2018

The Plausible and the Possible: A Bayesian Approach to the Analysis of Reasonable Suspicion

W. David Ball
Santa Clara University, wdball@gmail.com

ARTICLES

THE PLAUSIBLE AND THE POSSIBLE: A BAYESIAN APPROACH TO THE ANALYSIS OF REASONABLE SUSPICION

W. David Ball*

ABSTRACT

The United States Supreme Court uses the wrong approach to analyze reasonable suspicion. The Court asks whether, if criminal activity were afoot, an officer would be likely to see what she saw (e.g. furtive gestures or flight from police). The ultimate question we are interested in, however, is whether the conclusion about criminal activity was reasonable in light of what the officer saw. These two questions are different. Even if it is highly likely that an officer would make a set of observations when criminal activity is afoot, it does not follow that criminal activity is itself highly likely when an officer makes these observations.

In this Article, I propose that a Bayesian approach could improve judicial assessments of reasonable suspicion. Bayesian analysis is designed to deal with conditional probabilities (e.g., how likely is it that criminal activity is afoot given a furtive gesture) and relative likelihoods (of all the possibilities that might explain the furtive gesture, which is the most likely and by how much). Crucially, Bayesian analysis includes all relevant information—especially information about false positives that result from innocent people displaying the same behavior as those engaged in criminal activity. This Article will demonstrate how Bayesian tools might more easily reveal logical gaps in the analysis of reasonable suspicion and help to estimate the "reasonableness" of a given investigation more coherently, consistently, and accurately.

INTRODUCTION

Imagine that state-of-the-art criminological research has generated an extremely accurate test for detecting concealed weapons. Thanks to the research, police now know that if someone is carrying a concealed weapon, they will be able to observe a particular kind of visible, suspicious bulge 99% of the time. Because concealed

* Thanks to Santa Clara Law School for their financial support during my sabbatical, which I devoted to the study of Bayesian statistics. Many thanks to participants in the Santa Clara Faculty Workshop, the online Southwest Criminal Law Workshop, and to my co-panelists at the Katz Symposium for their helpful feedback. I wish to thank Jack Chin, Carissa Hessick, Eric Miller, and David Sloss in particular. All errors remain mine, of course.

weapons are illegal in this jurisdiction, police are anxious to go out into the field and use the new test to zero in on those breaking the law.

Under the United States Supreme Court's current method of analyzing reasonable suspicion, our newfangled test would easily pass Fourth Amendment scrutiny. *Terry v. Ohio* requires officers to provide articulable reasons giving rise to reasonable suspicion in order to justify a brief stop; if officers also suspect that the individual is armed, they may conduct a brief "frisk" of the outer clothing of the individual.[1] When asked to justify a stop, an officer could say that she relied on an observation—the new suspicious bulge test—that, in the presence of criminal activity, is observed 99% of the time. Though this is the answer courts accept, it answers the wrong question. We want the probability that a bulge we observe is a gun, not the probability that if there is a gun, we will observe a bulge. In a world in which we do not start off knowing that someone is engaged in criminal behavior—the world we are supposedly reviewing—we cannot know how a given set of observations relates to criminal activity without also analyzing how often those behaviors are associated with innocent activity.

Imagine one percent of people have guns, and that ten percent of people without guns nevertheless display the suspicious bulge. When an officer observes a suspicious bulge–even under our new, 99% accurate test–the probability that the bulge is a gun is nowhere close to 99% (even though most people, when surveyed, estimate that it is).[2] The actual probability that the bulge is a gun is 9.1%.[3] In other words, guns can be almost perfectly correlated with bulges even when most bulges aren't, in fact, guns.[4]

The Supreme Court's current method of assessing reasonable suspicion, then, is inaccurate and illogical. Even if a bulge is consistent with having a gun 99% of the time, that isn't the only relevant fact. We must, instead, consider the entire range of possibilities. Failing to consider false positives—bulges but no guns—is not

---

1. Terry v. Ohio, 392 U.S. 1, 21, 27 (1968).

2. *Cf.* NATE SILVER, THE SIGNAL AND THE NOISE: WHY SO MANY PREDICTIONS FAIL—BUT SOME DON'T 245, 491 (2012) (citing Dan M. Kahan et al., *The Polarizing Impact of Science Literacy and Numeracy on Perceived Climate Change Risks: Supplementary Information* 4, NATURE CLIMATE CHANGE (2012),https://www.nature.com/nclimate/journal/v2/n10/extref/nclimate1547-s1.pdf) (finding that 3% of people were able to calculate the risk of disease with true positives, false positives, and a prior estimate); JAMES V. STONE, BAYES' RULE: A TUTORIAL INTRODUCTION TO BAYESIAN ANALYTICS 3–4 (2013) (similar example).

3. I perform the calculation *infra* at page 521, but it is simply (.99*.01)/((.99*.01)+(.1*.99)). If your guess was off, don't feel bad. Given a similar setup, most doctors asked to estimate the probability of a disease given a positive test failed to adjust their estimates based on the prior estimate of the disease. Thomas Agoritsas et al., *Does Prevalence Matter to Physicians in Estimating Post-test Probability of Disease? A Randomized Trial*, 26 J. GEN. INTERNAL MED. 373, 376 (2016), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3055966/; *see also* Gerd Gigerenzer et al., *Helping Doctors and Patients Make Sense of Health Statistics*, 8 PSYCHOL. SCI. PUB. INT. 53, 55 (2008) (Using a hypothetical with a 1% base rate, 90% rate of true positives, and 9% false negative rate, the majority of 160 gynecologists surveyed "grossly overestimated the probability of cancer, answering '90%' or '81%.'").

4. Mark Schweizer, *The Law Doesn't Say Much About Base Rates* 13 (Mar. 15, 2013) (unpublished working paper), https://ssrn.com/abstract=2329387 (Confusing the two is "a common mistake known as the 'fallacy of the transposed conditionals,' the 'inverse fallacy' or 'the prosecutor's fallacy'").

only misleading. It also sidelines the crucial privacy interests of the Fourth Amendment: namely, the cost of searching and seizing (or stopping and frisking) those who have done nothing wrong.

Bayesian analysis gives us a framework for evaluating reasonable suspicion that comprehensively accounts for bulges of both the guilty and the innocent. To illustrate how Bayesian analysis calculates probabilities—and to illustrate how, in some circumstances, this is our normal and natural approach—let us turn to the example with which E.T. Jaynes begins his seminal Bayesian work *Probability Theory: The Logic of Science*[5]:

> Suppose some dark night a policeman walks down a street, apparently deserted; but suddenly he hears a burglar alarm, looks across the street, and sees a jewelry store with a broken window. Then a gentleman wearing a mask comes crawling out through the broken window, carrying a bag which turns out to be full of expensive jewelry. The policeman doesn't hesitate at all in deciding that this gentleman is dishonest. But by what reasoning process does he arrive at this conclusion?[6]

The policeman's conclusion strikes us as reasonable, but this conclusion is not based on reason alone. While the behavior is certainly consistent with criminal activity, it is also consistent with non-criminal activity: "there may have been a perfectly innocent explanation for everything."[7] Jaynes suggests that the man might have been the owner of the store and was wearing a mask because he was coming home from a costume party.[8] As he arrived, someone threw a rock through the window.[9] The man realized that his merchandise would be vulnerable to looting, so he gathered it up for safekeeping.[10] He was just leaving the store as the policeman arrived.[11] Most of us would reject this explanation because it is so unlikely—even though it is entirely consistent with the officer's observations. Both explanations are possible, but only one is plausible.

Why does it seem right to conclude that a theft is in progress, even though it is not the only logical possibility? Because, based on our real-world experiences and the background information at our disposal, we know that a theft is more likely than the string of coincidences beginning with a costume party and ending with the store owner happening to arrive at precisely the moment his shop window was smashed. Bayesians call this relevant background information the prior estimate: the probability we would assign to various explanations prior to gathering any

5. E.T. JAYNES, PROBABILITY THEORY: THE LOGIC OF SCIENCE (G. Larry Bretthorst ed. 2003).
6. *Id.* at 1.
7. *Id.*
8. *Id.*
9. *Id.*
10. *See id.*
11. *See id.*

particularized information.[12] The prior estimate is essentially what we would expect to be the case in the normal course of events.

What the jewelry store hypothetical illustrates is that we often have enough background information to choose among competing explanations, even when each explanation fits the data. Our assessment of the best explanation can, of course, be updated as we gather additional information. We might discover a costume party invitation in the man's pocket that leads us to conclude that his explanation is true, however unlikely it might have seemed to start with—but we would be ill-served to pretend that both explanations are equally plausible from the outset. Jaynes concludes that "[w]e are hardly able to get through one waking hour without facing some situation . . . where we do not have enough information to permit deductive reasoning; but still we must decide immediately what to do."[13]

In analyzing police officers' evaluations of reasonable suspicion, then, courts are prone to find them reasonable—but they do so by merely analyzing whether or not such evaluations are logically possible.[14] Courts analyze reasonable suspicion by focusing only on what Bayesian analysis calls the likelihood function—the likelihood that we would observe a given set of behaviors if a theory (say, criminal behavior) were true.[15] In other words, this analysis starts with the officer's conclusion and then determines whether the observed data fits the mold. The complete picture, however, requires us to look at all the information at our disposal. Bayesian analysis combines data (such as an officer's observations), the prior estimate (or base rate),[16] and the likelihood to arrive at the best explanation (the probability),[17] and it allows us to evaluate which explanation is most likely even in the absence of complete information.[18]

In this Article, I argue that courts would be well served to use a Bayesian framework in reasonable suspicion cases. Rather than start with an explanation and evaluate the likelihood of observing behaviors consistent with that explanation, a

---

12. Edwin T. Jaynes, *Prior Probabilities*, 4 IEEE TRANSACTIONS SYS. SCI. & CYBERNETICS, 227, 227–28 (1968).

13. JAYNES, *supra* note 5, at 1.

14. *See, e.g.*, United States v. Arvizu, 534 U.S. 266, 275–76 (2002) (finding reasonable suspicion on the basis of the officer's training and familiarity with the locality, overturning a Ninth Circuit ruling which held that factors capable of innocent explanation deserved little weight). I discuss these matters in greater detail in Part I, *infra*.

15. STONE, *supra* note 2, at 5–7 (defining likelihood).

16. Also known as the "prior probability," *id.*, and "prior information," *id.* at 4.

17. For simplicity's sake, I will use the term "probability" here to refer to what Bayesians call the posterior probability or posterior odds (the recalculated odds that combine the prior estimate and the data). *See, e.g.*, *id.* at 8.

18. I relied on a series of introductory level textbooks for my understanding of Bayesian analysis and would recommend them to lawyers who want to gain some familiarity with the subject. SILVER, *supra* note 2, at 240–61, has an accessible introduction to the subject. Both D.S. SIVIA & J. SKILLING, DATA ANALYSIS: A BAYESIAN TUTORIAL (2d ed. 2006) and STONE, *supra* note 2, provide book-length treatments of the subject at a slightly more advanced level. JOHN K. KRUSCHKE, DOING BAYESIAN DATA ANALYSIS: A TUTORIAL WITH R, JAGS, AND STAN (2d ed. 2015) provides practical instruction in how to compute Bayesian values. WILLIAM M. BOLSTAD, INTRODUCTION TO BAYESIAN STATISTICS (2d ed. 2007) and ANDREW GELMAN ET. AL, BAYESIAN DATA ANALYSIS (3d ed. 2013) provide more advanced treatments.

Bayesian approach would start with the data and ask which explanation is most plausible. The more we knew about a particular case, the less we would rely on our prior estimate, but both kinds of information would always be considered. The approach presented in this Article suggests a framework that can lead to a more coherent—and fair—estimate of reasonable suspicion. Bayesian reasoning has been used to address several legal problems, including identification evidence,[19] the burden of proof in tort cases,[20] the way judges think,[21] drug-sniffing dog alerts,[22] DNA evidence,[23] and probable cause.[24] This Article, however, focuses on the process of Bayesian analysis itself, rather than the results of applying that analysis. The discussion primarily centers on reasonable suspicion, because there is good data on it, but the issues and analysis presented apply equally to the probable cause standard. This Article is not so much designed to break new ground as to bring Bayesian insights to lawyers who know the *Terry* doctrine but who do not know much about the Bayesian method.

This Article proceeds in three parts. First, I will briefly explore the way in which reasonable suspicion cases are currently analyzed: by seeing whether the data are logically consistent with a theory that some form of criminal activity is afoot (what Bayesians call the likelihood function). Second, I will offer a brief primer on Bayesian analysis with an emphasis on how the likelihood function—the logical consistency of the data and criminal activity—can be misleading. Finally, I will explore both the promise of and potential problems with applying Bayesian analysis to questions of racial disparity and order-maintenance policing.

I. LIKELIHOOD ANALYSIS IN REASONABLE SUSPICION CASES

Terry stops carve out a permissible realm of police/citizen interaction that is less intrusive than an arrest and which need be justified only on the basis of reasonable suspicion.[25] The review of reasonable suspicion is objective[26]: an officer

19. Michael O. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83. HARV. L. REV. 489 (1970); *see also* Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971) (critiquing the use of Bayesian reasoning to address legal problems).

20. Edward K. Cheng, *Reconceptualizing the Burden of Proof*, 122 YALE L.J. 1254 (2013).

21. *See, e.g.*, RICHARD A. POSNER, HOW JUDGES THINK 65–67 (2008). *But see* Jack Knight et al., *How Bayesian Are Judges?* 16 NEV. L. J. 1061 (2016) (finding little support for the proposition that judges update their probability estimates based on evidence).

22. Richard E. Myers II, *Detector Dogs and Probable Cause*, 14 GEO. MASON L. REV. 1 (2006).

23. Ian Ayres & Barry Nalebuff, *The Rule of Probabilities: A Practical Approach for Applying Bayes' Rule to the Analysis of DNA Evidence*, 67 STAN. L. REV. 1447 (2015).

24. Max Minzner, *Putting Probability Back into Probable Cause*, 87 TEX. L. REV. 913 (2009).

25. The phrase was not used in the Terry majority itself, though it was used in Justice Douglas's dissent, *see* Terry v. Ohio, 392 U.S. 1, 37 (1968) (Douglas, J., dissenting), and in the *Terry* companion case *Sibron v. New York*, which concerned a New York stop-and-frisk statute. 392 U.S. 40, 72 (1968) (Harlan, J., concurring) ("Under the decision in *Terry* a right to stop may indeed be premised on reasonable suspicion and does not require probable cause, and hence the New York formulation is to that extent constitutional.").

26. *Terry*, 392 U.S. at 21.

must form her suspicions on the basis of specific, articulable facts.[27] A mere hunch is not enough.[28] At the same time, courts interpret facts in light of officer training and experience.[29] A fact which might not read as criminal to the average citizen might nevertheless be validly interpreted as criminal by a trained, experienced officer.[30] In *Terry*, Officer Martin McFadden observed two men, Terry and Chilton, walking in front of a row of stores in Cleveland around 2:30 pm.[31] McFadden initially focused on the two men because they "didn't look right" to him.[32] The two men were Black.[33] McFadden observed each of them walking past the stores five or six times, stopping in front of a particular window, for a total of around twelve times.[34] A third (White) man, Katz, arrived and joined the two men in conversation.[35] Katz then walked off and Terry and Chilton eventually followed.[36] McFadden suspected them of planning a robbery.[37] He stopped and frisked the men, finding guns on both Terry and Chilton.[38] Terry was convicted of carrying a concealed weapon.[39]

In *Terry*, the Court analyzed what Bayesian analysis calls the likelihood function: whether the observed actions "were consistent with McFadden's hypothesis that these men were contemplating a daylight robbery."[40] Courts today continue this approach, beginning with a hypothesis of criminal activity and then evaluating whether a defendant's observed behaviors and traits were consistent with criminal activity or whether they did anything to "negate that hypothesis."[41] Almost any set of facts, however, can be logically consistent with criminal activity. Indeed, an officer can be right that certain observations are consistent with criminal activity and still be wrong that criminal activity is the most likely (or even a reasonably likely) explanation of the observed behavior. But, "[a]t present, as long as officers articulate some objective facts justifying their suspicions, courts typically defer to their

27. *Id.*

28. *Id.* at 27.

29. *See, e.g.*, L. Song Richardson, *Police Efficiency and the Fourth Amendment*, 87 IND. L.J. 1143, 1155–61 (2012); *see also* Eric J. Miller, The Epistemology of Suspicion 62–64 (2011) (unpublished manuscript) (on file with author).

30. Richardson, *supra* note 29.

31. *Terry*, 392 U.S. at 5.

32. *Id.*

33. DETECTIVE MARTIN MCFADDEN, CLEVELAND POLICE DEP'T, REPORT ON THE ARREST OF JOHN WODALL TERRY, RICHARD D. CHILTON, AND CARL KATZ (1963),http://www.acluohio.org//assets/issues/PolicePractices/TerryVOhioPoliceReport1963_1031.pdf [hereinafter MCFADDEN POLICE REPORT].

34. *Terry*, 392 U.S. at 6.

35. MCFADDEN POLICE REPORT, *supra* note 33.

36. *Terry*, 392 U.S. at 6.

37. *Id.*

38. *Id.* at 7.

39. *Id.* at 4.

40. *Id.* at 28.

41. *Id.*

judgments."[42] For example, in *United States v. Arvizu*, an officer observed that a driver was sitting in a rigid, upright posture and that children were waving out of the back window as they passed by the officer.[43] These observations were deemed consistent with drug trafficking,[44] though most parents know that all kinds of bizarre behavior are consistent with a mini-van full of children. In *Ornelas v. United States*, a loose door panel and a rusty screw in a ten-year-old car gave rise to an inference that the car—driven by a man whose name was in a database of heroin dealers—was being used to smuggle drugs, rather than as a sign that such an old car (literally an Oldsmobile) merely needed repairs.[45] In *Illinois v. Wardlow*, a young man running from an unmarked police car speeding through a high-crime area gave rise to an inference of criminal activity, even though there are quite reasonable alternative explanations for running away (say, out of fear of a potential confrontation).[46] The same could be said for many other cases.[47] This is not to say that the reasonable suspicion standard requires that criminal activity be the most likely explanation of what is going on. It doesn't.[48] But even under a relaxed standard of probability, looking only at the likelihood is misleading. *Terry* analysis, which only examines whether the data is consistent with a given theory (the likelihood function), does not tell us all we need to know about whether the theory itself (criminal activity) is a plausible explanation.[49] To do that, we need to examine all the information at our disposal, including how often we observe indicia of criminal activity among the innocent.

II. UNDERSTANDING LIKELIHOOD REQUIRES CONTEXT

Bayesian analysis is often contrasted with frequentist statistics (also known as null hypothesis testing statistics), the dominant mode in college curricula and what many people think of as, simply, "statistics."[50] Bayesian analysis starts with the data and asks how probable it is that a theory is true. In the gun-bulge hypothetical that starts this Article, the question is this: given the bulge (the data), how probable is it that a person has a gun (the theory)? The frequentist, on the other hand,

42. Richardson, *supra* note 29, at 1166.

43. United States v. Arvizu, 534 U.S. 266, 270–71 (2002).

44. *Id.*

45. *See* 517 U.S. 690, 692–93 (1996).

46. *See* 528 U.S. 119, 121 (2000).

47. *See, e.g.*, Alabama v. White, 496 U.S. 325 (1990) (finding that an anonymous tip provided "sufficient indicia of reliability to justify" a traffic stop); United States v. Cortez, 449 U.S. 411, 419 (1981) ("[O]bjective facts, meaningless to the untrained, can be combined with permissible deductions from such facts to form a legitimate basis for suspicion of a particular person and for action on that suspicion.").

48. *See, e.g.*, United States v. Sokolow, 490 U.S. 1, 7 (1989) ("the level of suspicion required for a *Terry* stop is obviously less demanding than that for probable cause . . .").

49. The dissent in *Wardlow* made precisely this criticism. In *Wardlow*, Justice Stevens argued that there were many innocent reasons a pedestrian might flee, and that the Court needed to evaluate the stop based on all the circumstances. 528 U.S. at 128–34 (Stevens, J., dissenting).

50. *See* George W. Cobb, *The Introductory Statistics Course: A Ptolemaic Curriculum?*, TECH. INNOVATIONS STAT. EDUC. 1, 7 (2007); *see also* SILVER, *supra* note 2, at 251–52; KRUSCHKE, *supra* note 18, at 5.

typically asks about the likelihood function—that is, if it is true that someone has a gun (the theory), how likely is it that an officer will spot a bulge (the data)?[51] Bayesians start with data and then fit theories; frequentists start with theories and then fit data. Bayeisans tie probability to confidence in a particular theory, while frequentists view probability in terms of the "chance" that a theory is right.[52]

A detailed exploration of the differences between the two approaches is beyond the scope of this Article, but one key difference—how each views probability itself—sheds light on why Bayesians analyze probability the way they do. As James Stone describes it, frequentists think of probability as something in the world, the long-run frequency of a given result given enough repetitions (thus the name "frequentism").[53] So, to a frequentist, the probability of rolling the number three on a six-sided die is the proportion of times that the number three would show up in an infinite series of rolls.[54] Bayesians, however, think of probability as reflecting a state of knowledge.[55] We don't know what number will be rolled because we cannot accurately measure all the initial conditions of the die throw: its angle, the friction of the table, the speed and height from which it was thrown, etc.[56] A Bayesian assigns a certain probability to a result based on the limitations of the data, but "as the amount of information we have increases, our confidence in the probability of each possible outcome also increases. This suggests that probability is not a property of the physical world but is a measure of how much information an observer has about that world."[57]

---

51. Or, put another way, if the null hypothesis were true (you have a gun) how often would we expect to see a bulge?

52. KRUSCHKE, *supra* note 18, at 297–98; SIVIA & SKILLING, *supra* note 18.

53. SIVIA & SKILLING, *supra* note 18, at 9.

54. STONE, *supra* note 2, at 119.

55. SILVER, *supra* note 2, at 249 ("Bayes's theorem deals with *epistemological* uncertainty—the limits of our knowledge.") (emphasis in original); *see also* SIVIA, *supra* note 18, at 3–4, 9, 11.

56. STONE, *supra* note 2, at 119–20; *see also* KRUSCHKE, *supra* note 18, at 19 (discussing randomness as a function of "[e]xtraneous influences [that] contaminate" the observed data).

57. STONE, *supra* note 2, at 120. For example, many people have difficulty with what is known as the Monty Hall problem. Monty Hall was the host of a famous gameshow in the 1970's called "Make a Deal," and, in part of the show, a contestant would be given the chance to pick a prize from behind one of three curtains. Monty Hall would then reveal one of the unchosen curtains and offer the contestant a chance to either stick with her original choice or choose the other remaining unopened curtain. Marilyn vos Savant published this puzzle in a column in Parade magazine and, when she posted her result suggesting that the odds of switching gave you a two-in-three chance of winning, an outpouring of "mansplaining" resulted, challenging her analysis. Zachary Crockett, *The Time Everyone "Corrected" the World's Smartest Woman*, PRICEONOMICS (Feb. 19, 2015), https://priceonomics.com/the-time-everyone-corrected-the-worlds-smartest/.

The result is easily explained by the limitations of a frequentist approach. To a naïve frequentist, the chance of a prize being behind any given curtain is one in three. This is the long-run frequency. You either got it right the first time or you didn't, and it's either behind the original curtain (one in three) or the remaining curtain (one in three). We look only at the original setup. But, as vos Savant explained, that's not the way to look at it. Consider an alternative setup. Monty Hall has a million curtains, and you pick one. He then reveals 999,998 of the remaining curtains, leaving you with the one you chose and the one he didn't open. Do you still have a one in a million chance? Of course not. It's true that, ex ante, there is only a one in a million chance of the prize being behind any one curtain in particular, but after you have seen where the prize isn't, you know that you either got very lucky or that the only reason Monty didn't reveal the other curtain was that the prize was behind it. In other

Bayesianism's epistemological approach to probability has clear advantages over frequentism in analyzing one-off events. We can easily conceptualize the long-run frequency of repeatable events like dice throws or coin flips, but how can we envision long-run frequencies for unique events, like a police officer's assessment that criminal activity is afoot in a *Terry* stop?[58] What is the frequentist explanation for what we mean when we say, for example, that there is a ten percent chance of rain tomorrow, or of a team winning, or that contraband will be discovered? As the old joke goes, all probability is 50-50: it'll either happen or it won't. So it will rain tomorrow, or the team will win, or contraband will be discovered, or it won't. We don't have long-run frequencies because these events are not repeatable. Probability in the Bayesian sense—one's confidence in the outcome, based on what one knows—is a better conceptual fit for singular events.[59] You might consider probability for singular events as akin to the odds at which you would be willing to make a bet that an event will happen. A ten percent chance of rain means that you would bet on rain at any odds paying better than 1 in 10, because your expected return from that bet would be positive. You might also consider the confidence interval to reflect how precise your estimate is. For a sporting event you might say that a win is the likeliest but a tie is also possible, and allocate these probabilities among the outcomes. Or you might have a continuous range of outcomes ("how much rain will there be, ranging from zero to X inches") over which you put a range of probabilities.[60]

The Bayesian model incorporates all the information in the following way. Bayes's formula solves for conditional probabilities, the probability that something is true if something else is also true.[61] If we observe y, how probable is it that we will also observe x? Probability is given by the letter "p" and the observation or theory being estimated is within parentheses; the conditional probability is given using a vertical line.[62] So one would read p(x|y) as the probability that x is true if y

---

words, the two bets are one in a million and 999,999 in a million. In the case of three curtains, the bets are one in three that you chose right the first time, and two in three that you didn't. So you always switch.

58. For a more detailed treatment of these issues, see, e.g., SIVIA & SKILLING, *supra* note 18, at 10–11 and STONE, *supra* note 2, at 120.

59. *See* STONE, *supra* note 2, at 8–9; *see also* KRUSCHKE, *supra* note 18, at 72–73 (describing probability as a "degree of belief").

60. Laplace, who arguably did the most to popularize Bayes's theorem, for example, used Bayesian tools to estimate the mass of Saturn, putting it in a range consistent with the data and all of his prior information (including the laws of physics). But his range of possibilities was about the limits of his knowledge—the tools he had at his disposal and their accuracy. SIVIA & SKILLING, *supra* note 18, at 9–10. As the data improved, the range of probable masses of Saturn decreased—even though the mass of Saturn itself did not change. Sivia and Skilling describe the frequentist approach as assuming that there is a range of all the possible "true" values of Saturn in an infinite number of universes, and that probability has to do with which world we are in. *Id.* Bayesians like Laplace would say we live in this world—and our uncertainty is about measurement. *See id.* at 10–11. Again, the probability means not chance, but confidence. How much data was guiding you when you made your search?

61. SILVER, *supra* note 2, at 243; STONE, *supra* note 2, at 5.

62. The general formation of this probability is to use the Greek letter theta (θ) to stand for a given hypothesis, and to use the letter d to stand for data, but, in a nod to the math phobia of the average lawyer, I will spell these terms out.

is also true (alternatively, one could say "the probability of x given y"). In our gun-bulge hypothetical, it is the probability of someone having a gun (the theory or explanation) if an officer observes a bulge (the initial observation).[63] Note that this probability is not equivalent to the likelihood function, the focus of the Supreme Court's analysis of reasonable suspicion. The likelihood of observing a bulge if someone has a gun is not the same as the probability that someone with a bulge is carrying a gun. The likelihood of someone with drugs running in a high-crime area is not the same as the probability that someone running in a high-crime area is carrying drugs. The likelihood of observing the data under a given interpretation might be high, but the interpretation's probability given all the information could be low.

In Bayesian analysis, the probability of an explanation[64] depends on all the information you have about it, including your prior estimate (base rate). So, in the context of a *Terry* stop, we can model reasonable suspicion as: (a) the likelihood of observing behaviors if criminal activity is afoot (the current Supreme Court model) multiplied by (b) the prior estimate (base rate) of criminal activity itself divided by (c) the total number of observed behaviors (all bulges, whether from guns or something else). This will tell us the probability—how much confidence we should have in our theory that criminal activity is afoot, given the observation of a bulge. In this way, we look not just at when the bulge correctly indicates criminal behavior, but also when it does not.

The prior estimate—how many people carry guns—is a key difference between frequentism and Bayesianism, and it is where much of our background knowledge comes in.[65] A prior estimate can be based on prior research, such as population incidences of disease.[66] It can be made in varying degrees of strength or be more than simply one number; it can be a range of individual numbers or a function. We could assign various numerical weights to these prior estimates, and we could do so in a way that takes account of our uncertainty.[67] In our bulge hypothetical, we could construct our prior estimate by looking at gun sales, surveys about gun ownership, gun registries, gun recoveries on stops, or some combination of all of these.[68] For total observations, we know that we'll see bulges from those carrying guns (true positives) and bulges from those not carrying guns (false positives).

---

63. We could, of course, also look at, say, the probability that you have drugs if you are running in a high-crime area.

64. As noted *supra* in note 17, Bayesians refer to this as the posterior estimate, posterior odds, or posterior probability, but, for simplicity's sake, I will refer to it as the probability. *See, e.g.*, SILVER, *supra* note 2, at 244.

65. *See id.*

66. STONE, *supra* note 2, at 4–5. Generally, prior estimates are "established by appealing to publicly accessible and reputable previous research." KRUSCHKE, *supra* note 18, at 315.

67. Though the finer points of the analysis are beyond the scope of this Article, base rates can take a variety of forms, including both discrete and continuous functions. *See, e.g.*, STONE, *supra* note 2, at 81–84 (uniform priors); *see also* SIVIA & SKILLING, *supra* note 18, at 17–19.

68. We could also make an estimate across a range of functions, say, an equal likelihood of between 1% and 5%, or a bell-shaped distribution with a median of three percent and a standard deviation of .5%.

Remember, the likelihood only tells how likely one is to see a bulge if there's a gun. Nearly all guns have bulges in our example, but many bulges are not guns.

We combine the likelihood, prior estimate, and total observations to give us our final probability via Bayes's formula[69]:

$$p(theory|\text{data}) = \frac{p(data|theory) * p(theory)}{p(data)}$$

In plain terms, this means that the probability—our confidence in a given theory—is the product of the likelihood of observing the data given a particular theory and the prior estimate of that theory occurring in the world divided by the probability of obtaining that data in the world.[70] We can now demonstrate why, in our bulge hypothetical, a 99% likelihood that a person with a gun has a bulge does not translate into a 99% probability that a person with a bulge has a gun. We want to know the probability of having a gun if an officer sees a bulge—p(gun|bulge). The prior estimate that someone has a gun before we have any additional information, p(gun), is 1%. The likelihood of a bulge given that you have a gun is 99%. We also know that even though 99% of people do not have guns, ten percent of them have bulges. Our initial insight, then, is that false positives—ten percent of 99% of the population—are much more frequent than true positives—99% of one percent of the population. The bulge test for finding guns is accurate but guns are rare, and the rarity is what keeps the final probability low. That is, likelihood has to be contextualized with the rest of our knowledge.

Mathematically, we have:

$$p(gun|\text{bulge}) = \frac{p(bulge|gun) * p(gun)}{p(true\ positive) + p(false\ positive)}$$

Substituting the numerical values gives us

$$p(gun|\text{bulge}) = \frac{.99 * .01}{(.99 * .01) + (.1 * .99)} = .091$$

Under our newfangled test, the odds of having a gun once a bulge is spotted increase dramatically—from one percent (our prior estimate) to more than 9 percent.

---

69. An alternative way to express Bayes's formula is to create an odds ratio and multiply it by the prior estimate. This formula would be $p(theory|\text{data}) = \text{p(theory)} \frac{p(data|theory)}{p(data|not\ the\ theory)}$ with the fraction on the right serving as the likelihood ratio.

70. Again, frequentist analysis often stops at the question of whether or not a given set of data is likely for a given theory. That is, p values in frequentist statistics are just likelihoods—*p(data|theory)*—or the likelihood of observing data given a hypothesis (usually the null hypothesis, which is the hypothesis that there is no change between the starting condition and the test condition). This is a profound and often-repeated misunderstanding about what "p" values are in frequentist statistics: they do not measure the likelihood of a given theory being true, but only the likelihood of observing the data *if* the given theory were true. P values do not measure the truth of a theory. ALEX REINHART, STATISTICS DONE WRONG: THE WOEFULLY COMPLETE GUIDE 8–9, 40–42 (2015). For a more detailed critique of frequentism, see KRUSCHKE, *supra* note 18, at 297–329.

The test was highly relevant. But because the original odds of having a gun were so low, we still have many more false positives than true positives—ten times more, in fact. Thus, even if a given observation (e.g., a bulge) is associated with criminal possession of a weapon 99% of the time, it does not mean that the likeliest explanation for a bulge is a gun. The simple *Terry* analysis is insufficient.

We can, in fact, change the probability that someone has a gun without changing the accuracy of our test (the likelihood). Change the prior estimate of those with guns to ten percent, for example, and the probability that bulges are guns increases to 50%.[71] Prior estimates (base rates) are thus crucial determinants of probability, especially in situations where criminal behavior is rare. In other words, our confidence in the individual stop must be considered in terms of all that we know, including systemic information about other people and other stops. Merely asking about likelihood, as the Supreme Court currently does, is misleading.

Of course, a likelihood of 99% is unrealistically high, and our resulting probability might be even lower in the real world. Police will not accurately spot gun bulges 99% of the time. Data from the New York City stop-and-frisk program suggests that weapon recovery is even lower than our 9% probability (between 1% and 2% for those stops that resulted in a frisk).[72] Our false positive bulges might also, of course, be higher than the 10% figure used in the hypothetical.

Ultimately, looking simply at the likelihood—doing what judges do and asking whether an observation is consistent with criminal activity—tends to distort the odds that a given person is a criminal. We tend to think that if a suspect displays certain characteristics, then stopping him is reasonable. But what we should be asking is, if we observe these characteristics, how likely is it that someone is engaged in criminal behavior? Our current mode of analysis is ill-suited to answer this question.

III. APPLYING BAYESIAN ANALYSIS TO RACIALLY DISPARATE POLICING

The prior Part demonstrated that Bayesian analysis can help criminal justice actors evaluate reasonable suspicion with greater accuracy. Can Bayesian analysis also make policing fairer? Imagine that our suspect in the bulge hypothetical is Black, and that our officer stops Black suspects more often than White ones. Imagine further that stops of Black defendants by the department produce guns less frequently than stops of Whites. Is the stop still reasonable? Can we bring this background information into our analysis, or must we ignore this information unless we have evidence of racial animus? Put another way, if race operates on a system-wide level, should we start our analysis of reasonable suspicion with an assumption of race-neutrality?

---

71. (.99 * .1)/((.99 * .1)+(.1 * .99)) = .5.

72. Floyd v. City of New York, 959 F. Supp. 2d 540, 558, 573 (S.D.N.Y. 2013) ("52% of all stops were followed by a protective frisk for weapons. A weapon was found after 1.5% of these frisks.").

This Part explores the promises and pitfalls of a Bayesian approach to the question of racially disparate policing. Part A considers the relationship between systemic data and the individual case: despite overwhelming evidence of systemic racial disparity in our criminal justice system, it is extremely difficult to prove that any individual stop or frisk is the result of racial animus. Part B examines what might constitute evidence of racial disparity on a systematic level, focusing on the work of Sharad Goel and co-authors as a model of what can be discovered with sufficient data and analytical expertise. Part C discusses the problems with applying Bayesian analysis in an individual case and suggests that we should include a prior estimate of racial disparity into our analysis. Part D concludes with a case study of how Bayesian analysis might prophylactically undercut the logical foundations of the large-scale stop-and-frisk programs that have resulted in so much of the racial disparities observed.

A. Systemic Disparities and the Individual Case

Our knowledge about racial disparities in law enforcement seems to be divided between two separate, unbridgeable realms. We know that people of color (among other disenfranchised groups) are much more likely to suffer adverse criminal justice consequences at rates disproportionate to their share of the population. Just to cite a few examples, Black people are far more likely than White people to be arrested for marijuana possession,[73] to have their probation revoked,[74] and to be pulled over while driving.[75] This data no longer has the power to shock or surprise. Indeed, almost fifty years ago, the *Terry* opinion itself mentioned "[t]he wholesale harassment by certain elements of the police community, of which minority groups, particularly Negroes, frequently complain" even as the Court refused to suppress the guns taken from two "colored" men.[76] This harassment is, arguably, at the core of Fourth Amendment concerns: limiting police intrusion into the lives of the innocent.[77]

Despite evidence of the disparate impacts our system has on racial minorities, analysis of individual cases proceeds as if race-neutrality were the rule. For example, it is extremely difficult to prove that race is driving an officer's decision to

---

73. ACLU, THE WAR ON MARIJUANA IN BLACK AND WHITE (2013), https://www.aclu.org/files/assets/aclu-thewaronmarijuana-rel2.pdf.

74. JESSE JANNETTA ET AL., URBAN INST., EXAMINING RACIAL AND ETHNIC DISPARITIES IN PROBATION REVOCATION: SUMMARY FINDINGS AND IMPLICATIONS FROM A MULTISITE STUDY (2014), http://www.urban.org/sites/default/files/publication/22746/413174-Examining-Racial-and-Ethnic-Disparities-in-Probation-Revocation.PDF.

75. LYNN LANGTON & MATTHEW DUROSE, POLICE BEHAVIOR DURING TRAFFIC AND STREET STOPS, 2011 (rev. 2016), https://www.bjs.gov/content/pub/pdf/pbtss11.pdf; *see also id.* at 1 ("White drivers were both ticketed and searched at lower rates than black and Hispanic drivers.").

76. Terry v. Ohio, 392 U.S. 1, 14 (1968).

77. Among the many examples of Fourth Amendment scholarship to emphasize this is Akhil Reed Amar, *Fourth Amendment First Principles*, 107 HARV. L. REV. 757 (1994). After beginning his Article by describing Fourth Amendment jurisprudence as an "embarrassment," Amar observes that it allows "honest citizens" to be "intruded upon . . . with little or no real remedy." *Id.* at 757–58.

conduct a stop in an individual case. As Goel and co-authors have explained in *Combatting Police Discrimination in the Age of Big Data*, there are doctrinal problems with using courts to address racial disparities.[78] *Washington v. Davis* requires a showing of racial animus,[79] *Whren v. United States* says that the Fourth Amendment does not encompass equal protection concerns,[80] *United States v. Batchelder* establishes that prosecutorial discretion is essentially unbounded,[81] and *United States v. Armstrong* makes selective enforcement claims very difficult to prove.[82] There are also evidentiary problems. Rarely do government officials provide explicit evidence of racial animus,[83] and, even where such evidence is available, there are often alternative potential explanations for the official's behavior.[84] These explanations are not evaluated on the basis of plausibility, but possibility[85]—as long as there is a race-neutral way to explain, say, a suspicion of criminality or the use of force, the analysis comes to an end.[86]

The United States Supreme Court has made it difficult to find equal protection violations due to racial disparities in the enforcement of criminal law.[87] The most prominent case to so find—*Yick Wo v. Hopkins*—did so on facts that were overwhelmingly indicative of racial disparity in enforcement.[88] In *Yick Wo*, all 150 people arrested for violating a San Francisco ordinance concerning laundries in

78. Sharad Goel et al., *Combatting Police Discrimination in the Age of Big Data*, 20 NEW CRIM. L. REV. 181, 202–04 (2017). The analysis on the following pages mirrors theirs.

79. 426 U.S. 229 (1976).

80. 517 U.S. 806 (1996).

81. *See* 442 U.S. 114 (1979).

82. *See* 517 U.S. 456 (1996).

83. For a notable exception, see Foster v. Chatman, 136 S. Ct. 1737, 1744–45 (2016) (finding that prospective African-American jurors had their names circled and highlighted on jury selection materials found in the prosecution's file).

84. The evidence of this phenomenon is most pronounced in the requirement that prosecutors provide race-neutral reasons for striking minority jurors under *Batson v. Kentucky*, 476 U.S. 79, 88–89 (1986). Some prosecutor's offices responded to *Batson* with training to provide race-neutral justifications for striking Black jurors. Gilad Edelman, *Why Is It So Easy for Prosecutors to Strike Black Jurors?*, NEW YORKER (June 5, 2015), http://www.newyorker.com/news/news-desk/why-is-it-so-easy-for-prosecutors-to-strike-black-jurors.

85. I use these terms slightly differently than Kiel Brennan-Marquez does in his article *"Plausible Cause": Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249 (2017). In his usage, plausibility is akin to that which has explanatory value. My usage of plausible means only that it is somewhat probable or believable.

86. This is not to say that there are never legitimate reasons for stopping a member of a racial minority—there are, of course—only that the inquiry is not as searching as it might be. For a more in-depth treatment, see, e.g., Goel et al., *supra* note 78, at 200.

87. *See, e.g.*, David A. Sklansky, *Cocaine, Race, and Equal Protection*, 47 STAN. L. REV. 1283, 1303 (1995) (attributing the difficulty of establishing equal protection violations in crack prosecutions to "the rules the Supreme Court has developed for evaluating equal protection challenges. As the courts of appeals have recognized, those rules can be applied rather mechanically to the federal crack sentences, and all but require affirmance").

88. 118 U.S. 356 (1886). Though the case concerned a municipal ordinance, Yick Wo was arrested and jailed for violating the ordinance. *Id.* at 357. For a contrasting result on similar facts, see Brown v. City of Oneonta, 221 F.3d 329 (2d Cir. 2000) (Police stopping and questioning more than 200 Black men on the basis of a description of the suspect as a young Black male held not to violate the Equal protection clause without "other evidence of discriminatory racial animus.") *Id.* at 333–34.

wooden buildings (a putative fire hazard) were Chinese, while not a single non-Chinese person was arrested.[89] All 200 licensing requests made by those of Chinese descent to continue operation were denied; only one request by a non-Chinese applicant was denied.[90] With such disparities, the Court found that no possible explanation other than racial animus could explain the result.[91] But what of cases without such dramatic disparities? The evidence from *Yick Wo* does not admit of alternative explanations, but what if race only plays some role, not such a complete one?

In the non-extreme cases, where race is not the only possible explanation, we do not currently have useful analytical tools. We don't seem to be able to bridge our two types of knowledge: the system, with its obvious racial disparities, and the individual cases, where we have difficulty accounting for race as a motivating or determinative factor. But, of course, if we have systematic evidence that race matters, it is exceedingly unlikely—in fact, impossible—that we should find no racial effects in any individual case. We cannot get to the systematic effects without the concatenation of individual cases. The following parts suggest how we might begin to apply what we know about the system to the individual case.

B. Modeling Systemic Discrimination in Terry Stops

In a series of articles with co-authors, Sharad Goel has analyzed stop data for evidence of discrimination.[92] This work shows that great insights can come from comprehensive data sets. With systemic data, researchers can look at misses as well as hits, avoiding hindsight bias and giving us an accurate census of all stops, not just ones that resulted in an arrest, the detection of contraband, or the presence of a weapon.[93] In other words, systemic analysis provides the context that is necessary to evaluate probability, as demonstrated in Part II.

Several authors have attempted to construct theoretical models of disparate impact in prior works. Max Minzner and L. Song Richardson have suggested we look at success rates ("hit rates") for individuals or departments in the evaluation of suspicion,[94] meaning the percentage of stops that result in arrests, contraband, or

89. *Yick Wo*, 118 U.S. at 359.

90. *Id.*

91. *Id.* at 374 ("[T]he conclusion cannot be resisted, that no reason for it [the racial disparities] exists except hostility to the race and nationality to which the petitioners belong, and which in the eye of the law is not justified.").

92. Goel et al., *supra* note 78; Sharad Goel et al., *Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy*, 10 ANNALS OF APPLIED STAT. 365 (2016); Camelia Simoiu et al., *The Problem of Infra-Marginality in Outcome Tests for Discrimination*, 11 ANNALS OF APPLIED STAT. 1193 (2017).

93. Compare the Supreme Court's analysis of the universe of results from checkpoint stops, as in City of Indianapolis v. Edmond, 531 U.S. 32 (2000) and Delaware v. Prouse, 440 U.S. 648 (1979). In these cases, the Court not only had a very low standard of what kinds of hit rates count as efficient searches, but it also examined the stated "purpose" of the suspicionless stop. *See Edmond*, 531 U.S. at 37–48; *Prouse*, 440 U.S. at 653–55.

94. Minzner, *supra* note 24, at 920–22; Richardson, *supra* note 29, 1167–71.

weapons seizure. Minzner uses a Bayesian framework,[95] though his analysis differs from mine in that he focuses on the ways in which past results (by officer or department) should guide probable cause evaluations.[96] Richardson uses the hit rate framework in the context of evaluating racial disparities in stop-and-frisk programs.[97] Richardson critiques the "objective facts" framework of reasonable suspicion in light of social science research establishing that all human beings have implicit racial biases.[98] An officer's evaluation of whether a given behavior is suspicious, Richardson argues, is shaped by race.[99] Hit rates that take into account racial effects can illustrate the impact of these differences, and a police department, once made aware of the problem, can work to decrease the effect of racial bias.[100] In both works, a high hit rate is seen as proof that an officer's estimation of suspicion was, in some sense, more reasonable than that of an officer with a lower hit rate. Low hit rates might suggest that other factors not related to criminal activity were driving the decision to stop and/or frisk. Currently, officers suffer no adverse consequences for unsuccessful searches. Minzner suggests that if judicial approval of probable cause depended on hit rates, officers would seek to maximize approval by minimizing unsuccessful searches, resulting in a more socially optimal (and accurate) equilibrium between a level of suspicion and the concomitant likelihood of a search.[101]

One potential problem with hit rate analysis is that there might be heterogeneity or hidden variables within populations that might drive the results.[102] High hit rates do not necessarily mean officers are stopping fewer people than they should. Imagine that a small subset of White people wear t-shirts with the logo of the National Rifle Association (NRA), and that almost all of these people have guns. Imagine further that White people who don't wear NRA t-shirts are unlikely to have guns. Police stop only those with NRA t-shirts, giving them a high hit rate despite a low stop rate, and, in this example, stopping additional White people—those without t-shirts—would decrease the hit rate.[103] Saying that this is theoretically possible, however, does not mean it is the best explanation for differential results in the real world. Of course, our assumption about the non-shirt-wearers would need to be tested—otherwise, we could not be so sure that the failure to stop them was justified—and we could also assume that

95. Minzner, *supra* note 24, at 920–21 nn.32–35.

96. *Id.* at 915.

97. Richardson, *supra* note 29, at 1167.

98. *Id.* at 1145.

99. *Id.* at 1146–51.

100. *Id.* at 1167–71.

101. Minzner, *supra* note 24, at 928.

102. Schweizer, *supra* note 4, at 5 ("Relative frequencies of a property in a reference class should determine the subjective probability that a randomly chosen individual from the reference class has said property only if the reference class is *homogenous* with regard to the property.") (emphasis in original).

103. Simoiu and co-authors discuss this in their article; the technical term for this phenomenon is inframarginality. Simoiu et al., *supra* note 92.

people who wanted to avoid detection would adapt to the profile and stop wearing the shirts. Alternatively, imagine that nervousness is considered a sign of criminal activity, but that "[i]f innocent minorities anticipate being discriminated against, they might display the same behavior—nervousness and evasiveness—as guilty individuals, making it harder to distinguish those who are innocent from those who are guilty."[104]

Hits themselves are also subject to a variety of interpretations. The defendant in *Terry* was suspected of robbery but ultimately convicted of a weapons offense. Is being "right" (getting a conviction) for the wrong reasons (weapons, not robbery) a hit? Would it have been a hit if Terry had possessed illegal drugs, or had a bench warrant for an unpaid traffic ticket? Misses are also not necessarily a sign of poor policing: an officer could also act on the right reasons (a high likelihood) and get unlucky, just as an officer could try a long shot and get a lucky hit.[105] Individual officers are unlikely to have many stops and frisks[106]—and it would be difficult to have an officer-level factor with much predictive power without more officer-level data.[107]

Moving the analysis from officers toward either populations or types of police/citizen interactions poses its own set of problems. What is an equivalent population or an equivalent situation?[108] The more variables we track, the smaller the number of equivalent comparators, until, with enough ability to track the neighborhood, time of day, bulges, behavior, et cetera, every search becomes unique. In some sense, dog sniffs and DNA tests lend themselves more easily to Bayesian analysis than stops do, as demonstrated in research by Richard E. Myers II[109] and Ian Ayres and Barry Nalebuff,[110] respectively. These situations are reasonably standardized, and the results of the tests are either binary (in the dog case, detecting contraband or not) or quantifiable (in the DNA case, how many genetic matches).[111] Perhaps

104. *Id.* at 1198.

105. Miller, *supra* note 29, at 68–69.

106. In 2012, the New York City Police Department estimated that patrol officers made less than one stop a week. RAYMOND W. KELLY, REASONABLE SUSPICION STOPS: PRECINCT BASED COMPARISON BY STOP AND SUSPECT DESCRIPTION (2012), http://www.nyc.gov/html/nypd/downloads/pdf/analysis_and_planning/2012_sqf_final_04_02_2013.pdf. "In 2012 the NYPD conducted about 536,000 stops; with roughly 19,800 officers on patrol, this equates to less than one stop per officer per week." *2012 NYPD Reasonable Suspicion Stops Report*, NYPD (2013),c http://archive.is/k20dn#selection-587.0-587.43.

107. For a more in-depth criticism of the idea that differential hit rates justify differential stop rates, see Sonja B. Starr, *Testing Racial Profiling: Empirical Assessment of Disparate Treatment by Police*, 2016 U. CHI. LEGAL F. 485 (2016).

108. For a more detailed discussion see, e.g., Minzner, *supra* note 24, at 952–55 and Schweizer, *supra* note 4, at 5–6.

109. Myers, *supra* note 22.

110. Ayres & Nalebuff, *supra* note 23.

111. Indeed, DNA evidence, in one scholar's phrasing, is "inherently Bayesian." Christoph Engel, *Neglect the Base Rate: It's the Law!* 3, MAX PLANCK INST. FOR RES. ON COLLECTIVE GOODS (2012), https://www.coll.mpg.de/pdf_dat/2012_23online.pdf.

there are subcategories of stops that might lend themselves to similar kinds of standardization.[112]

The work of Goel and his co-authors is so useful because it operates on data sets that provide comprehensive information about both the breadth of departmental activity and the depth of factors involved in individual stops. In the earliest article, Goel, Rao, and Shroff used data from the New York City stop-and-frisk program, focusing on stops with a suspicion of weapons possession.[113] Because officers in New York City are required to fill out a form (UF-250) stating their reasons for the stop, and because the stop data also includes time, place, and race of the suspect, the authors were able to analyze what factors led to successful stops and how they related to racial impacts.[114] The authors used machine learning to "train" a model on several thousand stops, correlating factors listed on the UF-250 to hit rates.[115] They then tested this model on a remaining portion of the data to see if they could "predict" which stops resulted in the discovery of a weapon, their variable of interest.[116] What they found was that 43% of stops were made in circumstances that would yield a less than 1% chance of finding a weapon and that "blacks and Hispanics were disproportionately involved in low hit rate stops."[117] These racial effects, in turn, were due to two factors: stops in high-crime areas (high-crime areas correlate with poverty, which correlates with race) and discriminatory enforcement.[118] That is, once the researchers controlled for the fact that high-crime areas are home to more Black and Brown people, it was still true that Blacks and Hispanics stopped in these areas were less likely to have weapons than stopped Whites, suggesting "racial discrimination in stop decisions."[119]

A subsequent article by Goel, Perelman, Shroff, and Sklansky built on this work and proposed that these "stop level hit rates" be incorporated into the law.[120] The main obstacle to incorporating this analysis is the legal system's failure to analyze "*Terry* stops . . . as programs, not as isolated occurrences."[121] The authors applied the statistical analysis from the prior article and observed that the "low-odds stops"—those 43% of stops that had less than a one percent chance of discovering a weapon—"had a heavy racial tilt: 49 percent of the stops of blacks fell below the 1 percent probability threshold, as did 34 percent of the stops of Hispanics,

---

112. For one promising example, see Christopher L. Griffin Jr. et al., *Corrections for Racial Disparities in Law Enforcement*, 55 WM. & MARY L. REV. 1365 (2014) (examining case outcomes for intoxicated driving in North Carolina).

113. Goel et al., *supra* note 92, at 365–66.

114. *Id.* at 368.

115. *Id.* at 372.

116. *Id.* at 373–74.

117. *Id.* at 367.

118. *Id.*

119. *Id.*

120. Goel et al., *supra* note 78, 211–20. I might argue that the better term is stop factor hit rates, since they look at the odds of particular combinations of factors—though the authors surely mean that they are looking at the hit rates at the level of an individual stop.

121. *Id.* at 189.

compared with only 19 percent of the stops of whites."[122] The authors concluded that "the reasonableness of a stop-and-frisk under the Fourth Amendment, as well as its compatibility with the Equal Protection Clause of the Fourteenth Amendment, should depend in part on how police departments respond, or fail to respond, to what big data demonstrates about the department's policies and practices."[123]

The most recent of these articles, co-written with Camelia Simoiu and Sam Corbett-Davies, used a data set of 4.5 million traffic stops in North Carolina to analyze the evidence that stops were made on the basis of racial discrimination.[124] The authors used the data set to explore problems with hit-rate tests (what they call "outcome tests"): namely, that different outcomes can be the result of different rates of underlying behavior (risk of contraband, rates of offending, etc.).[125] The authors developed a new metric, the "threshold" test, which examines the conditions under which a stop is made.[126] This measure looks at the relationship between an officer's observations and the decision to stop.[127] Lower thresholds indicate that stops are more likely for a given population and a given set of observations.[128] That is, lower thresholds mean that officers are more willing to stop a member of a given race with fewer indicia of criminal activity. The data for the standard measures of stop rates and hit rates showed that Black and Hispanic drivers in North Carolina were stopped more often than Whites and that hit rates for Whites were higher, suggesting discrimination.[129] The same was true for the overall threshold rate: the threshold of observations that led to a search of Black and Hispanic drivers was lower than that for Whites.[130] The test was of critical importance, however, where there was ambiguous evidence. In one jurisdiction, both stop rates and hit rates were higher for Blacks than Whites, which, without the threshold test, might indicate that higher stop rates were justified because stops were more likely to uncover criminal activity.[131] The authors demonstrated, however, that these hit rates obscured the fact that "blacks still face[d] a lower search threshold (6%) than whites (9%), suggesting discrimination against blacks."[132] In other words, the amount of evidence triggering a search was less for Blacks than Whites, even though the overall higher stop rates would seem, at first glance, to be justified by higher hit rates.

---

122. *Id.* at 188.
123. *Id.* at 190–91.
124. Simoiu et al., *supra* note 92.
125. *Id.* at 1193.
126. *Id.*
127. *Id.* at 1194–95, 1202.
128. *Id.* at 1194–95.
129. *Id.* at 1202–04.
130. *Id.* at 1205–06.
131. *Id.* at 1207.
132. *Id.* at 1207–08.

With data that is both system-wide and stop-level deep, and with skilled analysts, we can get a much better idea of the patterns that emerge. While the approach of Goel and his co-authors is the gold standard, it is not without problems. The first, obviously, is that good data is hard to come by: it takes time and resources for law enforcement to collect, and time and resources to analyze. We need detail about individual stops as well as comprehensive system-wide data. We might also be concerned that officers might game the data, adding additional observations after the fact to make the search more justifiable.[133] And even if we have all the information, the question remains about the individual case: What is the relationship between the system and a particular stop?

C. Assuming Average Racism

Suppose we have evidence of systemic discrimination, whether it is the result of data analysis of the standard described in Part III.B or less detailed information. For example, nation-wide data indicate that African-Americans are 3.73 times more likely than Whites to be arrested for marijuana possession.[134] How, if at all, could we use this data in the analysis of an individual stop? An initial question is whether this disparity is evidence of discrimination or of differential rates of offending.[135] In this instance, however, we know that rates of marijuana usage are generally the same across races.[136] It is still possible, though, that patterns of that usage might be different (use in public, where police might see, versus use in a home, where it is more difficult to detect). Usage might be in locales that are more or less heavily policed, meaning that detection rates are higher even if usage rates are not. Higher arrest rates could also be the product of a higher stop rate (more Black people are stopped when marijuana might be present), a higher frisk rate per stop (more Black people get patted down, leading to the discovery of marijuana), or higher arrest rates for individuals in possession (more White people are released with a warning). We know that states, cities, and counties vary in their Black/White marijuana arrest rates,[137] so national averages are less useful than more localized ones, meaning the strength and shape of the prior estimate would have to

133. *See, e.g.*, Minzner, *supra* note 24, at 937. For a discussion of police perjury generally, including the ways in which narratives can be changed after the fact to increase the level of suspicion, see Christopher Slobogin, *Testilying: Police Perjury and What to do About It*, 67 U. COLO. L. REV. 1037 (1996).

134. ACLU, *supra* note 73, at 4.

135. *But see* Sonja Starr, *Explaining Race Gaps in Policing: Normative and Empirical Challenges* (Univ. of Mich. Law & Econ. Working Papers, Art. 110, 2015), http://repository.law.umich.edu/law_econ_current/110 (discussing problems with the notion that differential rates of offending justify differential treatment).

136. *See id.* at 13–17; *see also* Dylan Matthews, *The Black/White Marijuana Arrest Gap, in Nine Charts*, WASH. POST (June 4, 2013) (summarizing the ACLU report), https://www.washingtonpost.com/news/wonk/wp/2013/06/04/the-blackwhite-marijuana-arrest-gap-in-nine-charts/?utm_term=.67e70580fc86.

137. ACLU, *supra* note 73, at 17–20.

be tailored to an individual location.[138] We might also need to provide different types of analysis for stops where someone was correctly suspected of marijuana possession (a true positive) and stops where someone was falsely suspected of marijuana possession (a false positive). For a false positive, the harm would be a minimal intrusion on privacy (or perhaps some form of stigmatic harm). For a true positive, the harm would come from evidence, and the argument would need to be that the successful stop was based on race, not reasonable suspicion, and that the officer was lucky, not discerning. Remember also that the initial reason for the stop can be different from the ultimate reason for the hit, as in *Terry*, where suspicion of robbery resulted in a weapons charge, so we would need to account for these inadvertent hits in the data as well (where, say, an unrelated traffic stop leads to a marijuana arrest).

The difficulty in establishing this type of claim is evidenced by the litigation challenging New York City's stop-and-frisk program, *Floyd v. City of New York*.[139] The *Floyd* plaintiffs had access to reams of data, but they successfully established an equal protection violation in only one instance, where Cornelio McDonald could show that he was targeted because of race: "The only suspect description was 'black male,' the street was racially stratified, and other non-black individuals were present and presumably behaving no differently than McDonald—yet only McDonald was stopped."[140] McDonald had no contraband on him and was sent on his way.[141]Modeling a marijuana arrest involves a host of possible variables and outcomes. Even Goel and his co-authors' analysis of the New York stop-and-frisk program picked just one offense—criminal possession of a weapon—to focus on, since it has an obvious "correct" answer against which accuracy can be checked: the presence or absence of a gun.[142] The level of confounding variables in the analysis of an individual marijuana arrest, on the other hand, is large, and might depend on a number of factors (smell, movements, and the like). The data itself is impressionistic. We can state with confidence the chemical contents of DNA evidence, for example, and all scientists would agree. We cannot do the same for what constitutes a furtive gesture or the strong smell of marijuana.[143] So, despite the limitations of the hit rate analysis proposed by Minzner and Richardson, perhaps its main advantage is that there is a path to implement it. We could focus on the individual officer and use her record to adjust the suspicion level of the facts presented.

---

138. We can always adjust the prior or allow for some range of values (say from 2 to 5 times more likely, not 3.73 times more likely) in order to account for our uncertainty.

139. 959 F. Supp. 2d 540 (S.D.N.Y. 2013).

140. *Id.* at 633.

141. *Id.* at 631.

142. Goel et al, *supra* note 92, at 366.

143. We also cannot conclude that there is racism on "naked statistics" without running into Laurence Tribe's "blue bus" problem, where mere propensity is incorrectly used as proof. Tribe, *supra* note 19, at 1340–41. For more detail on these issues, see Minzner, *supra* note 24, at 956–58 and Engel, *supra* note 111, at 11–12.

Must we then throw up our hands when it comes to race in individual stops? Bayesian analysis suggests that we incorporate race via a prior estimate. We know that we cannot understand the probability of a successful individual stop without reference to all the information available, including everything we know about other stops. In Bayesian analysis, there is no such thing as the chance of finding a weapon in "just this stop" without estimates of the universe of false positives and true positives that surround it. Even absent evidence of the influence of race in a particular case, then, it would be a mistake to assume that there is no racial influence. We must, instead, assume that there is an average racial influence. When courts disregard race in their prior estimates of being stopped by police—analyzing the facts the officer reports without considering the ways in which race might also have played a role—they are, in effect, saying we have no prior evidence of racial influence on criminal law. That is decidedly not the case. One would be hard pressed to find anyone who would argue with a straight face that White people and people of color face equal chances of being stopped, frisked, arrested, sentenced, or executed. When we fail to assign some value to a racial prior estimate, we are not letting the data talk—we are letting only certain kinds of data talk. If we know racial differences exist throughout the system, we cannot then pretend that they don't exist in any particular case.[144]

Using a prior estimate would incorporate what we know about the system in the analysis of an individual stop, and, as demonstrated in Part II, we cannot understand the individual stop without incorporating prior knowledge. Sometimes the prior evidence of racism would be overwhelming. Sometimes it would be weaker. Sometimes there is a smoking gun one way or another in the actual data observed. Using prior estimates does not mean that all analysis would turn out differently, but it would be foolish to expect every *Terry* stop to spring *ex nihilo* from a place of race neutrality. Our goal should not be to avoid context, but to balance it with the individual situation. This is what we mean by reasonableness, in the main. Bayesian analysis allows us to break down our construction of reasonableness in a more precise way. Even if there aren't mathematical amounts attached to individual variables (and we might assign them using standard tests of "what odds would you accept if you were betting" or the like), using these variables would more accurately model what goes into the final evaluation of how probable a given hypothesis is (as demonstrated in Part II). At the very least, the existence of a prior estimate of racial disparity presents issues that could be addressed via our adversarial system, a point similar to that made by Ayres and Nalebuff in their Bayesian analysis of DNA evidence.[145] We know that likelihoods alone are misleading. If, *ceteris paribus*, race has relevance, then fact finders should attempt to account for it.

---

144. Race could also be a likelihood function. So it would estimate the likelihood of being stopped (or arrested) as a function of race, rather than a prior estimate of being stopped or arrested due to race.

145. Ayres & Nalebuff, *supra* note 23, at 1450.

In a 2007 article,[146] Bernard E. Harcourt and Jens Ludwig suggested how we might incorporate race doctrinally: via a burden-shifting model akin to that used in *Batson v. Kentucky*.[147] After a defendant makes a prima facie case that police have engaged in discriminatory conduct "across several layers of outcomes," they argue, the defendant need not prove actual discriminatory intent by the individual officer, but may, instead, infer discriminatory intent if the government does not provide credible explanations or justifications.[148] This would shift the burden to "the party with the most complete information": the police department.[149] We would therefore start, appropriately, in a world in which race played a role. Additional evidence about the case could certainly change our analysis, and, with additional evidence, the data would dominate the prior estimate—especially if the prior estimate were tenuous.[150]

We could easily imagine scenarios where more data would change the explanatory value of race. If, say, witnesses reported that a young Black man was arrested outside a concert for Phish, a band popular among young White male pot smokers, the probability of race being involved would increase—there were lots of White targets and police chose the Black one. If, in another case, evidence showed that a defendant was wearing a Bob Marley shirt standing amidst a sea of abstemious churchgoers, race would be a less probable explanation for his being targeted. In neither of these two examples would the prior estimate of race-based police conduct be dispositive. As we gathered more information about a stop, our confidence in our probability analysis would increase. The data would inform us more than our prior estimates. But prior estimates based on race would give our analysis a starting point. In our bulge hypothetical, without knowledge of the prior estimate, we would have thought we had a much greater chance of finding a gun if we saw a bulge. We would have been misled by small amounts of data, like a single observation of a suspect by a single officer in a single stop. Having a racial prior estimate puts a thumb on the scale where disconfirming racial effects requires data—but only because we already have data about racial effects. Naturally, we would want to make sure that the studies which inform our prior estimate are sound. To

146. Bernard E. Harcourt & Jens Ludwig, *Reefer Madness: Broken Windows Policing and Misdemeanor Marijuana Arrests in New York City, 1989-2000*, 6 J. CRIMINOLOGY & PUB. POL'Y 165, 167 (2007).

147. Batson v. Kentucky, 476 U.S. 79 (1986), addresses discriminatory patterns of peremptory dismissals of jurors. Once a discriminatory pattern of dismissals is observed, the dismissing party bears the burden of providing race-neutral explanations for the dismissals. *See id.* at 93–96. The Supreme Court most recently addressed *Batson* in Foster v. Chatman 136 S. Ct. 1737 (2016) (holding that the prosecution's race neutral explanations were insufficient in light of notes that, inter alia, highlighted all prospective Black jurors).

148. Harcourt & Ludwig, *supra* note 146.

149. *Id.*

150. SILVER, *supra* note 2, at 247, 259–60; SIVIA & SKILLING, *supra* note 18, at 19. *But see* SILVER, *supra* note 2, at 245 ("When our priors are strong, they can be surprisingly resilient in the face of new evidence.").

ignore this prior information, though, is to sacrifice accuracy, and it does so in a way that credits an individual story—of race-neutrality—that we know to be false in the aggregate.

D. Pushing Back on Programs Themselves

As demonstrated in Part C, applying Bayesian analysis to individual cases presents myriad initial theoretical obstacles even before the doctrinal obstacles are addressed. One response might be to evaluate police programs on an ongoing, systemic basis and provide feedback, as Goel and his co-authors have done, and as Minzner and Richardson have suggested.[151] Large-scale stop-and-frisk programs might make political sense but turn out to impose costs (on certain demographics) without any clear benefits. The most useful role that Bayesian analysis might perform, then, is in counteracting the case for these programs before they are implemented.

Stop-and-frisk programs are generally part of order maintenance policing. Bayesian analysis can be used to analyze the efficiency of order maintenance policing by constructing a probabilistic model. Note that there are a number of policing theories that are often lumped together to justify stop-and-frisk programs, including "Broken Windows theory" and "quality of life policing," two theories that the NYPD Office of the Inspector General points out "are not synonymous."[152] The following analysis does not address one of the causal theories of Broken Windows: that signs of disorder signal lawbreakers that they are unlikely to be apprehended, encouraging them to commit more crimes.[153] It will instead address the theory that serious offenders also commit minor crimes, and that focusing on minor crimes is an effective way to apprehend those serious offenders.[154] Under what conditions would it make sense to focus

---

151. Minzner, *supra* note 24, at 939–43; Richardson, *supra* note 29, at 1167–71.

152. MARK G. PETERS & PHILIP K. EURE, N.Y.C. DEP'T OF INVESTIGATIONS, OFFICE OF THE INSPECTOR GEN., AN ANALYSIS OF QUALITY-OF-LIFE SUMMONSES, QUALITY-OF-LIFE MISDEMEANOR ARRESTS, AND FELONY CRIME IN NEW YORK CITY, 2010-2015 10 (2016). But in his book, former New York City Police Commissioner, Howard Safir, equated the two. HOWARD SAFIR & ELLIS WHITMAN, SECURITY: POLICING YOUR HOMELAND, YOUR STATE, YOUR CITY xi–xix (2003).

153. *See* George L. Kelling & James Q. Wilson, *Broken Windows: The Police and Neighborhood Safety*, ATLANTIC, Mar. 1982, https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/. *But see* Bernard E. Harcourt & Jens Ludwig, *Broken Windows: New Evidence from New York City and a Five-City Social Experiment*, 73 U. CHI. L. REV. 271, 277 (2006) (finding "no support for the idea that broken windows enforcement activities, including order-maintenance policing or other measures designed to reduce the level of social or physical disorder within a community, represent the optimal use of scarce government resources").

154. As one of the authors of the Broken Windows theory, George L. Kelling, put it: "Not all fare beaters were criminals, but a lot of criminals were fare beaters. It turns out serious criminals are pretty busy. They commit minor offenses as well as major offenses." Shankar Vendantam et. al., *How a Theory of Crime and Policing Was Born, and Went Terribly Wrong*, NPR (Nov. 1, 2016, 12:00 AM),

http://www.npr.org/2016/11/01/500104506/broken-windows-policing-and-the-origins-of-stop-and-frisk-and-how-it-went-wrong. Harcourt and Ludwig, however, found no support for this mechanism—indeed, they "found that, if anything, increases in misdemeanor arrests were accompanied by *increases* in violent crime." Harcourt & Ludwig, *supra* note 146, at 173 (emphasis in original).

on minor crimes (say, subway turnstile jumping) as a means of catching more serious offenders (say, robbers)?

In order for this approach to be sound, we would need to establish the conditions under which the probability that a turnstile jumper is also a robber (written p(robber|jumper)) is maximized. Using Bayes's formula, we know that this probability depends on the likelihood that a robber jumps turnstiles (p(jumper| robber)), our prior estimate of how many robbers there are (p(robber)), and our total observations of turnstile jumpers, whether they are robbers or not (p(jumper)).

$$p(robber|\text{jumper}) = \frac{p(jumper|robber) * p(robber)}{p(jumper)}$$

We want to maximize the probability on the left, to make sure that our focus on turnstile jumpers yields the greatest probability of catching robbers, and we can do so by increasing the numerator and/or decreasing the denominator. That leaves us with three non-exclusive conditions that will increase our efficiency: a high likelihood of turnstile jumping by robbers, a high prior estimate of robbers, or a low total number of turnstile jumpers.

All other things being equal, going after small-time offenders in hopes of landing more serious offenders would be efficient only if there are lower overall rates of turnstile jumping and high rates of jumping among robbers—that is, if turnstile jumping were almost exclusively related to robbing. As the rate of turnstile jumping in the general population increases (p(jumper) increases), any individual who is apprehended is less likely to be a robber. At the extreme, if everyone jumped turnstiles, our focus on turnstile jumping would be no more efficient than a dragnet of the entire population. It's only if the overall rate of petty offenders (turnstile jumpers) is low that it makes sense to focus on them—and then only if people who rob are likely to jump turnstiles and don't, instead, try to avoid the attention of law enforcement. What Bayes's equation shows is that, essentially, the theory behind this kind of policing is a tautology: if only robbers jump turnstiles, you can catch robbers by catching turnstile jumpers—because they are essentially the same set of people.

As in Part II, putting the contextual variables together reveals the shortcomings of a standard model. The model's mistake, here, is to assume that even if most robbers are turnstile jumpers (likelihood), most turnstile jumpers must also be robbers (probability). Instead, the best way to solve serious crime—at least logically—is to focus on serious crime. There are not many returns from casting the net wide, as in stop-and-frisk, unless the behaviors are unique to a set of more serious offenders. Generalized suspicion is expensive and inefficient and has a host of other collateral costs. It only makes sense to crack down on turnstile jumping if that is your goal. If you want to

catch more robbers, investigate robberies—or find some other activity that only robbers engage in.[155]

CONCLUSION

This Article has endeavored to show how Bayesian analysis can aid the legal profession in its approach to reasonable suspicion problems. The current method of analyzing reasonable suspicion is problematic; it is misleading simply to look at whether individual behaviors conform to a theory that criminal justice is afoot. An accurate assessment requires us to look at alternative explanations and base rates of offending. Bayesian approaches provide logical and mathematical tools for understanding how individual cases cannot be accurately understood without reference to the system as a whole. I have argued that, in systems where there is evidence of racial disparity, we should assume that there is average racism in the individual case, but this approach is far from ready to be implemented. Even if this approach has logical appeal, formidable doctrinal obstacles remain, including whether systemic evidence of bias would be dismissed as mere propensity evidence (or included as evidence of a common plan) and under what circumstances evidence of disparate treatment could lead to an inference of animus. The larger point about the importance of context and the benefits of Bayesian analysis itself, however, remain.

This Article has focused on reasonable suspicion because there is an abundance of data about large-scale stop-and-frisk programs, but Bayesian approaches are equally well-suited to the analysis of probable cause justifying the issuance of a warrant or the execution of a warrantless arrest. In any instance where we need to evaluate the strength of our belief in a given explanation, we should structure our thinking in Bayesian ways. We want to know more than just whether a given theory is possible; we want to know which theory is the most plausible. To do that, we have to consider all of the information at our disposal: our background knowledge about the world, whether the data we have collected is consistent with a particular explanation, and all the other explanations which might explain the data. Simply looking for data that confirms our hypotheses has too much potential to mislead.

---

155. It could also be argued that a focus on petty offenses as a means of catching serious offenders might still be the best approach, even if it is inefficient, if there are no alternatives. After all, if we want to catch robbers, the probability of finding a robber among turnstile jumpers need only be greater than the probability of catching robbers using some other program. Our concern is not with absolute probability, but the relative strength of one approach versus another. This approach may, in the abstract, be true, but in practice, the opportunity cost of cracking down on petty offenses might not be justified. Consider that police departments across the country have allowed literally thousands of untested rape kits to pile up. Clare Sestanovich, *Untested Rape Kits: FAQ*, MARSHALL PROJECT (Feb. 27, 2015), https://www.themarshallproject.org/2015/02/27/untested-rape-kits-faq#.zfqVFZahH. Yes, we might catch a few rapists who jump turnstiles, but we would be much better served to test rape kits. Ironically, it was the case of John Royster, a serial rapist initially arrested for turnstile jumping, who provided a key talking point in favor of zero-tolerance policing. Jennifer R. Wynn, *Can Zero Tolerance Last? Voices from Inside the Precinct*, *in* ZERO TOLERANCE: QUALITY OF LIFE AND THE NEW POLICE BRUTALITY IN NEW YORK CITY 107, 114 (Andrea McArdle & Tanya Erzen, eds., 2001). *But see* SAFIR & WHITMAN, *supra* note 152, at xii ("[I]f you pay attention to small crimes, you will have a corresponding impact on more serious crime.").