Inappropriate actions can follow. Sometimes the algorithmic results are ignored or misused. Again, inappropriate actions can follow. And the decisions and actions may be inappropriate for reasons unrelated to an algorithm's output. But, until there are concrete actions, there can be no real harm and no real benefits. There is, therefore, an opportunity to right obvious wrongs before actions are taken.

The key point is this: The algorithmic tools are embedded in several concomitant activities that are typically far more consequential for the actions taken than the algorithms themselves. Remedies for problematic actions should be introduced where they can be most effective. Currently at least, this is unlikely to be for the algorithm.

PREDICTION

In this review, prediction and risk are the main focus. The other uses of algorithms need to be briefly mentioned because some are inaccurately folded into predictive policing or risk assessment. Algorithms are commonly used by law enforcement organizations for the following tasks:

- administrative functions like payroll or equipment tracking;
- personnel matters such as performance evaluations of sworn officers;
- forensics such as crime lab case management, DNA matching, and metabolic profiling (Dinis-Oliveira 2017);
- license plate recognition;
- organizing and analyzing evidence about specific crimes that have occurred;
- displaying crime statistics and time–space crime patterns on department websites;
- CCTV surveillance;
- organization, redaction, and retrieval of information, such as facial recognition, from CCTV videos or body cameras;
- crime mapping and spatial analysis;
- the use of algorithms in smartphones and computer operating systems;
- scheduling and deployment of human and material law enforcement assets;
- voice to written report translators; and
- performance evaluations for geographical areas (e.g., precincts) as in COMPSTAT.

The algorithms employed do not have to be fancy or sophisticated. A simple two-way table from an Excel spreadsheet qualifies. Also, there is no forecasting. The algorithms look back in time to what has already happened or provide prescriptive guidance for future activities.

At least for administration and scheduling, there is similar software for prosecutors' offices, public defenders' offices, administrative offices of the courts, and jail, prison, probation, and parole administrations. Here too, none of the algorithms are engaged in forecasting, and the calculations can be very simple.

Predictive Policing

Despite some wide-ranging expositions (Perry 2013), the foundations of predictive policing are forecasts of when and where future crimes will occur. Based in part on the "criminology of place"

(Weisburd et al. 2012), forecasts are made in time and space about crimes that will be committed. The spatial units are locations that can be quite large, such as an entire precinct or census tract. Locations also can be small, e.g., a particular street corner. Likewise, the temporal units can vary dramatically in size: an administrative reporting period such as a month, a week, a day, or even a particular time of day. One might forecast, for instance, a dramatic increase in assaults and robberies on a particular block on a Saturday night just after the bars close. The spatial units also can be types of locations, and the importance of the temporal units may be determined by social circumstances. For example, one might forecast an increase in residential burglaries along major bus routes on weekdays just after the school day ends. Or one might forecast an increase in jewelry stores break-ins over long holiday weekends.

From such forecasts, police assets can be allocated to prevent crime or apprehend criminals. An example of the former is patrolling at certain times and places to "show the flag." An example of the latter is staking out a particular jewelry store on Christmas Day. Sometimes both kinds of tactics are implemented at once. For instance, parking garages at airports late at night may benefit from a police presence to deter automobile thefts and/or catch would-be car thieves. In both cases, policing can be more surgical.

There is nothing in such police practices that requires powerful data analyses. Indeed, police in urban areas have been using their experience, tips from informants, complaints from the public, police craft lore, and very simple mapping procedures to do predictive policing for well over a century. Research on crime hot spots retrospectively helps justify this approach, although the demonstrable reductions in crime are typically modest (Braga et al. 2014).

The advanced use of data and data analysis by police departments is a relatively recent phenomenon. COMPSTAT procedures, employed primarily as an internal management tool to increase police accountability, are a good, early benchmark (Weisburd et al. 2004). Implicit forecasts became common. For example, a precinct with a large number of reported crimes typically was assumed to be a high-crime precinct in the immediate future. Performance was evaluated by whether, after various reforms and interventions, crime dropped. It was then a short step to begin formal forecasting coupled with efforts to improve the quality of local data.

As an empirical matter, virtually all modern predictive policing procedures begin with recorded crimes in time and space. These are often displayed on maps, much in the spirit of pin maps used by police for generations. For a given time period, the map shows where specific kinds of crimes have occurred (e.g., armed robberies in the third precinct). If the crimes are spatially concentrated, counts of the number of crimes are used. Such maps and the underlying data are available over time. This is the raw material for the data analytics of predictive policing.

Commonly, there is temporal and spatial autocorrelation that can be exploited for prediction. Crimes that are more proximate in space and/or time tend to be more similar. Crimes lagged in time and offset in space can be used as predictors. That is, a crime count or crime event at location i and time t becomes a function of lagged crime counts or crime events in time and offset crime counts in space. The time units might be months, weeks, or even days. The spatial units might be distance in miles or more simply, areas adjacent to location i. The metric of Euclidian distance has much the same statistical properties as the metric of time and often can be analyzed in much the same manner.

Expressed very generally,

$$Y_{it} = f(Y_{(\neq i,\neq t)}, \varepsilon_{(it)}), \tag{4.}$$

where ε is an IID (independent and identically distributed) perturbation that can also be lagged in time and offset in space. Y_{it} can represent crime in several different ways: as a count of the number of crimes, the presence or absence of a crime, crime probability/density, or a parameter

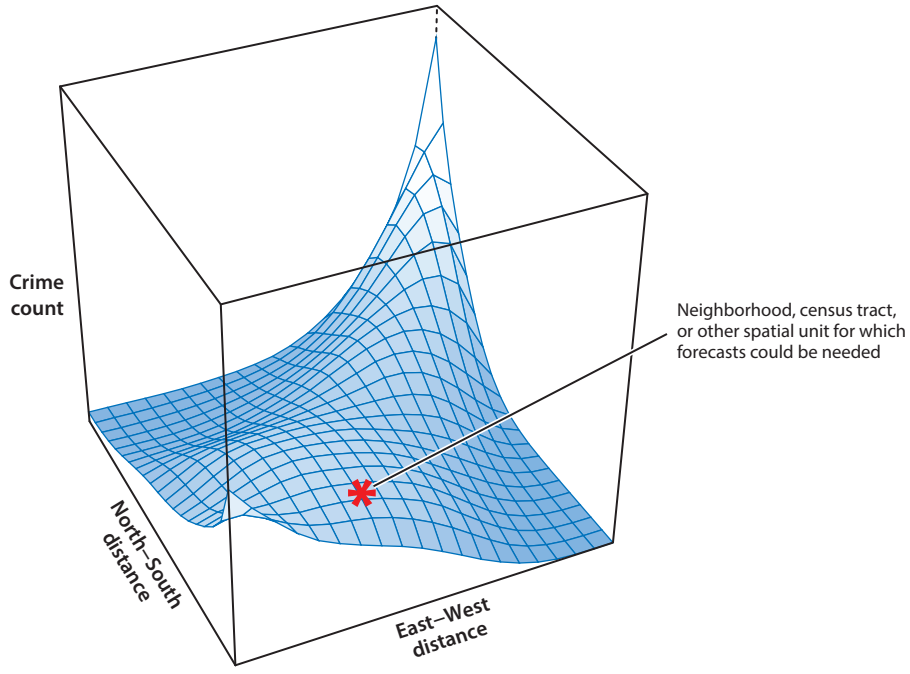

Figure 2

Distance density smoother resulting in a smoothed surface of crime counts.

of a statistical model (e.g., the rate parameter λ in Poisson models for counts). For didactic simplicity, crime counts of binary crime events are used in the next several pages. The challenge is to formulate the details of Equation 4 and then properly implement it with data. One must exploit the associations in time and space.

Figure 2 is an illustration of a smoothed surface of crime fitted values that can be produced by a detailed elaboration of Equation 4. Forecasting follows easily for any location, such as shown by the red asterisk. One has already obtained estimates of the parameters for the procedure responsible for **Figure 2**. Then, at time t, the crime counts at that location, and perhaps earlier times, along with crime counts at other locations, become predictor values. They are inserted in the procedure, and a new fitted value for time $t + 1$ is obtained; this is just a new fitted value. The fitted value at time $t + 1$ is the forecast. Sometimes criminogenic predictors can be used as well. For example, if one is trying to forecast for time $t + 1$ homicides and attempted homicides for location i, counts of the number of aggravated assaults and the local high school dropout rate for location i at time t might be included for all i when the fitting procedure is fashioned. Ideally, one can also compute a proper prediction interval at $t + 1$ to capture uncertainty in the forecast. There are some new techniques that have enormous promise for properly estimating forecasting uncertainty (Lei et al. 2018). Other recent work shows that both time and space can be smoothed at once using kernel-density estimation over those two dimensions (Hu et al. 2018). Forecasting accuracy may be improved.

A variety of very different methods have been used to flesh out Equation 4. An early treatment by Groff & La Vigne (2004) provides a good review that requires updating to highlight the differences between models and algorithms. **Figure 3** offers a visual summary of common forecasting methods organized into three broad categories. The first is model-based methods in which forecasts depend on a subject-matter theory of crime generation or a statistical, stochastic model. The

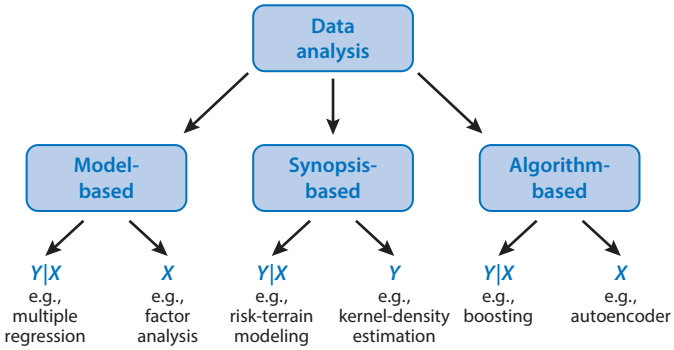

Figure 3

Methods of data analysis used by law enforcement and researchers for predictive policing.

second is synopsis-based methods in which forecasts are based on empirical summaries of crime in time and space, commonly displayed on various kinds of maps. The organizing goal is explanation, from which forecasts sometimes follow. The third is algorithm-based methods of the sort that can be treated as artificial narrow intelligence.

Sometimes particular predictive policing tools can be placed in more than one of the three boxes, and the literature has a penchant for calling all methods models. For example, in an article by D'Orsogna & Perc (2015) ambitiously titled "Statistical Physics of Crime: A Review," every procedure reviewed is called a model perhaps because, for each, equations are prominent. Such complications are addressed below when they matter.

Substantive and statistical models. The modeling literature distinguishes between two kinds of models. The more familiar kind of model is a mathematical statement of processes and relationships for some empirical phenomenon in at least the spirit of Equations 1 and 2. Classical econometrics, in which one or more algebraic equations are meant to formally represent economic theory, is the ideal case. "The science of model building consists of a set of tools, most of them quantitative, which are used to construct and test mathematical representations of portions of the real world" (Pindyck & Rubinfeld 1981, p. xiii). Causality typically is central (Angrist & Pischke 2009). Such models can then be used for forecasting, applying what are essentially extrapolations (Pindyck & Rubinfeld 1981).

The econometric perspective has been common in criminology for several decades (Tarling & Perry 1981). Equation 4 with added details becomes an algebraic theory of some crime generation processes. Models for pooled cross-section time-series data might then be appropriate (Cameron & Trivedi 2005), and there is recent work in econometrics focusing on crime prediction (Liesenfeld et al. 2017). The credibility of all such work depends on fulfilling the first- and second-order conditions discussed above.

A model does not have to literally represent how the world works. Rather than using a substantive model, a statistical model can be employed. The statistical model must only capture the stochastic process responsible for generating the data. There needs to be no accompanying subject-matter account. Consider a very simple integrated autoregressive moving average (ARIMA) model in time (only) with autoregressive and moving average terms but no differencing:

$$Y_t = \pi Y_{t-1} + \varepsilon_t + \phi\varepsilon_{t-1}, \qquad 5.$$

where Y_t is some response variable one seeks to forecast, π and ϕ are weights much like regression coefficients, and $\varepsilon_{(.)}$ is an IID (mean-zero) disturbance just as in Equation 2. Equation 5 is called

a model, although there rarely is any substantial subject-matter content. If, in principle, Y_t does not increase over time without limit (which places constraints on the weights), Equation 5 can be used to fit the observed values of Y_t. Ideally, the residuals become statistically indistinguishable from white noise; all the structure in Y_t has been removed. By that criterion, the model is correct. Then, a one-step-ahead forecast to time $t + 1$, for instance, is just πY_t. Proper prediction intervals can follow.

Much richer versions of ARIMA models may be specified, but they still have little or no subject-matter content (Box et al. 2016). There are no first- and second-order conditions as there are for models meant to capture how nature functions. But for valid use, statistical models also have assumptions that must be met. Equation 5 requires that the disturbances are really IID.

There are applications that fall between substantive models and statistical models. Sometimes a statistical model is broadly justified by subject-matter reasoning. For example, Aldor-Noiman and her colleagues (2016) offer a very sophisticated Bayesian, nonparametric, modeling approach for crime data with very low counts. Especially when the spatial and temporal units are small, many crime counts are 0 or integers of 2 or less. Most standard statistical methods will stumble.

Predictive policing forecasts are often a product of self-exciting point process models, first exposited for crime forecasting applications by Mohler and his colleagues (Mohler et al. 2011). Building on ideas from seismology (Helmstetter et al. 2003), the strength of earthquake aftershocks is greater if an earthquake is of greater magnitude at the epicenter. The strength of the aftershocks declines over time and with Euclidian distance from the epicenter. Similar notions are applied to crimes that are likely to be serial. Insofar as one crime begets another, the chances of that subsequent crime decline with time and distance from the ground-zero crime. This spatial-temporal process can be combined with a time-stationary component for each location akin to a hot spot, which can vary over space.

Point process models are a generalization of Poisson processes on a line to Poisson processes on a plane and in practice are a statistical model for counts in time and space. There are important assumptions. For example, the Poisson rate parameter is a deterministic function of spatial location and time, and the background crime events are themselves independent of one another (Mohler et al. 2011). Causal interpretations can be misleading because sequential connections between crimes may not be directly causal but might result from a confluence of similar forcing factors (e.g., burglaries committed independently by different groups of individuals after school near bus stops). Functional forms for the spatial and temporal declines usually are not specified by the statistical model but estimated with kernel-density procedures (Mohler et al. 2011). They come with no statistical guarantees. In practice, forecasting accuracy apparently can be better than predictions derived solely from identified hot spots, but that is not a mathematical requirement and depends on the spatial and temporal units of data being analyzed (Anselin et al. 2000), how distance is measured (Rosser et al. 2017), and even the shape of the spatial units.

The application of point processes to crime forecasting has been extended in several ways. Mohler (2014) predicts homicides and uses a marked point process model with other crime types in given locations as leading indicators. Reinhart & Greenhouse (2018) characterize locations with covariates thought to be related to crime. Rosser & Cheng (2016) show that direction as well as distance can matter, and removing the role of direction appears to eliminate important inaccuracies in the crime data. Flaxman and colleagues (2019) generalize point processes to a Cox process model while using regression kernels in a manner that is related to kriging (Cressie 1993). Mohler and colleagues (2015) conducted a randomized experiment in Los Angeles showing that their approach provides more accurate forecasts of burglaries than forecasts made by Los Angeles Police Department crime analysts using hot-spot maps.

In summary, models of various kinds have to date dominated modern crime forecasting. Variants of point processes have had some demonstrable success and gained many accomplished advocates. However, a close examination of the work makes clear that the models are statistical and in much the same spirit as ARIMA approaches that include exogenous predictors (Box et al. 2016). Statistical formulations fall in the gray area between substantive models and algorithms and, at least from a computer science perspective, are usually not included under the rubric of artificial narrow intelligence. One also must treat accuracy claims cautiously. Such claims rest on the performance of intricate, assumption-driven models with many features of convenience or novelty and, ultimately, all achievements are data-set dependent. No accuracy claims to date are derived from mathematical necessities.

Although far less salient, there are also modeling tools applied to the predictors alone. Perhaps the most common in crime forecasting are latent variable models of predictors formulated as a factor analysis. That is,

$$X_j = \alpha_{0j} + \alpha_{1j}U + \varepsilon_j, \quad j \in P. \tag{6.}$$

U is a latent variable, ε_j is mean-zero IID disturbance term, and X_j is the jth predictor from a set of P predictors (Hastie et al. 2009). Latent variables, seen as causes of observed variables, can be used in an otherwise conventional regression formulation. An important by-product can be dimension reduction of the predictor space.

Empirical summaries in time and space. **Figure 3** shows methods that rely on instructive summaries of data that at least implicitly support forecasting (Rey et al. 2012). Commonly, maps play a key role. Risk-terrain modeling, for instance, builds on the spatial and temporal crime patterns (Bowers et al. 2004, Caplan et al. 2010, Johnson et al. 2007, Ratcliffe 2014). Measures of crime are overlaid on mapping layers of criminogenic predictors. As Caplan & Kennedy (2016) describe the approach (see also Caplan 2011), predictor variables can be binary, empirical densities, or distance metrics from locations of concern. One might notice, for instance, that drug transactions are more common the closer one gets to homeless encampments. The location of homeless concentrations might, therefore, be a map overlay coded simply as a binary presence or absence.

Predictors are given weights that indicate the strength of their relationships with crime by time and place. These weights are, by and large, based on empirical associations, although the step from a measure of association to a weight is typically ad hoc. Statistical tests often are used as well but with no formal rationale. When forecasts are needed, the weighted predictors are aggregated such that higher aggregate values predict more crime. Apparently, seasonal factors frequently are overlooked; usefully accurate prediction for the summer months may stumble during the winter (Szkola et al. 2019). Uncertainty in the projections is not addressed in part because it presents difficult statistical challenges.

There are interesting variations in mapping-based methods. Rosser and her colleagues (2017) use network distance to measure the proximity between locations. Rather than using as-the-crow-flies distance, distance is the sum of street segments required to get from one location to another. Transit routes are also important. A next step might be to formally use Taxicab distance to measure proximity (Malkevitch 2007), which is closely related to Manhattan distance (Munoz et al. 2009).

Mapping methods also lead to tools for working with X alone. Often, a critical question is how to spatially aggregate values of predictors. Suppose one is trying to characterize a particular neighborhood by the number of bars. People in that neighborhood may patronize bars in adjacent neighborhoods as well as their own neighborhood, especially if they live near a neighborhood boundary. Yet the farther a bar is from the neighborhood, the less likely it is that people will make the trip. A statistical tool such as a kernel-density smoother can be used to characterize

the role of that distance. Then, the total number of bars frequented by residents of a particular neighborhood might be a weighted sum with weights derived from the kernel-density smoother: more distance, less weight.

For this discussion, synopsis methods are ways to summarize spatial and temporal patterns in the spirit of all summary statistics. Despite some naming conventions, there are no models; there is no attempt to formally represent how the data were generated. One might argue that synopsis methods are algorithms. There can be some overlap for procedures such as kernel-density smoothers, but mapping procedures are probably best understood as visualization techniques in which forecasting, at least to date, is an afterthought. Their main contribution is explanation, not prediction.

Algorithmic methods. **Figure 3** shows algorithmic methods that are likely to be considered artificial narrow intelligence. The applications to crime forecasting are very recent and relatively few. All would likely be seen as forms of supervised learning because the training data would contain values for predictors and one or more response variables. It is called supervised because the fitting process capitalizes on the known values of the response. The response variables are usually crime counts or crime events.

Supervised learning uses the same kind of data employed in crime forecasting models. But the functions linking inputs to outputs are not specified in advance, as they largely are in modeling approaches. The functions linking inputs to outputs are discovered by an algorithm in a highly inductive manner that computer scientists like to call learning. That learning requires minimizing some loss function, just as model-based methods do, but balancing bias against variance; the goal is to come as close as possible on average over hypothetical realizations of the data to the true relationship between inputs and outputs.

Bias refers to the difference between the expectation of the fitted values and the true response surface. Variance refers to the variability in the fitted values over new realizations of the data. With more complex fitted relationships between inputs and outputs, there is less bias but more variance. With less complex fitted relationships between inputs and outputs, there is more bias but less variance. Because ML algorithms figure out how to make these tradeoffs, they live nicely within the rubric of artificial narrow intelligence. All this really means is that the human-imposed structure that models force on the fitting process is replaced by computational procedures that extract structure from the data.

There is nothing mysterious in play. Consider stepwise regression, which can be seen as an early and clumsy form of ML. In its backward elimination implementation, stepwise regression begins with the largest linear model that the data analyst specifies. In the first step, the procedure (*a*) deletes a predictor, (*b*) stores how much model performance degrades (e.g., the reduction in the adjusted R^2), (*c*) puts the predictor back in the model, and (*d*) repeats the assessment with each predictor in the full model. The predictor that degrades performance the least is deleted from the model, and the evaluation process is repeated with those that remain. With two variables now deleted, the evaluation process moves on with the predictors not yet eliminated. The winnowing continues until dropping any remaining predictors degrades model performance too severely. That smaller model is the one that is used for the data analysis. Given the full model, an algorithm, not a human, determines the best model.

Modern ML is a far more powerful way to proceed and addresses far more than variable selection. Among other accomplishments, it can construct new predictors from old ones. There is generally nothing like a model to constrain how the algorithm proceeds, and the search for a good fit allows for a wide variety of relationships to be explored, many of which can be highly nonlinear.

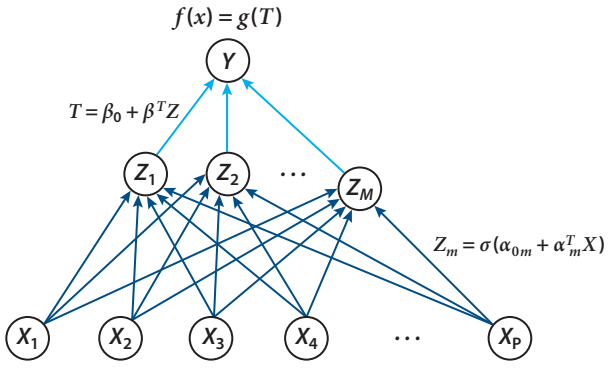

Figure 4

A simple deep-learning neural network with one hidden layer composed of M nodes [adapted from Berk (2020b) and Hastie et al. (2009)].

There are also a variety of safeguards to prevent problems before they appear or fix them after. For example, a form of regularization is commonly used to help protect against overfitting.

There are a large number of popular ML methods: random forests, gradient boosting, support vector machines, and others. Crime forecasting is a relatively recent application, which so far has been dominated by various forms of neural networks, often characterized as deep learning. The details are too lengthy to specify here and can vary depending on the kind of deep learning in play (Goodfellow et al. 2016).

Figure 4 shows a very simple example in which deep applies only as a formality. There are x_1 through x_p inputs, called predictors in statistics. Each is connected to all the nodes z_1 through z_m of a single hidden layer, called latent variables in statistics. The inputs are combined in a linear manner and then transformed by an activation function, σ. The linear combinations serve as regularizers, although there are several other regularization tools that are commonly used (Goodfellow et al. 2016). Activation functions play a similar role to link functions for the generalized linear model (e.g., in logistic regression). They introduce nonlinearities that for deep neural networks are compounded over hidden layers. Usually, there are at least several hidden layers and often many more. Highly nonlinear functions can be constructed.

The transformed linear combinations populate each node in the hidden layer. These transformed linear combinations are then combined linearly and can be transformed by g to provide fitted values for the output Y. Y may be numeric or categorical, which for predictive policing will usually be crime counts or (binary) crime events. There can be more than one Y included as outputs for the same network (e.g., crime numbers and victim numbers). In some applications, there are hundreds of inputs, hundreds of hidden layers, and, consequently, thousands of parameters to be estimated. They play roles similar to the αs and βs in **Figure 4**. Computational challenges can be daunting, in part because the loss function is not convex. But once the loss is minimized, one has an algorithmic structure that can be used for forecasting when new inputs are introduced and the output is unknown.

There are many different flavors of deep learning. Convolution neural networks (CNNs), for example, were developed to classify images (e.g., in facial recognition) and can be used for spatial data; a map is an image. The convolution processes can be seen as part of data preprocessing that extracts the essential features of the image to be passed along as predictors to a neural network (Goodfellow et al. 2016). Recurrent neural networks (RNNs) have been developed for pooled cross-section data (Goodfellow et al. 2016) so that the temporal structure can be exploited.

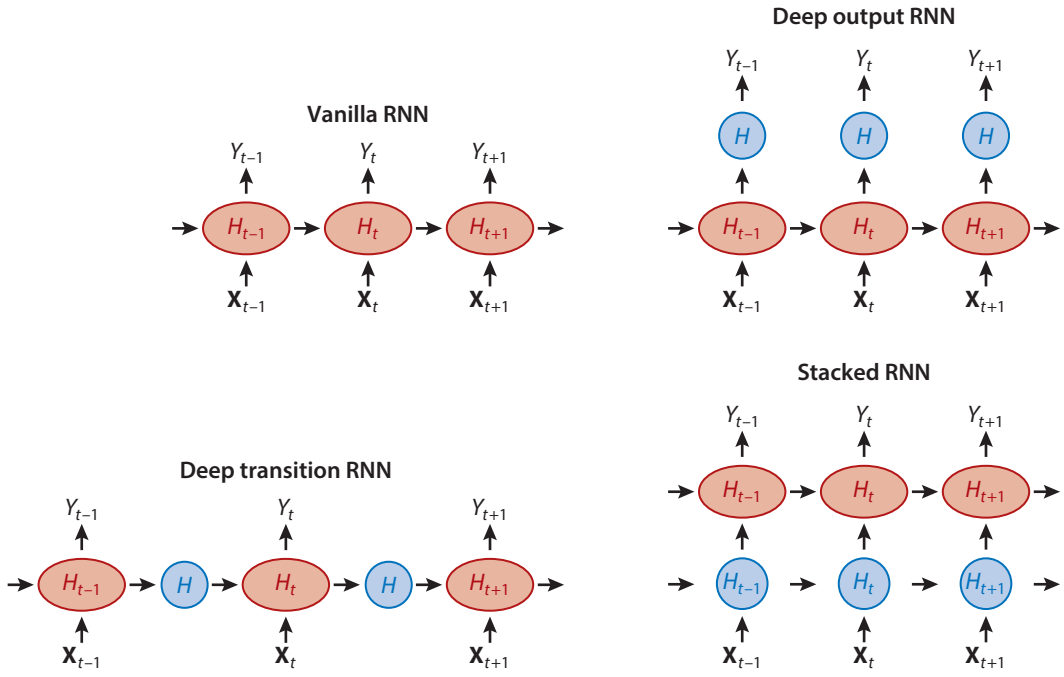

Figure 5

Examples of deep recurring neural network (RNN) structures. Red ellipses represent additional hidden layers beyond those in the vanilla RNNs.

Generative adversarial networks (GNNs) can be used as target hardeners for algorithms vulnerable to hacking (Krohn 2019).

Figure 5, drawing heavily on work by Pascanu and colleagues (2014), shows four different RNN structures that differ in complexity and structure. Each vertical slice has a structure abstracted from **Figure 4**. There is a set of inputs (X_t), a single hidden layer (H_t), and a single output (Y_t). RNNs allow for one such structure at each time point. Conventionally, they share the assumption that dependence over time in the response Y_t is solely a result of associations between the hidden layers over time. Hidden Markov models have the same rationale (Zucchini et al. 2016).

In **Figure 5**, the hidden layers in red ellipses introduce additional nonlinearities at particular points in the network. A greater number of hidden layers generally implies more complex fitted values, and their locations reveal where the additional complexity is needed. One can think of RNNs as nonlinear versions of analysis techniques for cross-sections over time.

Recent papers introduce powerful deep-learning procedures to crime forecasting (Wang et al. 2017a,b). They use the same kinds of data as employed by the more complex modeling approaches discussed earlier but are able to inductively allow for a large number of nonlinear relationships. Zhang & Cheng (2020) developed a variant of deep learning that exploits a network distance metric rather than Euclidian distance as an elaboration of RNNs for forecasting crime. Zhuang and colleagues (Zhuang et al. 2017) used a creative extension of RNNs that includes a spatial component to forecast crime hot spots defined by calls for service. The output predicted is categorical (i.e., hot spot or not), not a crime count. They claim that their approach forecasts more accurately than other deep-learning methods and traditional ML classifiers.

Finally, just as for other predictive policing methods, there are procedures for predictors alone. Some are a rebranding of statistical tools developed more than 50 years ago, typically

as unsupervised learning. Clustering and multidimensional scaling are examples. Autoencoder (Goodfellow et al. 2016) is a new tool that can be seen as a nonlinear version of principal component analysis. It too can be used for dimension reduction but also for a number of novel tasks like denoising data. For example, all handwritten offense reports should have a date that a police officer fills in. But different police officers will often use different date conventions (e.g., 1/6/20 versus 01/06/2020) and sometimes have unreliable penmanship. An autoencoder can be given a large number of inputs that are the dates as written and also the outputs correctly determined by humans. From these data, the algorithm can learn how to reliably go from the handwritten dates to the correct dates so that in future, handwritten dates can be automatically and accurately read by a machine or recoded to a consistent format for data processing.

Criticisms of Predictive Policing

Consider again **Figure 1**. Much of the pushback on predictive police is less about artificial intelligence and more about the activities surrounding it. Ratcliffe & McCullagh (2001) focused on the dissemination and utilization of crime forecasting information. They discovered that the road from forecasts to implementation can be rocky. A lot depends on how the information is transmitted to patrol officers and how it conforms to existing police officer priorities. Accurate forecasts alone are no guarantee of improved policing performance.

Bennett Moses & Chan (2018) write that police agencies are purchasing predictive policing software when evidence of forecasting accuracy is spotty and that, among many other things, algorithms may misdirect oversight concerns from police officers and their superiors to technical details of some forecasting algorithm. In addition, the data on which algorithms are trained can be badly flawed because crimes reported is not the same as crimes committed and reported crimes may be mischaracterized. There are also concerns that disadvantaged neighborhoods will be overpoliced, which can be seen as a form of racial bias.

Kaufmann and colleagues (2019) seek to go beyond the data and forecasting methods to the different ways in which the empirical patterns are translated into larger formulations about the nature of crime and the societal systemic responses. For example, a focus on crime forecasts leads to a societal emphasis on preemptive policing that can become an indictment of neighborhoods that are already disadvantaged. In his important book *The Rise of Big-Data Policing*: *Surveillance*, *Race*, *and the Future of Law Enforcement*, Ferguson (2017) considers a wider range of surveillance practices in the context of US constitutional protections and the growing role of the surveillance state.

There are several responses to the concerns raised. First, much of the uneasiness can be legitimate, but typically has little to do with artificial intelligence itself. In that sense, it is beyond the scope of our discussion.

Second, the baseline for forecasting accuracy is current practice, not perfection. The guiding question is whether predictive policing leads to improvements in accuracy beyond current procedures. There have been some successes, but so far we have no general conclusions one way or the other.

Third, it is surely true that much of the information used for algorithm training is shaped by historical patterns. But does that information lead to substantial racial bias in the forecasts made? Police patrols have long been substantially directed by calls for service. These come disproportionately from disadvantaged neighborhoods. It follows that police can be more densely present in some neighborhoods than others. In turn, for some neighborhoods, this can lead to more official reports of crimes, more apprehensions, and more arrests. But those are not necessarily a product of racial animus or racially motivated administrative practices. Police merely are concentrated where

they are summoned. Furthermore, because clearance rates are generally lower in disadvantaged neighborhoods, rap sheets for many of the neighborhood residents may actually underestimate criminal activities.

Finally, what about the crime victims? There is no debate that crime victims and crime perpetrators disproportionately live in disadvantaged neighborhoods. For example, in Philadelphia last year, there were 361 homicide victims. Most were African-Americans from low-income parts of the city. The same was true of most perpetrators. If one cares about crime victims, law enforcement is surely part of the way cities should respond even though that can mean arresting perpetrators from disadvantaged neighborhoods. Yet most critiques of predictive policing give crime victims no voice.

RISK ASSESSMENT

Because predictive policing and risk assessments are different, so are the data they use, the forecasting procedures they employ, and the jurisprudential issues raised. Predictive policing forecasts crimes by place and time. Risk assessment forecasts crimes to be committed by particular individuals without much regard to where or when. An actuarial approach naturally follows. The kinds of people who will commit crimes in the future are like those who have done so in the past. A key challenge, therefore, is to identify features of past offenders that are strongly associated with criminal behavior. This is little different from forecasting which individuals are likely to suffer from heart disease, which applicants for a job are unlikely to perform well, or which applicants for a mortgage are prone to default on the loan. Often overlooked, however, is that the alternative to a failure is a success; when individuals are not classified as high risk, they can be automatically classified as low risk. Good things can follow. For example, at an arraignment, an individual may be diverted into a drug treatment program.

In the United States, criminal justice risk assessment has a very long history. Parole decisions were in part informed by risk starting in the 1920s (Burgess 1928), and the use of artificial narrow intelligence in law enforcement has been an important component of risk assessment long before it was introduced into predictive policing. For example, Berk and colleagues (2005), building on work done for the California Department of Corrections in the 1990s, used ML to develop intimate partner violence forecasting tools for the Los Angeles County Sheriff's Department.

Risk assessment is now commonly employed to inform a variety of criminal justice decisions: release decisions at arraignment, post-conviction sentencing decisions, assignments to security levels in prison, release decisions by parole boards, and the intensity and nature of supervision on probation and parole (Berk 2018). There are a few jurisdictions in which police use risk assessments and other approaches to identify particular individuals at high risk of being shot or being a shooter (Hollywood et al. 2019).

Over time, quantitative risk assessment has gone from simple scales computed using craft lore and bivariate associations (e.g., between age and crime) to various kinds of traditional regression analysis (e.g., logistic regression) to modern ML. The main rationale for these changes has been demonstrable improvements in forecasting accuracy (Berk & Bleich 2013). The early scales were not only crude by today's standards but also not properly evaluated (Ohlin & Duncan 1949, Reiss 1951). It was extremely difficult to determine how well they forecasted. Regression models clearly have been an improvement and are still commonly used, but they are flawed by modern statistical standards. For example, because the response variable in logistic regression is the logarithm of the odds of some binary outcome (e.g., rearrest), a regression coefficient shows the average change in the logit of Y for a unit change in a given predictor, adjusted for its linear dependence on all other predictors. Yet criminal justice officials want to forecast the chances (i.e., probability) of rearrest. Altering the coefficient for the probability of a particular outcome means employing a nonlinear

transformation from which there is no longer a single coefficient to interpret. It is, therefore, not clear how to properly construct interpretable weights to assign the predictors when a forecast is needed.

Recalling **Figure 1**, risk assessment methods as well as predictive policing methods should be distinguished from the activities surrounding them. With this done, ML as a form of artificial narrow intelligence is proving to be superior to regression models along several dimensions when conducting criminal justice risk assessments.

- Regression models must meet a variety of formal assumptions that usually are implausible in practice (Freedman 2009). ML methods are not constrained in this manner (Hastie et al. 2009).
- ML forecasts as accurately as conventional regression models when inputs are related to outputs in a simple manner but substantially better when inputs are related to outputs in a complex manner, the predictors have distributions that can map well to that complexity, and the ML procedures are properly tuned (Berk 2018, Berk & Bleich 2013, Berk & Sorenson 2020). This point is elaborated on below.
- Criminal justice decisions based on risk assessment are typically categorical (e.g., detain at an arraignment or not). This means that in the end, assessments of risk should be categorical (e.g., high risk or low risk). For regression models, this is typically done, if at all, in an ad hoc manner. Some ML procedures automatically forecast outcome classes in a thoroughly principled manner and most others can be easily altered to do so (Berk 2018).
- Policymakers are of late less concerned about binary outcomes, e.g., recidivate or not, and more concerned with forecasting with three or even four outcome classes. In the case of domestic violence, for instance, at an arraignment one usefully can forecast one of three outcomes: no new reported domestic violence incidents, reported incidents not involving injuries, or reported incidents in which injuries result (Berk & Sorenson 2016). ML can easily accommodate outcomes with more than two categories (Berk 2018).
- Regression models treat the costs of false positives as the same as the costs of false negatives. ML procedures can easily incorporate asymmetric costs of forecasting errors. This can make forecasts more responsive to stakeholder needs (Berk 2020b).
- There is substantial sophistication and improvements in how to address forecasting unfairness when ML forecasting methods are used (Berk et al. 2018, 2020a). Regression models have yet to effectively address issues of racial bias.

Risk Assessment Machine Learning Methods

Perhaps the two most important and effective ML methods used for criminal justice risk assessment are random forests (Breiman 2001a) and stochastic gradient boosting (Friedman 2002). They share a strategy of constructing an ensemble of fitted values from classification or regression trees (Breiman et al. 1984) and then combining them in a sensible fashion for purposes of regularization. **Figure 6** is meant to convey the general idea.

The top row represents M passes over the data, each time addressing the conditional distribution of Y at time t given a set of predictors X also at time t. The setting might be a post-conviction sentencing decision. Y could be an arrest for a violent crime and X might be predictors such as age, prior record, and the current crime of conviction. The asterisks are meant to convey that one is not working with the original data in the form it was generated. Rather, one is working with a random sample with replacement, sometimes called a bootstrap sample, or with reweighted observations as described earlier. The M samples being analyzed can differ in other ways as well.

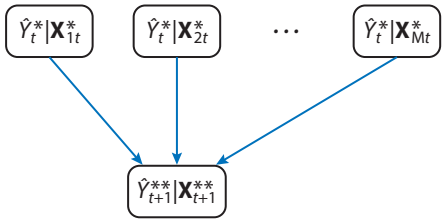

Figure 6

Stylized ensemble methods for machine learning.

Here are the algorithmic steps for random forests (Berk 2020b). Let N be the number of observations in the training data and assume for now that the response variable is binary.

1. Take a random sample of size N with replacement from the data.
2. Take a random sample without replacement of the predictors.
3. Construct the first recursive partition for a classification tree of the data as usual.
4. Repeat Step 2 for each subsequent split until the tree is as large as desired. Often this leads to one observation in each terminal node. Do not prune. Compute each terminal node proportion as usual.
5. Drop the out-of-bag (OOB) data down the tree. The OOB data were not used to grow the tree because they were by chance excluded when the random sample of the data was drawn with replacement. Save the class assigned to each OOB observation along with each observation's predictor values. The OOB data help reduce overfitting because they were not used for training,
6. Repeat steps 1–5 many times (e.g., 500).
7. Using only the class assigned to each observation when that observation is OOB, count the number of times over different trees that the observation is classified in one category and the number of times over different trees it is classified in the other category.
8. Assign each case to an outcome category by a majority vote over the set of trees when that case is OOB. Thus, if, over a large number of trees, 51% of the time or more a given case is classified as a 1, then that becomes its assigned class. If, over a large number of trees, 51% of the time or more a given case is classified as a 0, then that becomes its assigned class.
9. If forecasts are needed, the new IID cases from the same joint probability distribution are dropped down each tree and depending on the terminal node in which a case falls, a class is assigned as usual. There is then a vote over different trees through which the winning class becomes the forecasted class. For numeric response variables or if forecasted conditional proportions are desired, one simply averages the terminal node conditional means or conditional proportions for each case. These averages serve as forecasts.

Stochastic gradient boosting can also be sequentially summarized using a binary response variable here for simplicity (Berk 2020b).

1. There is a binary response variable in the training data coded as 1 or 0. The algorithm is initialized with some constant such as the overall proportion of 1s. This constant serves as the fitted values from which residuals are obtained by subtraction in the usual way. The residuals are then appended to the training data as a new variable.
2. A random sample of the data is drawn without replacement. One might, for example, sample half the data. Formally, the sampling is not necessary, but it seems to help performance.

3. Because the residuals are numeric, a regression tree, not a classification tree, is applied to the sample with the residuals as the response. Another set of fitted values is obtained and used to update the prior set of fitted values. From these, a new set of residuals is obtained and appended. In effect, the residuals serve as weights such that observations that are less well fit in one pass through the data are made more important for the next pass through the data.
4. Another random sample is taken from the training data, and the fitting process is repeated. As before, the sampling is not formally necessary.
5. The entire cycle is repeated many times: (*a*) fitted values, (*b*) residuals, (*c*) sampling, and (*d*) a regression tree. In the end, the fitted values from each pass through the data have been combined in a linear fashion. For classification, these aggregated fitted values can be interpreted as probabilities.
6. Commonly, observations with $\hat{y}_i > 0.5$ are assigned a 1 and observations with $\hat{y}_i \leq 0.5$ are assigned a 0. Numeric response variables are also permitted. The steps are largely the same but without an assigned class.

One central point is that weighting, which is so fundamental to boosting, occurs implicitly through the residuals from each pass. Larger positive or negative residuals imply that for those observations, the fitted values perform less well. As each regression tree attempts to maximize its fit, the tree responds more to the observations that in the previous iteration had larger positive or negative residuals.

Ensemble procedures can forecast very well. One important reason is that each decision tree can produce a very complex set of fitted values that is regularized by combining the fitted values over different trees. One can arrive at a good balance between the bias and the variance. Other reasons are discussed elsewhere (Berk 2020b).

Accuracy, Fairness, and Transparency

Almost regardless of the type of artificial narrow intelligence or the applications, there has been a range of concerns about matters that go well beyond the technology. Questions about accuracy address the algorithmic fitting and forecasting errors. Fairness involves treating similarly situated groups similarly. Transparency refers to the accessibility of ML technology to stakeholders.

Accuracy. Accuracy is usually treated in reverse: How often does the ML algorithm get it wrong? For criminal justice risk assessment, there are three popular definitions of error. The first is the overall fraction of cases for which the actual outcome class differs from the assigned outcome class. Given the known outcome classes in the training and/or test data, what proportion of the time does the algorithm identify them incorrectly? Sometimes this is interpreted as the overall probability of a classification error.

An important problem with the overall measure of error is that the costs of forecasting errors usually differ depending on the outcome class. One way to address this problem is to use the available data to compute an error rate separately for each observed outcome class. For example, given the subset of individuals who actually failed on parole, what proportion of the time does the algorithm not identify them correctly? Likewise, given the subset of individuals who actually did not fail on parole, what proportion of the time does the algorithm not identify them correctly? Sometimes these proportions are treated as if they are probabilities.

If failure on parole is a negative, the first is called the false-positive rate. If success on parole is a positive, the second is called the false-negative rate. The logic applies to more than two outcome

classes, but these naming conventions become unclear. If one outcome class is called a positive and another outcome class is called a negative, what names are attached to the other outcome classes?

In practice, an individual's outcome class is not known when a forecast is required (or there would be no need for a forecast). The first three definitions condition on the known outcome class and are used to measure algorithmic performance. In practice, more interest targets forecasting accuracy. Using test data, one conditions on the forecasted outcome class and ascertains the proportion of times the forecasted outcome class is wrong. Typically, forecasting error is computed separately for each outcome class, which is often interpreted as probabilities.

Because most risk algorithms minimize a loss function, anything that constrains that optimization reduces accuracy (Kearns et al. 2018a). There is a mathematically required price. Constraints to improve fairness (see below) are a common instance. For example, at arraignments, more offenders are detained who pose no risk to public safety, and more offenders are released who pose a threat to public safety.

Practitioners are sometimes flummoxed when ML risk accuracy is on occasion little different from that of conventional regression models. However, any conclusions about relative performance need to be based on a solid understanding of how ML risk forecasts are made.

- Sometimes the high-risk offenders and the low-risk offenders are patently obvious. Most risk assessments, even just a casual reading of a presentencing report, agree on which is which. The real challenge is accurately classifying offenders in the middle. Recent work (Berk 2017) suggests that for parole board decisions, ML earns its keep by correctly classifying the equivocal cases. Inmates prone to violence and those likely to become model citizens are apparent to all. In other words, most any of the more common risk tools would perform reasonably well over the full range of inmates, and that is often how performance comparisons are made. Narrowly crafted accuracy assessments can be much more instructive.
- Suppose an ML procedure estimates with a probability of 0.95 that a given offender will be rearrested. A linear model estimates that probability as 0.55. These are very different probabilities of risk. Yet both classify the offender as high risk because both probabilities exceed a threshold of 0.50. Classification, therefore, masks important performance differences, and, along with a forecast, the numerical values that determined the forecast should be provided. For example, measures like the vote from random forests should be coupled with each forecasted outcome class (even though those votes are not probabilities). The larger the fraction of the vote is for a winning outcome class, the more reliable the forecast; out-of-sample performance is generally better. Such performance differences will not be apparent from the forecasted class alone.
- When the true response surface is relatively simple, more complex sets of fitted values usually do not improve performance much. When the true response surface is complex, one needs complex fitted values. But a complex set of fitted values depends substantially on the quality of the predictors. If they are too weak to capture complex surfaces, ML may not perform better than conventional regression. For example, if a binary outcome has a strongly unbalanced distribution (e.g., the probability of a rearrest for a violent crime is only 0.05), having at least some relevant predictors with long right or left tails can dramatically help.
- In part because the primary outputs from ML procedures are sets of fitted values, and there is no pretense of a formally correct model, the bias–variance tradeoff becomes essential. When the true response surface is simple, the bias–variance tradeoff matters little. There is no need to trade variance for bias because making the fitted values more complex is not typically helpful. A variety of regression-like procedures perform equally well. When the

true response surface is complex, the bias–variance tradeoff can be usefully exploited, which is precisely what effective ML does and conventional regression does not.

- Finally, performance depends on how an ML algorithm is tuned. Often comparable performance with conventional procedures comes from faulty tuning. Relying on the default values of the tuning parameters, for instance, is often a serious error that cripples risk algorithms. It is also unwise to rely exclusively on automated fitting criteria such as cross-validation. Many ML risk algorithms are sample-size dependent, and the folds required by cross-validation necessarily reduce the effective sample size. Accuracy can suffer.

Fairness. The analysis of fairness for criminal justice risk assessment is much more developed than the analysis of fairness for predictive policing. Many different academic disciplines have weighed in (Berk et al. 2018, Corbett-Davies & Goel 2018, Goel et al. 2019, Hamilton 2016, Huq 2019, Kearns et al. 2018b, Kleinberg et al. 2017, Mayson 2019). In criminal justice settings, fairness represents equal treatment across groups. There is a related approach in which fairness is defined as equal treatment across people, which has been used in other kinds of applications such as privacy (Dwork & Roth 2014).

There are often direct links between common measures of accuracy and different definitions of fairness.

- In equality of outcome, each different group is represented in equal proportions for some outcome. For example, the same proportion of men and women are detained at an arraignment; e.g., 26% of the men and 26% of the women are detained.
- In equality of opportunity, for each outcome class, every group is classified with the same accuracy. For two outcomes, this means, for example, the same false-positive and false-negative rates for blacks and whites. The classifier works equally well for both races.
- In equal prediction accuracy, every group's outcome class is forecasted with equal accuracy. For example, individuals of Hispanic descent and individuals of Chinese descent have forecasts of rearrest that are correct 65% of the time and forecasts of no rearrest that are correct 72% of the time.
- In equality of treatment, the cost ratio of false positives to false negatives is the same for every group. For instance, there are three false positives for every false negative for American citizens born in the United States and American citizens born in Somalia. Equal treatment is implicated because the cost ratio indicates how the algorithm is trading one kind of classification error for another. In this 3 to 1 illustration, false negatives are three times more costly than false positives; thus, the algorithm works harder to avoid them.

Unfortunately, you can't have it all. Not only is accuracy affected by efforts to improve fairness, but there are often tradeoffs between different kinds of fairness. In particular, if "the base rates differ across groups, any [risk] instrument that satisfies predictive parity at a given threshold. . .*must* have imbalanced false positive or false negative errors rates at that threshold" (Chouldechova 2017, p. 5; emphasis in original).

A base rate is the distribution of the outcome classes. For example, if the proportion of men who fail to appear (FTA) when required to return to court is 0.65, the FTA base rate is 0.65. If the proportion of women who fail to appear when required to return to court is 0.55, the FTA base rate is 0.55. The threshold refers to some cutoff on a risk score such that those offenders who are above that score have a particular forecasted outcome class. An estimated probability of failing a drug test greater than 0.75 might be used to forecast failing a drug test on parole. The term predictive parity means the same as equal forecasting accuracy.

The message is that given different base rates, there is an inevitable tradeoff between equal false-positive or equal false-negative rates on one hand and equal forecasting accuracy on the other. More generally, the several kinds of accuracy and several kinds of fairness require a far more nuanced discussion than commonly undertaken. Much too often, stakeholders just talk past one another.

Transparency. ML algorithms are often characterized as a black box. The manner in which inputs are related to outcome is obscure. At the very least, black-box algorithms can make stakeholders uneasy.

There are several ways to constructively respond. First, the human brain is a black box too. For example, when a magistrate determines that some individuals are to be detained and others not, the reasons usually are not apparent. Even if at sentencing a judge's reasoning is explained, the content can be very abbreviated, subject to public relations concerns, especially if in that jurisdiction judges are elected, and shaped by unconscious bias (Rachlinski et al. 2009). Even more difficult to comprehend are the decisions by parole boards that are made behind closed doors. The transparency bar is very low, and algorithms can compete.

Second, there are ways to construct algorithms that are easier to understand. For example, Zeng and colleagues (2017) develop a linear approach that uses only binary predictors and integer weights. Each unique value for numeric variables has its own binary variable coded 1 or 0 so that nonlinear relationships can be computed, and the different units of the original numeric variables no longer complicate interpretations. The integer weights are also easier to interpret than conventional regression coefficients. A penalized discrete optimization approach is used that seeks to minimize the number of classification errors and, in the spirit of the lasso (Tibshirani 1996), produces a sparse set of predictors.

Third, there is a cottage industry of algorithms meant to complement ML procedures that extract useful information, often using visualizations, about otherwise obscure processes. For example, there are several ways in which the contributions to the fitted values can be calculated and displayed. One simple approach computes the decline in accuracy when each predictor in turn is precluded from affecting the fitted values (Breiman 2001a). Another approach displays how each predictor is related to the response holding all other predictors constant without the use of the standard covariance adjustments employed by linear regression (Hastie et al. 2009).

Fourth, one can often provide answers to specific questions about what is going on under the hood. For example, one can take a single case and obtain the fitted class. Then for the same case, the value of some predictor such as age is changed. One can learn whether the fitted class would have been meaningfully different if the individual were 10 years older, with everything else about the individual fixed. In this manner, the way the algorithm uses age can be easily demonstrated.

Finally, unlike the human brain, there is code that details exactly how risk is computed. If the software is not proprietary, the code can be studied. Most stakeholders will not have the skill or inclination to take on this task, but many universities and business establishments have individuals who can lend a hand.

In short, ML risk assessment has a public relations problem. The hype surrounding artificial intelligence encourages science fiction parallels that are very misleading. Unbridled enthusiasm from some ML proponents reinforces the narrative. Private firms that sell ML software capitalize on both through slick advertising and well-crafted sales pitches. It should be no surprise that many stakeholders feel overmatched.

FUTURE DEVELOPMENTS

Over the next decade, likely developments are easy to anticipate. Although the algorithms used for predictive policing and criminal justice risk assessment will continue to improve, the pace of

change will slow. The major algorithmic advances probably have been delivered already. What remains for the software is fine-tuning, although new risk outcomes will be predicted. For example, there is already work in progress to forecast whether an apprehension will lead to a shooting of either the suspect or any of the arresting officers. The increasing use of employment assessment algorithms in the private sector will migrate widely to police recruitment, promotion, and disciplinary actions. Beyond predictive policing and criminal justice risk assessment, some previously fanciful kinds of algorithmic applications will grow in importance and experience substantial performance improvements, particularly within robotic devices, as military applications lead the way (Kessel 2019). There are already piloted drone devices that can search an abandoned building and demonstrations that self-guided quadrotor swarms will soon be able to do the same much more quickly (Ackerman 2017).

Computer processing power will continue to grow rapidly. Concurrently, these developments will be provided in smaller housings so that, for instance, hand-held devices will have the capabilities of today's large desktop machines. Significant progress is already being made building computer code into the design of computer chips. The role of the software middle man is fading. Artificial narrow intelligence will soon be mobile, powerful, and part of the routine equipment used by police officers on patrol. A wide variety of policing tasks will be affected beyond those addressed in this review.

Databases will become much larger, more accurate, and dramatically richer. Existing data sets will be merged so that, for instance, detailed measures of performance on parole will be combined with data on employment, health, and social media use as well as with similar information on family and friends. Widespread surveillance data will also be readily available through devices like CCTV, cell phones, drones, automobile black boxes, police body cameras, and ankle bracelets. The accuracy and fairness of forecasts could be dramatically improved. Several Scandinavian countries are already far down this path; for instance, it seems quite possible to forecast with a useful level of accuracy whether a newborn infant will be arrested for a felony by the time he or she is 20 years old (Lyngstad et al. 2017).

As with all new technology, there are major risks attached to the significant benefits. Privacy concerns immediately come to mind. Misuse of the forecasts is also a potential hazard. For example, might forecasts of malfeasance be sold to health insurance firms or prospective employers or used for extortion? More generally, might the capacities developed for law enforcement be ported to other applications? Suppression of political dissent is an obvious example that seems well underway in some countries. None of these concerns, and others closely related, are new. Safeguards are needed. In the United States, Europe, and elsewhere, political institutions are beginning to respond. Citizen groups are forming and many NGOs are engaged. There is even counter-surveillance technology being developed that individuals can obtain. For example, eyeglasses are now available that diffuse the light from a CCTV camera to prevent facial recognition. But how all this will play out is unclear, and substantial worry is healthy. "We are playing cat and mouse. . .and that always ends poorly for the mouse" (Woodrow Hartzog, quoted in Hill 2020, p. 3).

All these developments will be implemented using artificial narrow intelligence applied to extensions of activities that are already being undertaken or that have been proposed. There is no need to consider science fiction scenarios. The real world is challenging enough.

SUMMARY AND CONCLUSIONS

Predictive policing and criminal justice risk assessment can be seen as examples of artificial narrow intelligence that do better what is already being done. Current practice is a very long way from

artificial general intelligence, let alone artificial superintelligence. Despite some media portrayals, the singularity is nowhere in sight.

Algorithms are very different from models. Models are mathematical expressions about how the world works. Algorithms are methods for computation. Concerns that may be central for one may be irrelevant, or even misleading, for the other. One powerful example is that there is no such thing as mean function misspecification for algorithms. Another powerful example is that for algorithms, explanation takes a distant back seat to prediction; accurate forecasting does not depend on subject-matter understanding.

Concerns about algorithms in criminal justice settings must allow that algorithms by themselves are rarely the problem. Accuracy and fairness stem far more from the data and how they are managed. And there can be no material benefit or harm until algorithm-informed decisions are made and actions are taken. Furthermore, the performance benchmark is not perfection but current practice. The goal of predictive policing and criminal justice risk assessment is just to do better.

Predictive policing forecasts crimes in place and time. To date, models have dominated practice and performance, especially models based on point processes and their extensions. Algorithms recently have made an appearance, typically as some form of deep learning. There are no definitive, defensible claims of superior accuracy for any model, algorithm, or synopsis method because all such claims depend on the setting and the data; all results are setting and data-set specific. Such dependence is almost universal in organized forecasting competitions as well. There are no formal mathematical results that preordain a winner. Although the balance of evidence seems to favor predictive policing compared to existing policing practices, the winning margin is often small and limited to certain kinds of crimes. Finally, as emphasized earlier, there are no guarantees that police officers will use the forecasts as intended or even have tactics that can capitalize on them.

Criminal justice risk assessments forecast which individuals will re-offend among those already being supervised or processed by the criminal justice system. Many legacy risk tools dating back decades are still in use, but ML classifiers are gaining proponents. There are well understood formal reasons why ML forecasts of risk will be at least as accurate as the model-based competitors and, with complex relationships between inputs and outputs, can perform better. The use of regularized sets of complex fitted values offers a superior tradeoff between the bias and the variance. At the same time, all the more powerful ML procedures offer comparable accuracy because they are operating with much the same underlying principles. Just as for predictive policing, however, there are no guarantees that a risk assessment will be properly used or, even if it is, that effective practices can be implemented.

The mind of a criminal justice decision-maker is a black box, and, relative to that baseline, ML risk assessments can be more transparent. Relative accuracy is more difficult to ascertain because there are several kinds. Nevertheless, research in cognitive science, as well as criminal justice, strongly favors quantitative risk assessment. There are also several different kinds of fairness and inevitable tradeoffs between them coupled with inevitable tensions between fairness and accuracy. You can't have it all. There can be no technical fix. Legal, political, and legislative processes must strike an acceptable balance between competing priorities, which is, in any case, the way we currently do things.

Finally, the future will bring further improvements in predictive policing and criminal justice risk assessment. The power of these tools will grow dramatically with new computer technology and far better data sets. This will lead to spillovers into applications outside of criminal justice, many of which may not be benign. There will also be an influx of high-tech tools from military applications, and some will pose significant risks (**https://www.cna.org/CAAI/reports**). All this will be substantially driven by algorithms that embody artificial narrow intelligence.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

A special thanks goes to Penn Cary Coglianese, Michael Kearns, and Aaron Roth for helping to introduce me to many of the issues discussed and to Penn colleagues Andreas Buja, Eric Tchetgen Tchetgen, Ed George, Linda Zhao, and Arun Kuchibhotla for hours of technical instruction.

LITERATURE CITED

Ackerman E. 2017. This autonomous swarm doesn't need GPS. *IEEE Spectrum*, Dec. 27. **https://spectrum.ieee.org/automaton/robotics/drones/this-autonomous-quadrotor-swarm-doesnt-need-gps**

Aldor-Noiman S, Brown LD, Fox EB, Stine RA. 2016. Spatio-temporal low count processes with applications to violent crime events. *Stat. Sin.* 26:1587–610

Angrist JD, Pischke J-S. 2009. *Most Harmless Econometrics*. Princeton, NJ: Princeton Univ. Press

Angwin A, Larson J, Mattui S, Kirchner L. 2016. Machine bias. *ProPublica*, May 23. **https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing**

Anselin L, Cohen J, Cook D, Goor W, Tita G. 2000. Spatial analyses of crime. In *Criminal Justice 2000*: *Measurement and Analysis of Crime and Justice*, ed. R Kaminsk, N La Vigne, pp. 213–62. Washington, DC: Natl. Inst. Justice

Arute F, Arya K, Babbuch R, Bacon D, Bardin JC, et al. 2019. Quantum supremacy using a programmable superconducting processor. *Nature* 575:505–10

Bennett Moses L, Chan J. 2018. Algorithmic prediction in policing: assumptions, evaluation, and accountability. *Polic. Soc.* 28(7):806–22

Berk RA. 2017. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *J. Exp. Criminol.* 13:193–216

Berk RA. 2018. *Machine Learning Forecasts of Risk in Criminal Justice Settings*. Cham, Switz.: Springer

Berk RA. 2020a. Almost politically acceptable criminal justice risk assessment. *Criminol. Public Policy*. In press

Berk RA. 2020b. *Statistical Learning from a Regression Perspective*. Cham, Switz.: Springer. 3rd ed.

Berk RA, Bleich J. 2013. Statistical procedures for forecasting criminal behavior: a comparative assessment. *J. Criminol. Public Policy* 12(3):515–44

Berk RA, Brown L, Buja A, George E, Pitkin E, et al. 2014. Misspecified mean function regression: making good use of regression models that are wrong. *Sociol. Methods Res.* 43:433–51

Berk RA, Buja A, Brown L, George E, Kuchibhotla AK, et al. 2019. Assumption lean regression. *Am. Stat.* **https://doi.org/10.1080/00031305.2019.1592781**

Berk RA, Heidari H, Jabbari S, Kearns M, Roth A. 2018. Fairness in criminal justice risk assessment: the state of the art. *Sociol. Methods Res.* **https://doi.org/10.1177/0049124118782533**

Berk RA, Sorenson SB. 2016. Forecasting domestic violence: a machine learning approach to help inform arraignment decisions. *J. Empir. Legal Stud.* 13(1):95–115

Berk RA, Sorenson SB. 2020. An algorithmic approach to forecasting rare violent events: an illustration based on intimate partner violence perpetration. *Criminol. Public Policy* 19(1):213–33

Berk RA, Sorenson SB, He Y. 2005. Developing a practical forecasting screener for domestic violence. *Eval. Rev.* 29(4):358–82

Bowers KJ, Johnson SD, Pease K. 2004. Prospective hot-spotting: the future of crime mapping? *Br. J. Criminol.* 44:541–658

Box GEP, Jenkins GW, Reinsel GC, Ljung GM. 2016. *Time Series Analysis*: *Forecasting and Control*. Hoboken, NJ: Wiley

Braga AA, Papachristos AV, Hureau DH. 2014. The effects of hot spots policing on crime: an updated systematic review and meta-analysis. *Justice Q.* 31(4):633–63

Breiman L. 2001a. Random forests. *Mach. Learn.* 45:5–32
Breiman L. 2001b. Statistical modeling: two cultures (with discussion). *Stat. Sci.* 16:199–231
Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and Regression Trees*. Monterey, CA: Wadsworth Press
Burgess EM. 1928. Factors determining success or failure on parole. In *The Working of the Indeterminate Sentence Law and the Parole System in Illinois*, ed. AA Bruce, AJ Harno, EW Burgess, EW Landesco, pp. 205–49. Springfield, IL: State Board Parole
Cameron AC, Trivedi PK. 2005. *Macroeconometrics*: *Methods and Applications*. New York: Cambridge Univ. Press
Caplan JM. 2011. Mapping the spatial influence of crime correlates: a comparison of operational schemes and implications for crime analysis and criminal justice practice. *J. Policy Dev. Res.* 13(3):57–83
Caplan JM, Kennedy LW, Miller J. 2010. Risk terrain modeling: brokering criminological theory and GIS methods for crime forecasting. *Justice Q.* 28(2):360–81
Caplan JM, Kennedy LW. 2016. *Risk Terrain Modeling*: *Crime Prediction and Risk Reduction*. Berkeley: Univ. Calif. Press
Cellen-Jones R. 2014. Stephen Hawking warns artificial intelligence could end mankind. *BBC News*: *Technology*, Dec. 2. **https://www.bbc.com/news/technology-30290540**
Chinoy S. 2019. The racist history behind facial recognition. *New York Times*, July 10. **https://www.nytimes.com/2019/07/10/opinion/facial-recognition-race.html**
Chouldechova A. 2017. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. arXiv:1610.075254v1 [stat.AP]
Coglianese C, Lehr D. 2017. Regulating by robot: administrative decision making in the machine-learning era. *Georget. Law J.* 105:1147–223
Corbett-Davies S, Goel S. 2018. *The measure and mismeasure of fairness*: *a critical review of fair machine learning*. Paper presented at the 35th International Conference on Machine Learning, Stockholm
Cressie NAC. 1993. *Statistics for Spatial Data*. Hoboken, NJ: Wiley
Dinis-Oliveira RJ. 2017. Metabolic profile of flunitrazepam: clinical and forensic toxicological aspects. *Drug Metab. Lett.* 11(1):14–20
Domingos P. 2015. *The Master Algorithm*: *How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books
D'Orsogna MR, Perc M. 2015. Statistical physics of crime: a review. *Phys. Life Rev.* 12:1–21
Dwork C, Roth A. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9(3–4):211–407
Ferguson AG. 2017. *The Rise of Big-Data Policing*: *Surveillance*, *Race*, *and the Future of Law Enforcement*. New York: NYU Press
Flaxman S, Chirico M, Pereira P, Loeffler C. 2019. Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: a winning solution to the NIJ "Real-Time Crime Forecasting Solution." *Ann. Appl. Stat.* 13(4):2564–85
Freedman DA. 2009. *Statistical Models*: *Theories and Practice*. Cambridge, UK: Cambridge Univ. Press. 2nd ed.
Friedman JS. 2002. Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38:367–78
Goel S, Shroff R, Skeem JL, Slobogin C. 2019. The accuracy, equity, and jurisprudence of criminal risk assessment. *Risk-Resil. Res.* In press. **http://risk-resilience.berkeley.edu/sites/default/files/journal-articles/files/manuscript_the_accuracy_equity_and_jurisprudence_of_criminal_risk_assessment_1.2.19_.pdf**
Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. Cambridge, MA: MIT Press
Grahm O, Dowe D. 2019. The Turing test. In *The Stanford Encyclopedia of Philosophy*, ed. EZ Zalta. Stanford, CA: CSLI
Groff ER, La Vigne NG. 2004. Forecasting the future of predictive crime mapping. *Crime Prev. Stud.* 13:29–57
Hamilton M. 2016. Risk-needs assessment: constitutional and ethical challenges. *Am. Crim. Law Rev.* 52(2):231–92
Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. New York: Springer. 2nd ed.
Helmstetter A, Ouillon G, Sornette D. 2003. Are aftershocks for large California earthquakes diffusing? *J. Geophys. Res.* 108(B10):2483

Hill K. 2020. Activate this 'bracelet of silence,' and Alexa can't eavesdrop. *New York Times*, Febr. 16, Sect. BU, p. 3

Hollywood JS, McKay KN, Woods D, Agneil D. 2019. *Real-Time Crime Centers in Chicago*. Santa Monica, CA: RAND

Hu Y, Wang F, Guin C, Zhu H. 2018. A spatial-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Appl. Geogr.* 9:89–97

Huq AZ. 2019. Racial equality in algorithmic criminal justice. *Duke Law J.* 68(6):1043–134

Johnson SD, Birks DJ, McLaughlin L, Bowers KJ, Pease K. 2007. *Prospective crime mapping in operational context*. Home Off. Rep. 19/07, Res. Dev. Stat. Dir., London

Kaufmann M, Egbert S, Leese M. 2019. Predictive policing and the politics of patterns. *Br. J. Crimiol.* 59:674–92

Kearns M, Neel S, Roth A, Wu Z. 2018a. An empirical study of rich subgroup fairness for machine learning. arXiv:1808.08166v1 [cs.LG]

Kearns M, Neel S, Roth A, Wu Z. 2018b. Preventing fairness gerrymandering: auditing and learning subgroup fairness. arXiv 1711.05144v4 [cs.LG]

Kearns M, Roth A. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford, UK: Oxford Univ. Press

Kessel JM. 2019. Killer robots aren't regulated. Yet. *New York Times*, Dec. 13. **https://www.nytimes.com/2019/12/13/technology/autonomous-weapons-video.html**

Kleinberg J, Mullainathan S, Raghavan M. 2017. *Inherent tradeoffs in the fair determination of risk scores*. Paper presented at the 8th Innovations in Theoretical Computer Science Conference, Berkeley

Krohn J. 2019. *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*. Boston: Addison-Wesley

Kroll JA, Huey J, Barocas S, Felten EW. 2017. Accountable algorithms. *Univ. Pa. Law Rev.* 165(3):633–705

Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. 2018. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* 113(523):1094–111

Liesenfeld R, Richard J-F, Vogler J. 2017. Likelihood-based inference and prediction in spatio-temporal panel count models of urban crimes. *J. Appl. Econom.* 32:500–620

Lyngstad T, Skardhammar T, Berk R. 2017. *Predicting future crime at birth*. Paper presented at the 73rd Annual Meeting of the American Society for Criminology, Philadelphia

Malkevitch J. 2007. Taxi! *American Mathematical Society*, October. **https://www.ams.org/publicoutreach/feature-column/fcarc-taxi**

Mayson SG. 2019. Bias in, bias out. *Yale Law J.* 128:2218–300

Metz C. 2019. Google claims a quantum breakthrough that could change computing. *New York Times*, Oct. 23. **https://www.nytimes.com/2019/10/23/technology/quantum-computing-google.html**

Mitchell M. 2019a. *Artificial Intelligence: A Guide for Thinking Humans*. New York: Farrar, Straus and Giroux

Mitchell M. 2019b. We shouldn't be scared of "superintelligent A.I." *New York Times*, Oct. 31. **https://www.nytimes.com/2019/10/31/opinion/superintelligent-artificial-intelligence.html**

Mohler G. 2014. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* 30(3):491–97

Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, Tita GE. 2011. Self-exciting point process modeling of crime. *J. Am. Stat. Assoc.* 493:100–8

Mohler GO, Short MB, Malinowski S, Johnson M, Tita GE, et al. 2015. Randomized controlled field trials of predictive policing. *J. Am. Stat. Assoc.* 110(512):1399–411

Mullainathan S. 2019. Biased algorithms are easier to fix than biased people. *New York Times*, Dec. 6. **https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html**

Munoz D, Bouchereau F, Vargas C, Enriquez R. 2009. Heuristic approaches to the position location problem. In *Position Location Techniques and Applications*, pp. 103–52. Cambridge, MA: Academic

Ohlin LE, Duncan OD. 1949. The efficiency of prediction in criminology. *Am. J. Sociol.* 54:441–52

Pande V. 2018. Artificial intelligence's 'black box' is nothing to fear. *New York Times*, Jan. 25. **https://www.nytimes.com/2018/01/25/opinion/artificial-intelligence-black-box.html**

Pascanu R, Gulcehere C, Cho K, Begio Y. 2014. How to construct deep recurrent neural networks. arXiv:13126026v5 [cs.NE]

Perry WL, McInnis B, Price CC, Smith SC, Hollywood JS. 2013. *Predictive Policing: The Role of Crime Forecasting in Law Enforcement and Operations*. Santa Monica, CA: RAND

Pindyck RS, Rubinfeld DL. 1981. *Econometric Models and Economic Forecasts*. New York: McGraw Hill. 2nd ed.

Poole D, Mackworth A, Goebel R. 1998. *Computational Intelligence: A Logic Approach*. Oxford, UK: Oxford Univ. Press

Rachlinski JJ, Johnson SL, Wistrich AJ, Guthrie C. 2009. Does unconscious racial bias affect trial judges? *Notre Dame Law Rev*. 84(3):1195–252

Ratcliffe JH. 2014. The hotspot matrix: a framework for spatio-temporal targeting of crime reduction. *Police Pract. Res.* 5(1):5–23

Ratcliffe JH, McCullagh MJ. 2001. Chasing ghosts? Police perception of high crime areas. *Br. J. Criminol.* 41:330–41

Reinhart A, Greenhouse J. 2018. Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *J. R. Stat. Soc. C* 67(5):1305–29

Reiss AJ. 1951. The accuracy, efficiency, and validity of a prediction instrument. *Am. J. Sociol.* 17:268–74

Rey SJ, Mack EA, Koschinsky J. 2012. Exploratory space-time analysis of burglary patterns. *J. Quant. Criminol.* 28:509–31

Rosser G, Cheng T. 2016. Improving the robustness and accuracy of crime prediction with a self-exciting point process through isotropic triggering. *Appl. Spat. Anal.* 12:5–25

Rosser G, Davies T, Bowers K, Johnson SD, Cheng T. 2017. Predictive crime mapping: arbitrary grids or street networks? *J. Quant. Criminol.* 33:569–94

Scharre P. 2018. *Army of None: Autonomous Weapons and the Future of War*. New York: Norton

Smith CS. 2019. Dealing with bias in artificial intelligence: three women with extensive experience in A.I. spoke on the topic and how to confront it. *New York Times*, Novemb. 19. **https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html**

Spielkamp M. 2017. Inspecting algorithms for bias. *MIT Technology Review*, Jan. 12. **https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias**

Starr S. 2014. Sentencing by the numbers. *New York Times*, Aug. 10. **https://www.nytimes.com/2005/01/02/magazine/sentencing-by-the-numbers.html**

Szkola J, Piza EL, Drawve G. 2019. Risk terrain modeling: seasonality and predictive validity. *Justice Q*. **http://doi.org/10.1080/07418825.2019.1630472**

Tarling R, Perry JA. 1981. Statistical methods in criminological prediction. In *Prediction in Criminology*, ed. DP Farrington, R Tarling, pp. 210–30. Albany, NY: SUNY Press

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58(1):267–88

Turing A. 1950. Computing machinery and intelligence. *Mind* 59(237):433–60

Wang B, Yin P, Bertozzi AL, Brantingham PJ, Osher SJ, Xin J. 2017a. Deep learning for real-time crime forecasting and its ternarization. arXiv:1711.08833v1 [cs.LG]

Wang B, Zhang D, Zhang D, Brantingham PJ, Bettozzi AL. 2017b. Deep learning for real time crime forecasting. arXiv:1707.03340v1 [math.NA]

Weisburd D, Groff ER, Yang S-M. 2012. *The Criminology of Place*. Oxford, UK: Oxford Univ. Press

Weisburd D, Mastrofski SD, Greenspan R, Willis JJ. 2004. *The growth of Compstat in American policing*. Rep., Police Found., Washington, DC

Zeng J, Ustun B, Rudin C. 2017. Interpretable classification models for recidivism prediction. *J. R. Stat. Soc. Ser. A* 180(3):689–722

Zhang Y, Cheng T. 2020. Graph deep learning model for network-based predictive hotspot mapping of sparse spatio-temporal events. *Comput. Environ. Urban Syst.* 79:101403

Zhuang Y, Almeida M, Morabito M, Ding W. 2017. *Crime hot spot forecasting: a recurrent model with spatial and temporal information*. Paper presented at the 8th IEEE International Conference on Big Knowledge, Hefei, China

Zucchini W, MacDonald IL, Langrock R. 2016. *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton, FL: CRC Press. 2nd ed.