CRITIQUE OF PURE RISK ASSESSMENT OR, KANT MEETS *TARASOFF*

*Douglas Mossman, M.D.**

ABSTRACT

This Article takes a "critical" approach to the assumptions underlying current practices of violence risk assessment. The Article first explicates a fundamental difference between how mental health researchers now interpret the phrase "violence prediction" and how they understood that phrase in the 1970s. Applying a "critical" approach, the Article then shows that courts and mental health professionals may need to abandon the hope that more accurate methods of predicting violence will help clinicians make better decisions about potential violence. Although this result may seem disappointing, it can liberate mental health professionals from regarding patients as statistical sources of risk and can encourage them to treat patients instead as sources of initiative and moral worth. The Article then provides a moral and legal framework for reformulating the Tarasoff *rule. Taking the view that consequentialist, predictive approaches to formulating the duty to protect are ethically questionable (as well as scientifically impractical), the Article suggests that assigning therapists a duty to intervene in response to a patient's explicit, credible threat is consistent with the view that patients (along with other humans) ought to be treated not merely as means, but always as ends in themselves.*

* Professor and Director, Division of Forensic Psychiatry, Wright State University Boonshoft School of Medicine; Administrative Director, Glenn M. Weaver Institute of Psychiatry and Law, University of Cincinnati College of Law. B.A. 1976, Oberlin College; M.D. 1981, University of Michigan Medical School. The author thanks Michael L. Perlin, Kathleen J. Hart, Glenn M. Weaver, Victor Knapp, and Jerald Kay for their advice and support.

I. INTRODUCTION

A. The Importance of Tarasoff

On March 17, 2006, the Glenn M. Weaver Institute of Law and Psychiatry of the University of Cincinnati College of Law marked the three decades since issuance of the *Tarasoff* decision[1] with a symposium entitled, "The Future of the 'Duty to Protect': Scientific and Legal Perspectives on *Tarasoff*'s Thirtieth Anniversary." I had the honor of being among the speakers who gave presentations to the symposium's attendees, and this Article, along with others that appear in this issue of the *University of Cincinnati Law Review*, expands upon my oral remarks.

Because I am an academic psychiatrist who devotes most of his professional attention to forensic issues, it seems "just natural" to me to take special note of any "significant" anniversary of *Tarasoff*. Yet I recognize that persons from other disciplines or with other perspectives—that is, most readers of this Article—may not share this sentiment. For them, it may seem "just natural" to ask, "Why should an Institute of Law and Psychiatry devote an entire symposium to the thirtieth anniversary of *Tarasoff*?" At least two good reasons suggest themselves.

1. *Tarasoff*'s Impact

First, thirty years after its promulgation, *Tarasoff* remains, to mental health professionals, the most influential ruling in mental disability law.[2] One might argue that the *Jackson*[3]*-O'Connor*[4] line of cases and the

1. I refer specifically to the second *Tarasoff* decision, issued on July 1, 1976. Tarasoff v. Regents of Univ. of Cal. 551 P.2d 334 (Cal. 1976). The California Supreme Court issued its first *Tarasoff* decision on December 23, 1974. Tarasoff v. Regents of Univ. of Cal. 529 P.2d 553 (Cal. 1974). In general, when this article refers to the *Tarasoff* decision, rule, or doctrine (without qualification), it is the second and ultimate decision that is being designated. When it is necessary to distinguish between the two, the Article follows the custom of calling the 1974 decision "*Tarasoff I*" and the 1976 decision "*Tarasoff II*."

2. "*Tarasoff* . . . represents one of the most significant developments in medico-legal jurisprudence of the past century. . . . The original case . . . in 1974 burst like a bomb over the clinical scene." Thomas G. Gutheil, *Moral Justification for* Tarasoff*-Type Warnings and Breach of Confidentiality: A Clinician's Perspective*, 19 BEHAV. SCI. & L. 345, 345 (2001).

3. Jackson v. Indiana, 406 U.S. 715 (1972) (subjecting petitioner to a more lenient commitment standard and to a more stringent standard of release than those generally applicable to other persons not charged with offenses violated equal protection clause; indefinite commitment of a criminal defendant solely because of his incompetence to stand trial violated due process).

4. O'Connor v. Donaldson, 422 U.S. 563 (1975) (a state may not constitutionally confine,

Ford[5]*-Penry*[6]*-Atkins*[7] line of cases are more significant in that they address the liberty or life-and-death concerns of persons with mental disabilities, and therefore must be far more important than a tort issue that involves mere monetary damages.[8] Without minimizing the significance of these or other cases, however, *Tarasoff* outranks them in several respects, at least as far as mental health professionals are concerned. Although some of the death penalty cases generate more legal citations or "hits" on Google,[9] queries of the PsycInfo and Web of Science[10] databases show that articles mentioning or citing *Tarasoff* vastly outnumber articles that refer to the other cases,[11] providing vivid testimony to the unique impact of *Tarasoff* on law-and-mental-health scholarship.

---

without more, a non-dangerous individual who could survive safely in the community by himself or with assistance from willing and responsible family members or friends).

5. Ford v. Wainwright, 477 U.S. 399 (1986) (Eighth Amendment prohibits the State from inflicting the death penalty upon a prisoner who is "insane").

6. Penry v. Lynaugh, 492 U.S. 302 (1989) (jury must, upon request, receive instructions that allow them to give effect to mental retardation as mitigating evidence in determining whether to impose the death penalty, but the Eighth Amendment does not categorically prohibit the execution of mentally retarded capital murderers).

7. Atkins v. Virginia, 536 U.S. 304 (2002) (executing mentally retarded criminals constitutes "cruel and unusual punishment" and is prohibited by the Eighth Amendment).

8. For this observation, I am indebted to Professor Perlin. He elaborates a bit on this comment in Michael L. Perlin, *"You Got No Secrets to Conceal": Considering the Application of the* Tarasoff *Doctrine Abroad*, 75 U. CIN. L. REV. 611 (2006).

9. Here are the results of my February 22, 2006 web searches and "unrestricted" Shepardizations using the Lexis database:

Case (search phrase)	Google Hits	Citations
Tarasoff v. Regents	18,500	1,604
Jackson v. Regents	12,800	1,381
O'Connor v. Donaldson	14,800	1,438
Ford v. Wainwright	28,400	1,123
Penry v. Lynaugh	41,400	2,283
Atkins v. Virginia	62,000	1,200

10. PsycINFO covers titles, authors, and abstracts of worldwide professional and academic publications in psychology and related publications in medicine, psychiatry, nursing, sociology, education, pharmacology, physiology, and linguistics. The database used for this search covers the years 1967 to early 2006. The Web of Science indexes publications from 5,900 major journals in more than 150 scientific disciplines (including mental-health-related fields such as psychiatry and neuroscience, and frequently, law review articles), with all cited references captured. The database used for this search covered the years 1980 to early 2006.

11. Results of my February 22, 2006 searches of PsycInfo and Web of Science databases:

Case (search phrase)	PsycInfo	Web of Science
Tarasoff	569	146
Jackson v. Indiana	3	2
O'Connor v. Donaldson	50	2
Ford v. Wainwright	36	5
Penry v. Lynaugh	34	4
Atkins v. Virginia	58	9

Moreover, no court ruling has had a broader or more enduring impact on day-to-day mental health practice. A minority of mental health professionals deals with individuals who face civil commitment, and very few clinicians evaluate or treat individuals potentially subject to capital punishment. But rare is the practicing mental health professional who is not acquainted with the *Tarasoff* decision and the discomfort that arises in clinical situations that trigger a duty to protect.[12] Because of the *Tarasoff* decision in California and its legal "progeny"[13] in other U.S. jurisdictions, mental health clinicians across the country regularly break confidentiality and take other actions to prevent patients from harming members of the public.

The influence of *Tarasoff* in mental health practice goes beyond therapists' mere knowledge of the case to permeate their current views of what mental health clinicians ought to do as part of their everyday practice.[14] Psychotherapists accept the fact that, while they may regard themselves as the care-givers of individual patients, they sometimes must function as agents for social protection.[15] Mental health clinicians

12. For example: I have been teaching psychiatry residents for nearly two decades. In a talk that I give each year to first-year residents, I ask how many of them have heard of *Tarasoff*. I have encountered just one resident (out of the more-than-200 whom I've asked) who had not already heard of the case. In 1996, Professor Monahan observed that "*Tarasoff* . . . has become a familiar part of the clinical landscape." John Monahan, *Violence Prediction: The Past Twenty and the Next Twenty Years*, 23 CRIM. JUST. BEHAV. 107, 110 (1996) [hereinafter Monahan, *Twenty Years*]. *See also* Samuel Knapp & Leon VandeCreek, *Real-life Vignettes Involving the Duty to Protect*, 1 J. PSYCHOTHERAPY INDEP. PRAC., 83, 83 (2000) (discussing "the widespread acceptance of the duty to protect . . . and . . . three common dilemmas faced by psychotherapists" related to the duty).

13. Although it is common now to encounter references to the "*Tarasoff* progeny," the first publication that appears to have done so is ALAN A. STONE, LAW, PSYCHIATRY, AND MORALITY: ESSAYS AND ANALYSIS 161 (Am. Psychiatric Press 1984).

14. "[N]o court decision in the last generation has succeeded in so raising the anxieties of mental health professionals. The ill-defined nature of the duty to protect has led to great confusion about clinicians' obligations." Paul S. Appelbaum et al., *Statutory Approaches to Limiting Psychiatrists' Liability for Their Patients' Violent Acts*, 146 AM. J. PSYCHIATRY 821, 821 (1989). Professor Monahan observes, "The duty to protect, in short, is now a fact of professional life for nearly all American clinicians . . ." Monahan, *Twenty Years*, *supra* note 12, at 111.

15. A few courts have rejected *Tarasoff*. *See, e.g.*, Thapar v. Zezulka, 994 S.W.2d 635 (Tex. 1999) (no duty to warn where Texas Health and Safety Code mandates confidentiality); Gregory v. Kilbride, 565 S.E.2d 685 (N.C. Ct. App. 2002) (North Carolina does not recognize a psychiatrist's duty to warn third persons); Boynton v. Burglass, 590 So.2d 446 (Fla. Dist. Ct. App. 1991) (psychiatrist who has no right or ability to control a voluntary outpatient's behavior can not be held liable for failure to warn the patient's victim).

Nonetheless, mental health professionals see *Tarasoff* as establishing a national standard. *See, e.g.,* JAMES BECK, *The Psychotherapist and the Violent Patient*, *in* THE POTENTIALLY VIOLENT PATIENT AND THE TARASOFF DECISION IN PSYCHIATRIC PRACTICE 9, 33 (James Beck ed., 1985); Alan Felthous, *Duty to Warn or Protect: Current Status for Psychiatrists*, 21 PSYCHIATRIC ANNALS 591 (1991).

See also American Psychiatric Association, THE PRINCIPLES OF MEDICAL ETHICS WITH ANNOTATIONS ESPECIALLY APPLICABLE TO PSYCHIATRY, Section 4, No. 8, http://www.psych.org/

do not explicitly think about stopping violence during every treatment encounter, but the implicit obligation to protect the public is present in every clinical contact[16]—something Justice Clark noted in his dissent in *Tarasoff*.[17] Not surprisingly, therefore, seminars on how to manage potential *Tarasoff* liability remain among the most popular and best attended of the continuing education offerings available to mental health professionals.[18] Books dealing with liability prevention and violence prediction frequently receive awards from forensic mental health professionals' organizations.[19]

The *Tarasoff* doctrine's influence on legal decision-makers consists of more than just numbers of citations. It also manifests itself in how courts view the social role of mental health professionals. The notion that public protection is the *raison d'être* of patients' interactions with psychiatrists, psychologists, and other mental health clinicians is vividly exemplified by the Ohio Supreme Court's 1997 statement that "the relationship between the psychotherapist and the patient in the outpatient setting constitutes a special relation justifying the imposition of a duty

psych_pract/ethics/ppaethics.cfm (last visited Jan. 8, 2006): "When, in the clinical judgment of the treating psychiatrist, the risk of danger is deemed to be significant, the psychiatrist may reveal confidential information disclosed by the patient."; *and* American Psychological Association, ETHICAL PRINCIPLES OF PSYCHOLOGISTS AND CODE OF CONDUCT, Section 4.05(b), http://www.apa.org/ethics/code2002.html (last visited Jan. 8, 2006): "Psychologists disclose confidential information without the consent of the individual only as mandated by law, or where permitted by law for a valid purpose such as to . . . protect the client/patient, psychologist, or others from harm . . .".

16. "Medical students and psychiatric residents are commonly taught to assess for 'homicidal ideation' as a part of a psychiatric work-up . . . ." David M. Gellerman & Robert Suddath, *Violent Fantasy, Dangerousness, and the Duty to Warn and Protect*, 33 J. AM. ACAD. PSYCHIATRY L. 484, 484 (2005).

In psychiatric hospital charts that I read in my clinical work, I commonly encounter daily progress notes in which medical students or residents write, "o S/I, H/I, A/H, V/H," physician short-hand for "no suicidal ideation, homicidal ideation, auditory hallucinations, or visual hallucinations." The last two entries usually are relevant entries for a previously psychotic patient, but the first two entries reflect efforts to avert liability, should the patient harm himself or herself or someone else.

17. "Now, confronted by the majority's new duty, the psychiatrist must instantaneously calculate potential violence from each patient on each visit." *Tarasoff II*, 551 P.2d 334, 361 (Cal. 1976) (Clark, J., dissenting).

18. In the week before the March 17, 2006, *Tarasoff* symposium, I received two brochures about (among other topics) "risk assessment of the mentally ill individual" in my postal mail. For additional examples, *see* http://www.specializedtraining.com/seminars.htm (last visited Mar. 15, 2006).

19. Examples include the following winners of the Manfred S. Guttmacher Award, cosponsored by the American Psychiatric Association and the American Academy of Psychiatry and the Law: JOHN MONAHAN, THE CLINICAL PREDICTION OF VIOLENT BEHAVIOR (1981); JAMES C. BECK, CONFIDENTIALITY VERSUS THE DUTY TO PROTECT: FORESEEABLE HARM IN THE PRACTICE OF PSYCHIATRY (1990); JOHN MONAHAN ET AL., RETHINKING RISK ASSESSMENT: THE MACARTHUR STUDY OF MENTAL DISORDER AND VIOLENCE (2001). A full listing of Guttmacher Awardees appears at http://www.psych.org/public_info/libr_publ/guttmacher.cfm (last visited Feb. 25, 2006).

upon the psychotherapist to protect against and/or control the patient's violent propensities."[20] Despite the traditional common law position that exempts individuals from liability for the easily preventable acts of others,[21] courts have expanded the *Tarasoff* doctrine in many directions, to include protective obligations related from risks posed by mentally disabled drivers[22] and by medical patients who may transmit HIV[23] or other infectious diseases.[24]

2. Scientific Advances

A second reason to mark the thirtieth anniversary of *Tarasoff* is that this decision remains the inspiration for ongoing scholarship in the areas of violence prevention, risk assessment, and communication about risks. The result is that we have much better knowledge than we possessed in the 1970s concerning these subjects and other areas of scientific study that relate directly to the core concerns of *Tarasoff*.[25] From the vantage

20. Morgan v. Fairfield Family Counseling Ctr., 673 N.E.2d 1311, 1327 (1997). Five years before *Morgan*'s blunt statement, Professor Monahan had noted:

> Throughout history and in all known societies, people have believed that mental disorder and violence were somehow related. . . . [T]here can be little doubt that this assumption has played an animating role in the prominence of *dangerous to others* as a criterion for civil commitment and the commitment of persons acquitted of crime by reason of insanity, in the creation of special statutes for the extended detention of mentally disordered prisoners, and in the imposition of tort liability on psychologists and psychiatrists who fail to anticipate the violence of their patients.

John Monahan, *Mental Disorder and Violent Behavior: Perceptions and Evidence*, 47 AM. PSYCHOLOGIST 511, 511 (1992).

21. "This is true although the actor realizes that he has the ability to control the conduct of a third person, and could do so with only the most trivial of efforts and without any inconvenience to himself." RESTATEMENT (SECOND) OF TORTS § 315 (1965).

22. Schuster v. Altenberg, 424 N.W.2d 159 (Wis. 1988) (psychiatrist could be liable for damages resulting from his patient's auto accident).

23. Reisner v. Regents of the Univ. of Cal., 37 Cal. Rptr. 2d 518 (Cal. Ct. App. 1995) (failure to inform patient led to actionable transmission of HIV).

24. Bradshaw v. Daniel, 854 S.W.2d 865 (Tenn. 1993) (failure to inform concerning diagnosis and symptoms of Rocky Mountain Spotted Fever could be basis for wrongful death action).

25. Writing at the twentieth anniversary of *Tarasoff*, Professor Monahan observed:

> Twenty years ago, American law asked very different questions about violence prediction than it does today. The methodologies by which social scientists went about answering those questions were unlike those in current use, and the conclusions drawn by the researchers were not the same as those drawn now.

Monahan, *Twenty Years, supra* note 12, at 107. Before *Tarasoff*, concerns about the constitutionality of detaining mentally ill persons framed both the legal questions asked about violence prediction and the way mental health professionals thought about detention. After *Tarasoff*, "[l]iability, rather than constitutionality" became "the concern that motivate[d] interest in the prediction of violence . . . ." *Id*. at 111.

point of the past thirty years of research on violence prediction, we can ask whether that knowledge should reshape our views about therapists' duties in fact situations similar to those that faced the therapists who treated Tatiana Tarasoff's eventual killer. Put another way, we can ask whether and how our current knowledge about assessing the risk of violence—and our remaining areas of ignorance—should influence and inform courts' formulations of psychotherapists' duties or the policies enacted through legislation regarding psychiatric patients and their potential risk to the public. We can also ask whether available scientific knowledge about violence prediction should influence the attitudes about risk assessment held by mental health professionals who serve potentially violent clients. We can speculate about what advances in violence prediction mental health professionals and legal decision-makers might expect in coming decades, about the limits on such advances, and about what such limits imply for clinicians' efforts to predict and limit their patients' violent behavior.

Finally, *Tarasoff* in California and its progeny in other states have resulted in action by many state legislatures to define and limit duty-to-protect obligations only to situations where patients have made explicit threats.[26] We can ask whether, given our scientific knowledge, these statutes appropriately address the policy rationale—the perceived need to protect the public from the danger posed by mentally ill persons—that justified *Tarasoff*.

B. The Goals of This Article

This Article has two major goals. First, it explicates a fundamental difference between how mental health researchers interpret the phrase "violence prediction" now and how they understood that phrase in the 1970s. In fact, this development is so fundamental that researchers now use the phrase "violence risk assessment" to refer to the kinds of activities that they called "violence prediction" a few decades ago. Because I shall be examining our understanding of the underlying properties of risk assessment—absent the empirical content of particular individuals' risks for violence or any particular method of assessing

26. *See, e.g.*, ARIZ. REV. STAT. ANN. § 36-517.02, CAL. CIV. CODE § 43.92, COL. REV. STAT. ANN. § 13-21-117, 16 DEL. CODE ANN. § 5402, FLA. STAT. ANN. § 456.059, IDAHO CODE ANN. § 6-1901 *ET SEQ.*, 405 ILL. COMP. STAT. 5/6-103, KY. REV. STAT. ANN. § 202A.400, LA. REV. STAT. ANN. § 9:2800.2, MD. CODE ANN. 5-6 § 5-609, MASS. GEN. LAWS ch. 123, § 36A, MICH. COMP. LAWS ANN. § 330.1946, MONT. CODE ANN. § 27-1-1102, N.H. REV. STAT. ANN. § 329:31, N. J. STAT. ANN. § 2A:62A-17, OHIO REV. CODE ANN. § 2305.51, TENN. CODE ANN. § 33-3-207, UTAH CODE ANN. § 78-14A-101, VA. CODE ANN. § 54.1-2400, and WASH. REV. CODE § 71.05.120.

those risks—, I have titled the Article "Critique of Pure Risk Assessment," in playful (and possibly grandiose) homage to the approach that Immanuel Kant adopted for contemplating questions of metaphysics.[27]

My analogy to Kant goes a step further: a "critical" approach to the problem of violence prediction leads us to important conclusions, one of which is that we may need to abandon the hope that more accurate methods of predicting violence or assessing patients' level of violence risk will prove useful to practicing clinicians. Abandoning this hope may be cause for initial disappointment. But if courts and my mental health colleagues agree with my conclusion, abandoning the hope for useful risk assessments will ultimately liberate[28] us from obligations that we cannot carry out rationally, and will allow us to refocus our attention on treating patients. Just as Kant believed that his critical philosophy would limit speculative reason and thereby remove an obstacle that would otherwise preclude "an absolutely necessary *practical* employment of pure reason" (the grasping of moral issues),[29] I hope that a critique of pure risk assessment will liberate mental health professionals from regarding patients as statistical sources of risk so that we can approach patients instead as sources of initiative and moral worth.

27. In this Article, citations to English translations of Kant's works follow this scheme:

"KrV" = KRITIK DER REINEN VERNUNFT (1781/87), *translated in* CRITIQUE OF PURE REASON (Norman Kemp Smith trans., St. Martin's Press 1965).

"MAR" = Metaphysische Anfangsgründe der Rechtslehre (1797), *translated in* The Metaphysical Elements of Justice (John Ladd's trans., Macmillan 1965).

"SRTL" = *On a Supposed Right to Tell Lies from Altruistic Motives* (1797), in KANT'S CRITIQUE OF PRACTICAL REASON AND OTHER WORKS ON THE THEORY OF ETHICS 361 (T.K. Abbott trans., 1889), *available at* http://oll.libertyfund.org/Home3/ Book.php?recordID=0435.

"KpR" = KRITIK DER PRAKTISCHEN VERNUNFT (3rd ed. 1788), *translated in* CRITIQUE OF PRACTICAL REASON (Lewis White Beck trans., Macmillan 1993).

"GMS" = Grundlegung zur Metaphysik der Sitten (1785), *translated in* GROUNDWORK OF THE METAPHYSIC OF MORALS (H. J. Paton trans., Harper & Row 1964).

"MS" = DIE METAPHYSIK DER SITTEN (1797), *translated in* THE METAPHYSIC OF MORALS (Mary Gregor trans., Cambridge University Press 1996).

Unbracketed page numbers refer to pagination in the above English translations. Bracketed page numbers apply the standard method of reference to Kant's writings, using pagination in the Königlich preußische Akademie der Wissenschaften edition of KANTS GESAMMELTE SCHRIFTEN.

28. For the notion that this insight can be "liberating," I am indebted to Dr. Robert Simon, who made this comment about one of my previous expressions of these ideas. *See* Douglas Mossman, *How a Rabbi's Sermon Resolved My* Tarasoff *Conflict*, 32 J. AM. ACAD. PSYCHIATRY L. 359 (2004) [hereinafter Mossman, *Rabbi's Sermon*].

29. KrV *supra* note 27, at 26 [B xxv].

This leads to this Article's second goal: the description of a moral and legal framework for reformulating the *Tarasoff* rule. Here, my approach is explicitly Kantian. Taking the view that consequentialist, predictive approaches to formulating the duty to protect are ethically questionable (as well as scientifically impractical), I suggest that assigning therapists a duty to intervene in response to a patient's explicit, credible threat is consistent with the view that patients (along with other humans) ought to be treated not merely as means, but always as ends in themselves.

My argument proceeds as follows. In Section II, I recount the facts[30] behind the lawsuit that led to the *Tarasoff* decisions, and summarize key points of the California Supreme Court's ultimate ruling.[31] Although the main "story" behind *Tarasoff* is well known to most mental health clinicians, some of the case's motivating facts may be unfamiliar to legal audiences and younger mental health clinicians who have "grown up" professionally with *Tarasoff*. Moreover, although the facts behind the case may well have justified a potentially more effective action to protect Tatiana Tarasoff (*i.e.*, warning her) than the defendant therapists took, I also think that the California Supreme Court should have arrived at a different "major premise" under which, given the facts of the case, a warning was obligatory. Thus, understanding the facts behind *Tarasoff* is crucial to understanding the standard that the California Supreme Court should have formulated and the standard to which mental health professionals should be held.

Section III describes what *Tarasoff* implies about therapists' competing obligations and the structure of therapists' knowledge about future violence by their patients. Section IV discusses how researchers conceptualized knowledge about future violence when *Tarasoff* was issued and compares this to how researchers now think about anticipating and predicting future violence. Section V describes an implication of this new conceptualization: it is nearly impossible to achieve agreement on how and when to implement the duty-to-protect as *Tarasoff* defines it, because the ruling includes a duty to know whether an undefinable threshold probability of violence has been reached.

Section VI provides a Kantian perspective on the issues raised by the clinical encounter that brought about *Tarasoff*. Section VII uses this perspective to argue for defining the duty to protect as a duty to respond to things that a patient has said and done. Section VIII summarizes the

30. I present the facts as they exist in published court accounts. Although many forensic mental health professionals know additional information about the case, I have no personal knowledge concerning any of the parties involved in the criminal or civil cases discussed here.

31. That is, *Tarasoff II*, 551 P.2d 334 (Cal. 1976).

Article's arguments and describes their implications for legal decision-makers and clinicians.

II. THE *TARASOFF* DECISION REVISITED

A. Events Leading to the Lawsuit[32]

Prosenjit Poddar was born into the Harijan ("untouchable") caste in Bengal, India.[33] In September 1967, he came to the University of California at Berkeley as a graduate student. In the fall of 1968, Poddar met Tatiana Tarasoff while attending folk dancing classes. They saw each other weekly throughout the fall, and on New Year's Eve, Tatiana[34] kissed Poddar. Poddar interpreted the kiss as signifying that the two had a serious relationship, though this was not what Tatiana had meant to imply. When she learned of Poddar's belief, Tatiana told him that she was involved with other men and otherwise indicated that she did not want to have an intimate relationship with him.[35]

The rebuff led Poddar to undergo a severe emotional crisis, in which he had periods of depression and neglected his appearance, meals, studies, and health. During occasional contacts with Tatiana, Poddar audiotaped some of their conversations, trying to figure out why she did not love him. He told others about his being in love with Tatiana and his thoughts about killing her, saying that he could not control himself.[36]

---

32. For more detailed accounts, *see* Glenn S. Lipson & Mark J. Mills, *Stalking, Erotomania, and the* Tarasoff *Cases*, *in* J. REID MELOY, THE PSYCHOLOGY OF STALKING: CLINICAL AND FORENSIC PERSPECTIVES 259–73 (Meloy ed., 1998); Robert F. Schopp & Michael R. Quattrocchi, Tarasoff*, the Doctrine of Special Relationships, and the Psychotherapist's Duty to Warn*, 12 J. PSYCHIATRY & L. 13 (1984); and LEON VANDECREEK & SAMUEL KNAPP, TARASOFF AND BEYOND: LEGAL AND CLINICAL CONSIDERATIONS IN THE TREATMENT OF LIFE-ENDANGERING PATIENTS 2–7 (2001).

33. People v. Poddar, 518 P.2d 342, 344 (Cal. 1974) [hereinafter *Poddar II*]. Poddar's cultural background may be relevant to his psychological responses (discussed in the following paragraphs). At Poddar's trial, defense counsel had wanted to present expert testimony of an anthropologist concerning the cultural stresses that would have affected Poddar's adjustment "from the simple culture in which he had lived . . . to the sophisticated milieu of an American university." The trial court ruled that the anthropologist could testify about cross-cultural difficulties, but that only the psychiatric experts could answer hypothetical questions related to Poddar's diminished capacity. The trial court's decision was sustained on appeal. People v. Poddar, 103 Cal. Rptr. 84, 88 (Cal. Ct. App. 1972) [hereinafter *Poddar I*]. *See also* Leslie Bender, *Teaching Tort Stories*, 55 J. LEGAL EDUC. 108, 113 (2005) (discussing potential significance of Poddar's background and other cultural issues).

34. At several places in this article, I refer to Poddar's victim as "Tatiana." Use of her first name implies no disrespect toward Ms. Tarasoff. Rather, I am trying to make it easy for the reader to distinguish references to Poddar's victim from references to her parents or the legal cases that bear her last name.

35. *Poddar II*, 518 P.2d at 344.

36. *Poddar I*, 103 Cal. Rptr. at 86.

Over the next several months, as his mental condition deteriorated, Poddar became socially isolated, spoke disjointedly, and often cried.[37]

In the summer of 1969, while Tatiana was traveling in South America, Poddar began to improve psychologically. At the suggestion of a friend, he sought outpatient treatment through the university's mental health service[38] and became the voluntary outpatient of a psychologist employed by Cowell Memorial Hospital at the university.[39]

In August 1969, Poddar told the psychologist of his intent to kill someone (readily identifiable as Tatiana) when she returned from Brazil. The psychologist and two psychiatrist colleagues agreed that Poddar should be committed for observation in a mental hospital. The psychologist notified the campus police orally and by letter. Three police officers took Poddar into custody, but, satisfied that Poddar was rational, released him on his promise to stay away from Tatiana. Dr. Powelson, the psychiatrist who directed Cowell Memorial Hospital's department of psychiatry, then asked the police to return the psychologist's letter and ordered that no further action be taken to hospitalize Poddar.[40]

Poddar stopped seeing his psychologist after the police detained him.[41] He continued to follow[42] Tatiana, however, and at one point overheard her talking about a relationship with another man.[43] On October 27, 1969, Poddar went to Tatiana's home to speak with her. Tatiana was not there, and her mother told Poddar to leave. He returned later that day, however, armed with a pellet gun and a kitchen knife, and found Tatiana alone. When she refused to speak with him and ran from the house, Poddar caught up with her and stabbed her to death. He then returned to the Tarasoffs' home and called the police.[44]

37. *Poddar II*, 518 P.2d at 344.

38. *Id*.

39. *Tarasoff II*, 551 P.2d 334, 340 (Cal. 1976).

40. *Id.; Tarasoff I*, 529 P.2d 553 (Cal. 1974).

41. *Tarasoff I*, 529 P.2d at 559.

42. Several clinicians regard Poddar's behavior as what the Anglophone world now calls "stalking." *See, e.g.,* Robert Lloyd-Goldstein, *De Clérambault On-Line: A Survey of Erotomania and Stalking from the Old World to the World Wide Web*, *in* MELOY, *supra* note 32, at 198; Louis B. Schlesinger, *Stalking, Homicide, and Catathymic Process: A Case Study*, 46 INT'L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 64, 69 (2002) (citing the *Tarasoff* case as an example of homicide following stalking). Concerning the social context that led to the emergence of stalking as a distinct criminal offense and a recognized pattern of behavior, *see* Paul E. Mullen et al., *Stalking: New Constructions of Human Behaviour*, 35 AUSTL. N. Z. J. PSYCHIATRY 9 (2001).

43. *Poddar I*, 103 Cal. Rptr. 84, 86 (Cal. Ct. App. 1972).

44. *Poddar II*, 518 P.2d 342, 345 (Cal. 1974).

At his criminal trial, Poddar offered a diminished capacity defense, arguing (through testimony by three psychiatrists and one psychologist) that his paranoid schizophrenia precluded his harboring malice aforethought at the time of the killing. In rebuttal, a court-appointed psychiatrist testified that Poddar was merely schizoid and could harbor the mental states requisite for first- or second-degree murder.[45]

A jury found Poddar criminally responsible and guilty of second-degree murder. Poddar filed appeals in which he argued that the trial court erred in giving implied malice and second-degree murder instructions. Following a decision in which a lower court held that the trial court had erred in its jury instructions relating to the effect on any implied malice of Poddar's possible diminished capacity (implied malice being an element of second-degree murder),[46] the California Supreme Court reversed Poddar's conviction and ruled that he should be retried.[47] Rather than prosecute Poddar again, however, the state of California released him on condition that he go back to India and not return to the United States.[48]

Tatiana's parents did not just mourn the loss of their daughter. They sued Poddar's therapist, associates, and their employing institution, alleging (*inter alia*) therapist negligence for failure to detain Poddar and failure to warn the Tarasoffs of the grave danger to Tatiana that Poddar represented.[49] They also sought punitive damages for Dr. Powelson's behavior following the therapists' attempt to have Poddar hospitalized, alleging that his actions constituted "malicious and oppressive abandonment of a dangerous patient."[50]

The Alameda County trial court dismissed the Tarasoffs' suit. An appeals court also ruled against the Tarasoffs, holding that (a) their case could not proceed because an action for failure to detain was statutorily barred, and (b) the lack of a special relationship between the clinicians and either Tatiana or her parents precluded any duty to warn them.[51] The Tarasoffs then appealed to the California Supreme Court.

45. *Id*.

46. *Poddar I*, 103 Cal. Rptr. at 93.

47. *Poddar II*, 518 P.2d at 350.

48. Fillmore Buckner & Marvin Firestone, "*Where the Public Peril Begins": 25 Years after* Tarasoff, 21 J. LEGAL MED. 187, 195 (2000).

49. *Tarasoff II*, 551 P.2d 334, 431 (Cal. 1976).

50. Tarasoff v. Regents of the Univ. of Cal., 108 Cal. Rptr. 878, 881 (Cal. Ct. App. 1973).

51. *Id.* at 880–87.

B. The Facts: A Summary

For the Tarasoffs' lawsuit to proceed and liability to be imposed, the California Supreme Court would need to declare a rule under which, given the events that preceded Tatiana's death, the behavior of Poddar's psychologist and his psychiatrist associates would be negligent. As we have just seen, published case law reports that the following events gave rise to the Tarasoffs' lawsuit:

- In August 1969, Poddar told his treating psychologist that he was thinking about killing someone, and the psychologist could readily identify this person as Tatiana.
- Shortly after hearing this, the psychologist, with the concurrence of two psychiatrist colleagues, notified campus police and asked them to pick up Poddar so that he could be hospitalized.
- The police detained and spoke to Poddar, decided that he was rational, and released him after telling him to stay away from Tatiana.
- Dr. Powelson, the psychiatric director at the hospital where Poddar received outpatient treatment, ordered that no further action be taken to hospitalize Poddar.
- Poddar stopped treatment and apparently had no further contact with his therapist.
- On October 27, 1969, Poddar came to Tatiana's home, and her mother told him to go away.
- Later that same day, Poddar returned and killed Tatiana.

Let us assume that a majority of the California Supreme Court judges, viewing the allegations of the Tarasoffs, reacted by concluding that if those allegations were true, then Poddar's clinicians should have to answer to the plaintiffs. The majority's task would then consist of having to fashion a legal rule,[52] in the form of a major premise, under

52. I take the position here that in thinking about complex moral and legal issues, judges follow the same psychological process that everyone else does: they come to a conclusion, then fill in a rationale. Here's one law professor's summary of the process:

> Almost all justices vote almost all of the time in accordancewith their own personal, political and religious views . . . .
>
> Though presidents, senators and judicial nominees loudly proclaim that justices should merely apply "the law" in a neutral manner, every experienced lawyer understands that the best predictors of a justice's actual votes are his or her personal, political and religious predilections. Any lawyer who ignores this reality is doomed to failure. . . . That is why good lawyers check the biographical material about judges even before they read their cases.

Alan Dershowitz, *What Kind of Justice Will Alito Be?* (Jan. 13, 2006) FORBES.COM

which the behavior of the clinicians and other events that led to the *Tarasoff* lawsuit would make the treating clinicians liable. To create this major premise, the majority would have to fill in the variables in the following conditional statement:

- If a patient does A, and the therapist does not do B, and the patient later harms C, then the therapist will be liable for the harm to C.

The minor premise, in the *Tarasoff* case, would be …

- Poddar (the patient) did A, his psychologist (the therapist) did not do B, and Poddar then harmed C (Tatiana).

… from which it follows, *modus ponens*, that

- the therapist (Poddar's psychologist) is liable for the harm to C (Tatiana).

C. The Tarasoff *Rule*

1. Overcoming Obstacles

In formulating a rule under which Poddar's clinicians might be liable, the majority first had to address three objections raised by the defendant clinicians: (a) the lack of a treatment relationship between the clinicians and Tatiana, (b) the difficulty of predicting violence, and (c) the fact that warning her would have required the clinicians to breach therapeutic confidentiality.

a. Lack of Treatment Relationship

Notwithstanding the absence of a treatment relationship involving Tatiana, the treatment relationship between Poddar and the defendant clinicians, said the majority, "may support affirmative duties for the benefit of third persons."[53] The majority cited a California case that gave physicians duties to control the danger of hospitalized patients[54] and a Washington state case requiring doctors to warn patients

http://www.forbes.com/columnists/2006/01/12/alito-confirmation-dershowitz-comment-cx_ad_0113alito.html (last visited Jan. 14, 2006). *See also* Richard A. Posner, *Foreword: A Political Court*, 119 HARV. L. REV. 31, 34 (2005) ("the impressions that I have gleaned from being a federal appellate judge for the last twenty-four years" lead to the conclusion "that, viewed realistically, the Supreme Court, at least most of the time, when it is deciding constitutional cases is a political organ").

53. *Tarasoff II*, 551 P.2d at 343.

54. *Id,* (citing Vistica v. Presbyterian Hosp., 432 P.2d 193 (Cal. 1967)).

themselves about conditions or medications that might result in danger to others.[55] The majority also cited decades-old decisions from other jurisdictions that gave physicians duties to warn potential contacts (*e.g.*, family members) about a patient's contagious disease.[56]

b. Difficulty of Predicting Violence

Interestingly, the *Tarasoff* majority made two seemingly contradictory arguments to deal with and dismiss problems with anticipating future violence. On the one hand, the court "recognize[d] the difficulty that a therapist encounters in attempting to forecast whether a patient presents a serious danger of violence."[57] The court felt that to resolve this, it needed only to apply the traditional negligence standard to the problem of violence prediction. This would require a mental health professional to "forecast a serious danger" only with

> 'that reasonable degree of skill, knowledge, and care ordinarily possessed and exercised by members of [that professional specialty] under similar circumstances.' . . . [T]he therapist is free to exercise his or her own best judgment without liability; proof, aided by hindsight, that he or she judged wrongly is insufficient to establish negligence.[58]

On the other hand, the court said that it was not necessary to base its ruling on any assumed accuracy of predictions. "In the instant case," the majority said,

---

55. *Tarasoff II*, 551 P.2d at 344 (citing Kaiser v. Suburban Transp. Sys., 398 P.2d 14 (Wash. 1965).

56. *Tarasoff II*, 551 P.2d at 344 (citing Wojcik v. Aluminum Co. of America, 183 N.Y.S.2d 351 (N.Y. Gen. Term 1959); Davis v. Rodman 227 S.W. 612 (Ark. 1921); Skillings v. Allen, 173 N.W. 663 (Minn. 1919); Jones v. Stanko, 160 N.E. 456 (Ohio 1928)).

57. *Tarasoff II*, 551 P.2d at 345. Among the publications the Court cited in acknowledging the problem of prediction was a work of Professor Monahan's. *See Tarasoff II*, 551 P. 2d at 344, *citing* Monahan, *The Prevention of Violence*, *in* COMMUNITY MENTAL HEALTH IN THE CRIMINAL JUSTICE SYSTEM (Monahan ed., 1975).

58. *Tarasoff II*, 551 P. 2d at 345 (citations omitted). The Court apparently did not recognize how high a standard it actually set. Viewed "in hindsight," *i.e.*, after a patient had harmed someone, any judgment by a therapist that the patient did not pose a "serious danger" might easily appear—to opposing counsel, opposing experts, and the factfinder—to have fallen short of the therapist's "best judgment."

In other words, the court seems unaware of the impact of hindsight bias, a well-documented tendency for individuals who know an outcome to exaggerate the ease with which the outcome was predictable in advance, or to exaggerate the advance likelihood of an event once it has already occurred. The classic articles include Baruch Fischhoff, *Hindsight ≠ Foresight: The Effect of Outcome Knowledge on Judgement under Uncertainty*, 1 J. EXPER. PSYCHOL. HUM. PERCEPTION & PERFORMANCE 288 (1975), and Scott A. Hawkins & Reid Hastie, *Hindsight: Biased Judgements of Past Events after the Outcomes Are Known*, 107 PSYCHOL. BULL. 311 (1990).

> the pleadings do not raise any question as to failure of defendant therapists to predict that Poddar presented a serious danger of violence. On the contrary, the present complaints allege that defendant therapists did in fact predict that Poddar would kill, but were negligent in failing to warn.[59]

The *Tarasoff* majority even takes the position that concern about prediction accuracy may be absurd. Imagine that a patient talks to his therapist about his plans for a political killing. "We would hesitate to hold that the therapist who is aware that his patient expects to attempt to assassinate the President of the United States would not be obligated to warn the authorities because the therapist cannot predict with accuracy that his patient will commit the crime."[60]

c. Confidentiality

The *Tarasoff* majority addressed the problem of confidentiality through a cost-benefit analysis. While recognizing the public health benefits of having psychotherapeutic treatment occur in contexts with assurances of privacy, the court thought the cost to the patient of sacrificing privacy was merely "conjectural," while the "peril to the victim's life" was not.[61] Accordingly, society's safety outweighs therapeutic confidentiality. Despite the existence of a "public policy favoring protection of the confidential character of patient-psychotherapist communications," this protection "must yield to the extent to which disclosure is essential to avert danger to others."[62]

---

59. *Tarasoff II*, 551 P. 2d at 345.

This statement deserves two comments. First, one might maintain that the Court's positions are not contradictory by saying that adherence to professional standards is the issue, not accuracy. In other words, suppose that mental health professionals made predictions by reading palms, viewing tea leaves, and divining the future from entrails. For purposes of assigning liability, all that would matter is whether the defendant-therapist followed his profession's customary procedures for reading palms, viewing tea leaves, or examining entrails; the fact that these forecasting methods are useless would be irrelevant.

Second, notice that the *Tarasoff* majority states that Poddar's therapist and his colleagues "did in fact *predict* that Poddar would kill . . . ." *Id*. (emphasis added). One could argue, however, that the clinicians did no such thing—they responded to Poddar's uttered statements. As Section VII emphasizes *infra*, predicting violence and responding to a verbalized threat are two very different actions.

60. *Tarasoff II*, 551 P.2d at 346. Notice, again, that the majority has characterized a therapist's being "aware" that "his patient expects to attempt to assassinate the President" as a prediction, when in fact what most likely would have occurred was that the therapist *heard* the patient *describe* his plans. *Id*. Clearly, however, hearing a patient make a threat is an activity that differs crucially from making a prediction of the patient's behavior.

61. *Id.* at 346.

62. *Id.* at 347.

2. The Major Premise

Recall that, under the assumption that a majority of the California Supreme Court wished to substantiate its perception that Poddar's clinicians had a responsibility to respond to his statements about future violence, the court needed to formulate a major premise from which the facts in *Tarasoff* could be grounds for finding the clinicians liable. The general form of this major premise was, "If a patient does A and the therapist does not do B and the patient later harms C, the therapist will be liable for the harm to C." Under the *Tarasoff* majority's ruling, A, B, and C take on these meanings:

- A =displays behavior that a therapist should recognize as indicating that the patient presents a serious danger of violence
- B =(1) apply the standards of the therapist's profession to determine that the patient presents a serious danger of violence, and then (2) use reasonable care to protect an intended victim[63] from the danger
- C =the victim

The minor premise then becomes

- Poddar (the patient) made a threat, that is, displayed behavior that a therapist should have recognized as indicating that Poddar presented a serious risk of violence, and his psychologist (the therapist) did not both (1) apply the standards of his profession to determine that Poddar presented a serious danger of violence and then (2) use reasonable care to protect an intended victim from the danger, and Poddar later harmed Tatiana (the victim).

The conclusion must be

- The therapist (Poddar's psychologist) is liable for the harm to Tatiana (the victim).

Here is how the *Tarasoff* majority expressed the major premise equivalently, if more succinctly:

63. The phrase "intended victim" has some ambiguity. The *Tarasoff* court probably meant to point to the specific individual(s) whom the patient intended to harm and whom a therapist could readily identify as an intended victim. This interpretation is supported by subsequent California decisions that limited the scope of a therapist's duty to readily identifiable victims. *See* Thompson v. County of Alameda, 614 P.2d 728, 734 (Cal. 1980) (declining to impose "blanket liability" for harm to any conceivable victim); Mavroudis v. Superior Court, 102 Cal. App. 3d 594, 600–01 (Cal. Ct. App. 1980) (duty triggered only when patient poses "an imminent threat of serious danger to a readily identifiable victim"). As we shall see, however, later decisions relying on *Tarasoff* effectively interpreted this phrase as designating those individuals who were harmed as a result of a patient's intentional, harm-inducing actions, whether or not a therapist could have identified them in advance.

> When a therapist determines, or pursuant to the standards of his profession should determine, that his patient presents a serious danger of violence to another, he incurs an obligation to use reasonable care to protect the intended victim against such danger.[64]

This formulation leaves open the question, "What does it mean to 'use reasonable care to protect the intended victim'? Were not the actions that Poddar's clinicians' took—which included asking the police to assist with involuntary hospitalization two months before the homicide—'reasonable care'? What more could you ask of them?" The court gave an open-ended answer: it "depend[s] upon the nature of the case."[65] The therapist might have "to warn the intended victim or others likely to apprise the victim of the danger, to notify the police, or to take whatever other steps are reasonably necessary under the circumstances."[66] Thus, "under the circumstances" faced by Poddar's clinicians, mere efforts to hospitalize Poddar might not have sufficed, and they could be liable for Tatiana's death.

III. WHAT *TARASOFF* IMPLIES

Two key assumptions lie behind the *Tarasoff* majority's ruling. These assumptions involve implicit beliefs about (1) the nature of policy judgments and (2) relationships between certain kinds of clinical facts and clinicians' knowledge of those facts.

A. Balancing of Confidentiality Against Public Safety

The majority's conclusion that "the public policy favoring protection of the confidential character of patient-psychotherapist communications must yield to the extent to which disclosure is essential to avert danger to others" suggests that it is possible to decide, through a utilitarian calculus, whether the benefits of assuring therapeutic confidentiality are worth more or less than having therapists serve as a line of defense against violence. As Section I of this Article[67] notes, *Tarasoff* declares the therapist-patient relationship an exception to the general rule that individuals have no duty to control the conduct of other persons. The decision employs a balancing test that considers several factors, including "'the policy of preventing future harm, the extent of the

64. *Tarasoff II*, 551 P.2d at 340.

65. *Id*.

66. *Id*.

67. *See supra* note 21 and accompanying text.

burden to the defendant and consequences to the community of imposing a duty to exercise care with resulting liability for breach, and the availability, cost and prevalence of insurance for the risk involved.'"[68]

In making the case for an exception to the presumption of nonresponsibility for the acts of others, *Tarasoff II* cites a series of cases that involved mental patients who acted violently.[69] Thus, one factor that seems to have tipped the scale in favor of a "policy decision" making therapists liable was the unique type of risk posed by mental patients. As the following paragraph suggests, that risk, if avoidable, was one the court found unacceptable:

> Our current crowded and computerized society compels the interdependence of its members. In this risk-infested society we can hardly tolerate the further exposure to danger that would result from a concealed knowledge of the therapist that his patient was lethal. If the exercise of reasonable care to protect the threatened victim requires the therapist to warn the endangered party or those who can reasonably be expected to notify him, we see no sufficient societal interest that would protect and justify concealment. The containment of such risks lies in the public interest.[70]

In his dissent, Justice Clark concluded that the majority had clearly gotten this calculation wrong. "Overwhelming policy considerations" he wrote, "weigh against" the duty imposed by the majority.[71] Without an assurance of confidentiality, those needing treatment might not seek it because of the stigma attached to getting mental health care, and treatment would be less effective for those who did get it.[72] Further, an absence of confidentiality would lead to an absence of trust in one's therapist, "the very means by which treatment is effected."[73] Rather than reduce violence, Justice Clark concluded, "the duty to warn imposed by the majority will cripple the use and effectiveness of psychiatry. Many people, potentially violent—yet susceptible to

68. *Tarasoff II*, 551 P.2d at 342 (quoting Merrill v. Buck, 375 P.2d 304 (Cal. 1962); Biakanja v. Irving, 320 P.2d 16 (Cal. 1958); Walnut Creek Aggregates Co. v. Testing Eng'rs Inc., 56 Cal. Rptr. 700 (Cal. Ct. App. 1967)).

69. *See Tarasoff II*, 551 P.2d at 344 (citing Vistica v. Presbyterian Hosp., 432 P.2d 193 (Cal. 1967); Semler v. Psychiatric Inst. of D.C., 538 F.2d 121 (4th Cir. 1976); Underwood v. U.S., 356 F.2d 92 (5th Cir. 1966); Fair v. U.S., 234 F.2d 288 (5th Cir. 1956); Greenberg v. Barbour, 322 F.Supp. 745 (E.D. Pa. 1971); and Merchs. Nat'l Bank & Trust Co. of Fargo v. U.S., 272 F.Supp. 409 (D.N.D.1967)).

70. *Tarasoff II*, 551 P.2d at 347.

71. *Tarasoff II*, 551 P.2d at 358 (Clark, J., dissenting).

72. *Id.* at 359.

73. *Id.*

treatment—will be deterred from seeking it."[74]

Subsequent scholarship has suggested that, as an empirical matter, both the dissent and the majority were mistaken. Eight years after predicting, in 1976,[75] that the *Tarasoff* decisions would gravely impede therapists' treatment efforts, psychiatrist Alan Stone recognized that "the duty to warn is not as unmitigated a disaster for the enterprise of psychotherapy as it once seemed to critics like myself."[76] Although the ambiguous duties imposed by *Tarasoff* caused anxiety among mental health professionals, therapists nonetheless adjusted to what they perceived as the law's new expectations. The study that has done the best job of examining the impact of implementing *Tarasoff* suggested that when therapists had issued warnings, "in most cases issuing the warning had a minimal or a positive effect on the psychotherapeutic relationship."[77] In another study, the same researchers found "that almost half of the targets of patients' threats were family members, spouses, boyfriends, or girlfriends," a finding that they thought supported the view "that the *Tarasoff* type of situation," rather than being detrimental for treatment, "may hold promise for family-oriented therapeutic interventions."[78]

But if warnings have not undermined psychotherapy, they may also not accomplish the public protection goal that motivated the majority's holding in *Tarasoff*. Psychiatrist Thomas G. Gutheil suggests that often, courses of action that require therapists to breach confidentiality can be expected to have *worse* outcomes than courses of action that preserved confidentiality. Dr. Gutheil demonstrates this by reviewing the original facts of *Tarasoff* from the vantage point of a "modern clinician." That is, Dr. Gutheil assumes that a "hypothetical earlier case" had already established a *Tarasoff*-like duty to protect, and then looks at how effective—as public protection measures—various options for protective intervention might have been in averting the danger posed by Poddar. Trying to commit Poddar to a hospital would not have protected Tatiana; even if treaters had known Poddar would attack her in October 1969, her

74. *Id.* at 360.

75. Alan Stone, *The* Tarasoff *Decisions: Suing Psychotherapists to Safeguard Society*, 90 HARV. L. REV. 358 (1976).

76. STONE, *supra* note 13, at 181.

77. Renee Binder & Dale McNiel, *Application of the* Tarasoff *Ruling and Its Effect on the Victim and the Therapeutic Relationship*, 47 PSYCHIATRIC SERVICES 1212, 1212 (1996) (following California's enactment of a statute prescribing specific conditions for dealing with threatened violence, almost half of psychiatry residents had issued *Tarasoff*-type warnings).

78. Dale McNiel et al., *Management of Threats of Violence Under California's Duty to Protect Statute*, 155 AM. J. PSYCHIATRY 1097, 1100 (1998).

absence from the country when Poddar uttered his threat meant that no danger was imminent, and no court would have authorized his involuntary hospitalization. Because Tatiana was out of the country, warning her directly would have been "difficult to impossible." Warnings to Tatiana's family members would have been of questionable value because whether and when family members would convey this information "might be difficult to predict."[79]

Although Poddar may have left "treatment precisely because of the breach of confidentiality," Gutheil believes that Poddar's therapist could instead have maintained confidentiality, trying "to keep him in treatment aimed at decreasing his shame, rage, and dangerousness." If Poddar had still felt intense rage toward Tatiana after she returned, his treating psychologist might have encouraged and helped Poddar to notify her himself, which would have obviated any reason for the therapist to breach therapeutic confidentiality. Gutheil's point is that one cannot generalize about the costs and benefits of warnings or other protective measures; without analyzing the facts of a particular case, one cannot conclude what course of action best addresses the concerns of the patient, the therapist, and the larger public.[80]

The difficulty of deciding whether a policy of preserving therapeutic confidentiality provides more value than a policy that mandates warnings or other actions based on probabilistic judgments is a matter to which this Article will return shortly.[81] For now, it suffices to notice that the reasoning in *Tarasoff* assumes that weighing and balancing of benefits is possible and that such a process can guide legal policy-making.

B. Implicit Model of Patients and of Therapists' Knowledge

Both the 1974 and 1976 *Tarasoff* decisions contain the sentence, state: "The protective privilege ends where the public peril begins."[82] This occurrence prompts Buckner and Firestone to observe, "[t]he court obviously liked the ring of this phrase."[83] Besides being a pithy, alliterative summation of the court's views on confidentiality versus preventing violence, the sentence also suggests that one could, in theory, establish a clear demarcation between those clinical situations in which

79. Gutheil, *supra* note 2, at 353.

80. *Id*.

81. *See* discussion *infra*, Part V (describing broad disagreements in individuals' perceptions of the desirability of various outcomes).

82. *Tarasoff I*, 529 P.2d 553, 561 (Cal. 1974), *Tarasoff II*, 551 P.2d 334, 347 (Cal. 1976).

83. Buckner & Firestone, *supra* note 48, at 198.

patients pose "a serious danger of violence" that constitutes a "public peril," and those situations in which patients pose no such danger. "[A] therapist should not be encouraged routinely to reveal such threats," says the court. "To the contrary, the therapist's obligations to his patient require that he not disclose a confidence unless such disclosure is necessary to avert danger to others."[84]

This point about the implications of the *Tarasoff* court's thinking deserves more precise delineation. In referring at one point to the "concealed knowledge of the therapist that his patient was lethal,"[85] *Tarasoff* implies that patients either are "lethal" or they are not. Presumably, because the *Tarasoff* rule refers to "a serious danger of violence," the court's implication extends to forms of violence that are not lethal, but that result in significant injury. Some reflection suggests that this makes intuitive sense: because death either occurs or it does not, one can logically dichotomize people who become victims of patient violence as those who die and those who do not. Given a clear definition of "serious injury,"[86] one could also divide victims into those who were injured seriously and those who were not. Similarly, reflection suggests that one could logically dichotomize therapists' choices about patients: faced with a particular clinical situation, a therapist can either to take some form of protective action, or not do so.

Tarasoff carries this dichotomization beyond the realm of facts about the world—a patient either does or does not commit violence, a therapist either takes or does not take protective action—to the realm of therapists' *knowledge* about those facts. *Tarasoff* assumes that, because death or serious violence are either/or phenomena and because taking action is something a therapist either does or does not do, a therapist therefore either has knowledge about future violence or does not. By virtue of such knowledge, a therapist can either realize that a duty to warn or take other protective action has arisen, or can fail to realize this.

84. *Tarasoff II*, 551 P.2d at 347.

85. *Id.*

86. In any study of violence prediction, researchers must decide what actions "count" as violent events so that they can decide whether a particular person should be regarded as having acted violently. A moment's reflection will reveal that individuals' actions come in degrees of violence, from more minor (and questionably violent) incidents such as pushing and shoving, to major (and unquestionably violent) incidents such as lethal assaults with firearms. I discuss this problem further in Douglas Mossman, *Assessing Predictions of Violence: Being Accurate about Accuracy*, 62 J. CONSULTING CLINICAL PSYCHOL. 783, 784 (1994) [hereinafter Mossman, *Accuracy*]. An example of how violence is defined for purposes of such a study appears in Henry J. Steadman et al., *Violence by People Discharged From Acute Psychiatric Inpatient Facilities and by Others in the Same Neighborhoods*, 55 ARCHIVES GEN. PSYCHIATRY 393, 395 (1998) (violence defined as battery that results in physical injury, sexual assaults, assaultive acts with weapons, or threats made while holding a weapon).

As framed by the *Tarasoff* decision, the duty to protect takes this form:

- there are naturally only two types of patients—those who will do violence and those who will not;
- therapists therefore must categorize patients into those about whom action to protect a third party is warranted, and those about whom no such action is warranted;
- therapists then should take the action that their categorization dictates.

True, *Tarasoff* does not explicitly require therapists to get every judgment right, but only to use their "best" judgment and to exercise reasonable care in their employment of professional standards when making these judgments. The point, however, is that therapists' judgments are assumed to take the form of binary, "yes-or-no" assessments about whether "a serious danger of violence" exists. More succinctly: because violence either will occur or will not, and because a therapist can either take a protective action or not, therapists' judgments about violence are assumed to take the form of predictions that violence either will or will not occur.[87]

C. Subsequent Cases

As troublesome as *Tarasoff* was, its legal progeny were even more unsettling for therapists. Thorough review of these cases would take us far beyond the scope of this article. The cases described in the following paragraphs illustrate (1) how other courts have perceived clinical violence assessments as binary predictions and (2) the expansion of mental health professionals' prediction duties beyond situations in which patients make threats to harm specific persons to include all "foreseeable" victims harmed by intentional actions—an expansion that flows logically from the liability-defining major premise created by the *Tarasoff* court.

87. Quattrocchi and Schopp believe that viewing violence assessments as binary arises from *Tarasoff*'s origins in negligence law, where foreseeability is central to the existence of a duty. They believe that, in the case of possible future violence, prediction foreseeability "generates protective obligations that reflect dichotomous classification of persons. Those classified as not dangerous trigger no protective obligation on the part of the clinician and those classified as dangerous trigger a protective obligation." Michael R. Quattrocchi & Robert F. Schopp, *Tarasaurus Rex: A Standard of Care That Could Not Adapt*, 11 PSYCHOL. PUB. POL'Y & L. 109, 177 (2005).

1. *Lipari v. Sears*

The first major case articulating the expanding duty to protect was *Lipari v. Sears Roebuck & Company*.[88] Ulysses Cribbs had been committed to a psychiatric hospital and had received psychiatric care from the Veterans Administration (the VA). He purchased a shotgun from Sears while he was undergoing a month-long episode of day treatment[89] at the VA. A month after Cribbs stopped the treatment, he walked into a nightclub, shot Dennis Lipari to death, and seriously wounded Lipari's wife, Ruth Ann. Mrs. Lipari sued Sears, alleging negligent sale of a shotgun to a mentally ill person; Sears and Mrs. Lipari sued the VA on the basis of negligent failure to detain the patient who should have been recognized to be dangerous. As in *Tarasoff*, the alleged tort involved failure to take adequate action to protect a third party.

The VA's motion for dismissal was denied by District Court, because, under its reading of Nebraska law and § 315 of the Restatement (Second) of Torts,

> the relationship between a psychotherapist and his patient gives rise to an affirmative duty for the benefit of third persons. This duty requires that the therapist initiate whatever precautions are reasonably necessary to protect potential victims of his patient. This duty arises only when, in accordance with the standards of his profession, the therapist knows or should know that his patient's dangerous propensities present an unreasonable risk of harm to others.[90]

Unlike the psychologist who treated Tatiana Tarasoff's killer, Cribbs's care-givers had never heard their patient threaten the Liparis or anyone else. But as was the case in the *Tarasoff* ruling, the liability trigger in *Lipari* was not a threat or other action of the patient. Instead, *Lipari* followed *Tarasoff* in implicitly requiring the therapist to take action following a judgment about the patient's future behavior, a judgment embodied in a yes-or-no prediction concerning whether a patient represents "an unreasonable risk of harm."

88. 497 F.Supp. 185 (D. Neb. 1980).

89. Day treatment "[is] a form of partial hospitalization [that] can be helpful for patients who do not require inpatient care but who may benefit from more intensive care than is possible for outpatients." John S. Ogrodniczuk & Paul I. Steinberg, *A Renewed Interest in Day Treatment,* 50 CAN. J. PSYCHIATRY 77, 77 (2005).

90. *Lipari*, 497 F.Supp. at 193.

2. *Jablonski v. U.S.*

The ability to distinguish violent from nonviolent patients is an important implicit premise underlying *Jablonski v. United States*.[91] On July 7, 1978, Phillip Jablonski threatened and attempted to rape Isobel Pahls, the mother of Melinda Kimball, with whom Jablonski had been living. Three days later, Jablonski agreed to undergo an outpatient psychiatric evaluation at the Loma Linda (California) VA Hospital, and Kimball went with him. Pahls had called local police, who in turn had called the VA about Jablonski's prior criminal record, but no one relayed this information to the psychiatrist who evaluated Jablonski. During the evaluation, Jablonski acknowledged receiving prior psychiatric treatment, but the evaluating psychiatrist did not obtain records of Jablonski's treatment at other VA facilities in California and other states. Records from Jablonski's treatment ten years earlier in El Paso, Texas would have shown that at that time, Jablonski had homicidal thoughts about his wife and had "on numerous occasions . . . tried to kill her."[92]

The Loma Linda psychiatrist offered Jablonski voluntary admission, which Jablonski refused; the psychiatrist concluded there were no grounds for involuntary hospitalization. The psychiatrist recommended that Kimball stay away from Jablonski, but she would not do so despite telling the psychiatrist she felt unsafe with him. Two psychiatrists met with Jablonski again a few days later and again concluded that he did not meet criteria for involuntary hospitalization. While Jablonski was being evaluated, a third VA psychiatrist again recommended that Kimball stay away from Jablonski. Kimball did not, however, and two days later, Jablonski murdered her.[93]

Kimball's daughter sued the VA for malpractice, and the district court found liability by the hospital psychiatrists for failing to record and transmit the information from the police, failing to obtain Jablonski's past psychiatric records, and failing to adequately warn Kimball; each failure, ruled the district court, was a proximate cause of Kimball's death.[94] The U.S. Court of Appeals upheld the district court's ruling, finding that the *Tarasoff* decision was "on point."[95] Though Jablonski had not threatened Kimball, his "previous history," said the appeals

91. 712 F.2d 391 (9th Cir. 1983).

92. *Jablonski*, 712 F.2d at 393.

93. *Id*. at 394. Phillip Jablonski's history of violence, before and after this killing, is described in *People v. Jablonski*, 126 P.3d 938, 944–52 (Cal. 2006).

94. *Jablonski*, 712 F.2d at 394.

95. *Id.* at 398.

court, "indicated that he would likely direct his violence against Kimball." His past acts of violence against his former wife and "[h]is psychological profile indicated that his violence was likely to be directed against women very close to him. This, in turn, was borne out by his attack on Pahls."[96]

Psychiatrists have interpreted *Jablonski* as a lesson telling them that seeking adequate information, especially readily available past records, is crucial to avoiding malpractice liability.[97] But the case also assumes that given adequate information, there can be a crystal-clear distinction between those patients who do and do not have "psychological profiles" portending particular types of violence.

3. *Petersen v. State*

Tarasoff, *Lipari*, and *Jablonski* all had stemmed from *intentional* acts by current or former psychiatric patients. But because these cases had been framed in terms of foreseeable harm created by patients, it was a natural extension to make clinicians liable for reckless acts as well. *Petersen v. State*[98] was the first case to do this.

On May 14, 1977, Cynthia Petersen was injured in an automobile accident caused by Larry Knox, who was speeding, ran a red light, and appeared to be under the influence of drugs. Five days earlier, Knox had been released from Western State Hospital, where he had undergone a two-week involuntary hospitalization for what his doctors ultimately diagnosed as a psychotic reaction to "angel dust." His psychiatrist had prescribed the antipsychotic drug thiothixene for Knox while he was at the hospital. The evening before his discharge, hospital security personnel had apprehended Knox while he was driving his car recklessly on the hospital grounds. But his psychiatrist discharged him believing that he had recovered from the drug reaction, was not psychotic, and had returned to his usual type of personality and behavior. After the accident, Petersen sued the State alleging that the hospital negligently treated Knox by failing to protect her from his dangerous propensities by, for example, seeking additional confinement. The jury agreed and awarded Petersen $250,000.[99]

---

96. *Id.*

97. *See, e.g.*, THOMAS G. GUTHEIL & PAUL S. APPELBAUM, CLINICAL HANDBOOK OF PSYCHIATRY AND THE LAW, 196 (3rd ed. 2000) (discussing the frequent difficulty of obtaining such records and observing, "A discouragingly common element in litigation involves the imputation that the clinician failed to obtain significant old records that allegedly would have altered the treatment plan.").

98. 671 P.2d 230 (Wis. 1983).

99. It may have helped the plaintiff's case to have introduced, at trial, evidence about the Knox

The Washington Supreme Court, citing *Tarasoff* and *Lipari*, upheld the lower court's verdict by finding that the requisite special relationship[100] existed between the hospital and Knox such that they had a duty to control Knox. Because of this, the Washington Supreme Court concluded that the psychiatrist-patient relationship had indeed created "a duty to take reasonable precautions to protect anyone who might foreseeably be endangered by Larry Knox's drug-related mental problems," and that the psychiatrist's decision to discharge Knox without seeking additional involuntary hospitalization was subject to civil sanction via a malpractice action.[101] Implicit in *Petersen* is the notion that, because the decision to seek commitment or not is binary, assessments of whether individuals warrant commitment—that is, assessment of whether there is a need to protect "anyone who might foreseeably be endangered" by patients—are binary as well.

IV. OUR MATHEMATICAL UNDERSTANDING OF VIOLENCE PREDICTION

A. Conceptualizations of the Psychiatric Prediction of Violence: The 1970s and 1980s

The notion that knowledge about future violence should take the form of binary predictions was a feature not of just the *Tarasoff* ruling and its progeny, but of much of the social science research about violence prediction that existed when *Tarasoff* was issued and in the decade that followed. Table 1 explains the meaning of several indices of accuracy used in describing the accuracy of binary predictions.

subsequent actions and psychiatric treatment after the automobile accident with Petersen. The jury learned that seven months after the accident, Knox killed a couple and raped their daughter. The malpractice jury also heard evidence from other psychiatrists who treated Knox after the accident. They testified that Knox had schizophrenia, though they disagreed about what type of schizophrenia Knox suffered. *See Petersen*, 671 P.2d at 242–44.

100. RESTATEMENT (SECOND) OF TORTS § 315 (1965); *see also Petersen*, 671 P.2d at 237.

101. *Petersen*, 671 P.2d at 237.

Table 1. — Chief methods of characterizing the accuracy of binary violence predictions.

	Patients' Actual Behavior:		
Predictions:	Violent ($V+$)	Not violent ($V–$)	*Sums:*
"Will be violent" ($T+$)	E = true positives	F = false positives	$E+F$
"Will not be violent" ($T–$)	G = false negatives	H = true negatives	$G+H$
Sums:	$E+G$	$F+H$	

$$\text{base rate} = BR = \text{probability of violence} = P(V+) = \frac{A+C}{A+B+C+D}$$

$$\text{correct fraction} = \frac{E+H}{E+F+G+H}$$

$$\text{percent correct} = \frac{E+H}{E+F+G+H} \times 100\%$$

$$\text{sensitivity} = \text{true positive rate } (TPR) = P(T+|V+) = \frac{E}{E+G}$$

$$\text{specificity} = \text{true negative rate } (TNR) = P(T–|V–) = \frac{H}{F+H}$$

$$\text{false positive rate } (FPR) = 1 - TNR = P(T+|V–) = \frac{F}{F+H}$$

$$\text{positive predictive value} = PPV = P(V+|T+) = \frac{E}{E+F}$$

$$\text{negative predictive value} = NPV = P(V–|T–) = \frac{H}{G+H}$$

Note that:

$$PPV = P(V+|T+) = \frac{E}{E+F} = \frac{\frac{E}{E+F+G+H}}{\frac{A}{E+F+G+H} + \frac{B}{E+F+G+H}}$$

$$= \frac{\frac{E+G}{E+F+G+H} \times \frac{E}{E+G}}{\left(\frac{E+G}{E+F+G+H} \times \frac{E}{E+G}\right) + \left(1 - \frac{E+G}{E+F+G+H}\right) \times \left(\frac{E+G}{E+F+G+H} \times \frac{F}{F+H}\right)}$$

$$= \frac{BR \times TPR}{(BR \times TPR) + ([1 - BR] \times FPR)} = \frac{P(V+)P(T+|V+)}{P(V+)P(T+|V+) + [1 - P(V+)]P(T+|V–)},$$

which is a statement of Bayes's Theorem.

In Table 1, violence either occurs (*V*+) or does not (*V*–), and—by analogy with diagnostic tests in medicine—the "test" (that is, a prediction about whether violence will occur) for violence is either "positive" (*T*+) or negative (*T*–). If a clinician predicts that a subject will act violently and that subject later does commit violence, that prediction is a "true positive." If a clinician predicts that a subject will act violently but the subject does not, that prediction is a "false positive." Similarly, predictions of nonviolence can turn out to be "true negative" or "false negative" predictions. The results of a study of violence prediction then are succinctly summarized by knowing four values:

- *E* = the number of true positives
- *F* = the number of false positives
- *G* = the number of false negatives
- *H* = the number of true negatives

From *E*, *F*, *G*, and *H*, one can calculate all the accuracy indices listed below the 2×2 contingency matrix in Table 1. To illustrate how *Tarasoff*-era studies evaluated and characterized accuracy, let us review data from a study by Kozol and colleagues[102] that *Tarasoff* cited.[103]

Kozol and colleagues examined outcomes of 592 convicted men sent to the Center for the Diagnosis and Treatment for Dangerous Persons in Bridgewater, Massachusetts. Clinicians at the center reached initial conclusions that 304 of these men were not dangerous. These men were therefore released into the community after completing their sentences, and twenty-six of them later committed serious crimes. Courts agreed with clinicians' diagnoses of dangerousness in 226 cases and committed these men to the center for treatment. Following treatment, eighty-two patients were discharged on recommendation of the clinical staff, and five later committed serious crimes. Courts also released forty-nine committed patients against advice of clinicians, and seventeen of these men later committed serious crimes.

102. Harry L. Kozol et al., *The Diagnosis and Treatment of Dangerousness*, 18 CRIME & DELINQ. 371, 390 (1972).

103. *Tarasoff II*. 551 P.2d 334, 360 (Cal. 1976) (Clark, J., dissenting). Kozol and colleagues' study was an important data source in Professor Monahan's famous monograph, JOHN MONAHAN, THE CLINICAL PREDICTION OF VIOLENT BEHAVIOR 44, 48 (1981). Their study continues to be cited in legal documents. *See, e.g.*, Heller v. Doe, 509 U.S. 312, 323 (1993); SHIRLEY A. DOBBIN & SOPHIA I. GATOWSKI, A JUDGE'S DESKBOOK ON THE BASIC PHILOSOPHIES AND METHODS OF SCIENCE 200 (1999), *available at* http://www.unr.edu/bench; and Alexander Scherr, *Daubert & Danger: The "Fit" of Expert Predictions in Civil Commitments*, 55 HASTINGS L. J. 1, 90 (2003).

Table 2. — Data and interpretations of findings, based on Kozol and colleagues' data.[104]

Recommendations (Treated as Predictions):	Discharged Patients' Actual Behavior: "Serious Crime" = Violent (*V*+)	Discharged Patients' Actual Behavior: No "Serious Crime" = Not violent (*V*–)	*Sums:*
Do not release = Will be violent (*T*+)	17 = true positives	32 = false positives	49
Release = Will not be violent (*T*–)	31 = false negatives	355 = true negatives	386
Sums:	48	387	435

$$\text{base rate} = BR = \text{probability of violence} = P(V+) = \frac{48}{435} = 0.110$$

$$\text{correct fraction} = \frac{17+355}{435} = 0.855$$

$$\text{percent correct} = 85.5\%$$

$$\text{sensitivity} = \text{true positive rate } (TPR) = P(T+|V+) = \frac{17}{48} = 0.354$$

$$\text{specificity} = \text{true negative rate } (TNR) = P(T-|V-) = \frac{355}{32+355} = 0.917$$

$$\text{false positive rate } (FPR) = P(T+|V-) = \frac{32}{32+355} = 0.083$$

$$\text{positive predictive value} = PPV = P(V+|T+) = \frac{17}{17+32} = 0.347$$

$$\text{negative predictive value} = NPV = P(V-|T-) = \frac{355}{31+355} = 0.920$$

For purposes of my (and Professor Monahan's[105]) analysis, the relevant subjects were the 304+82+49 = 435 patients about whom clinicians had made judgments and who had been released to the community. Committing a serious crime following release constituted being violent. A clinical recommendation for release was treated as a prediction of no violence (a "negative" prediction), and a judgment that a patient should not be released was a prediction of violence (a "positive" prediction). The results allow preparation of Table 2.

104. *See supra* note 102.

105. MONAHAN, *supra* note 103, at 44, 48.

The various calculations in Table 2 permit several different interpretations of the Bridgewater clinicians' accuracy. The traditional medical indices of binary test accuracy are sensitivity and specificity. If one uses these indices to evaluate the Table 2 data, one concludes that the clinicians did not detect most of the violent patients (*i.e.*, their sensitivity was only 0.354), but they were very good at detecting nonviolent patients (*i.e.*, their specificity is 0.917). However, studies of violence prediction in the *Tarasoff* era often focused on another index—the positive predictive value (*PPV*). This index provides the answer to the question, "If a clinician predicted that an individual would be violent, what is the probability that the clinician was right?" Using the data from Kozol and colleagues, one sees that the response is, "About one-third of the time," or equivalently, "clinicians' predictions of violence were wrong about two-thirds of the time."

Because of findings and interpretations like these, discussions of violence prediction in the 1970s and 1980s reached these conclusions:

- "[P]sychiatrists have absolutely no expertise in predicting dangerous behavior—indeed, they may be *less* accurate predictors than laymen—and . . . they usually err by overpredicting violence."[106]
- "[T]he 'best' clinical research currently in existence indicates that psychiatrists and psychologists are accurate in no more than one out of three predictions of violent behavior over a several-year period among institutionalized populations that had both committed violence in the past . . . and who were diagnosed as mentally ill."[107]
- "In general, mental health professionals . . . are more likely to be wrong than right when they predict legally relevant behavior. When predicting violence, dangerousness, and suicide, they are far more likely to be wrong than right."[108]
- "The American Psychiatric Association (APA), participating in this case as *amicus curiae*, informs us that '[t]he unreliability of psychiatric predictions of long-term future dangerousness is by now an established fact within the profession.' . . . The APA's best estimate is that *two out of three* predictions of long-term future violence made by psychiatrists are wrong. . . . [T]he APA's Draft

106. Bruce J. Ennis & Thomas R. Litwack, *Psychiatry and the Presumption of Expertise: Flipping Coins in the Courtroom*, 62 CAL. L. REV. 693, 734–35 (1974).

107. MONAHAN, *supra* note 103, at 47–49 (1981).

108. Stephen J. Morse, *Crazy Behavior, Morals, and Science: An Analysis of Mental Health Law*, 51 S. CAL. L. REV. 527, 600 (1978). Author's comment: I do not agree with everything Professor Morse writes, but after more than a quarter-century, this remains a *great* article.

> Report of the Task Force on the Role of Psychiatry in the Sentencing Process (1983) . . . states that '[c]onsiderable evidence has been accumulated by now to demonstrate that long-term prediction by psychiatrists of future violence is an extremely inaccurate process.'"[109]

Thus, at the time of the *Tarasoff* decisions and during the decade afterward, available research was taken to imply that mental health professionals could not predict violence over long periods of time. Moreover, mental health professionals also believed, based on then-available evidence, that individuals who suffered from mental disorders did not commit violence at rates greater than the rates of the general population, if one statistically "controlled" for those sociodemographic factors (*e.g.*, being young and poor) that were known to increase the risk of violence.[110] Thus, if the prevailing beliefs were correct, *Tarasoff* imposed a dual unfairness on psychiatric patients and their therapists. The decision stigmatized persons with mental disorders as especially dangerous and, if they sought treatment, left them subject to potentially embarrassing breaches of confidentiality. As for their therapists, *Tarasoff* insisted that mental health clinicians follow "professional standards" for detecting "serious danger" when, in fact, mental health professionals seemed to have no ability to do this.

B. Advances in the 1990s

1. Statistical Advances

In a highly influential 1984 article,[111] Professor Monahan opined that the apparently-dismal results from then-available violence prediction studies might reflect problems with the definition of violence, imprecise follow-up and ascertainment, and lengthy periods of time (usually years) covered by the studies. Professor Monahan hoped that a "second generation" of violence prediction studies, involving shorter time periods, better technology, and better identification and quantifying of

109. Barefoot v. Estelle, 463 U.S. 880, 920 (1983) (Blackmun, J., dissenting).

110. John Monahan, *Mental Disorder and Violent Behavior: Perceptions and Evidence*, 47 AM. PSYCHOLOGIST 511, 512–13 (1992) (summarizing previous beliefs and giving examples of sources); Elizabeth Walsh & Thomas Fahy, *Violence in Society*, 325 BRIT. MED. J. 507, 507 (2002) (citing H. Hafner & W. Boker, *Mentally Disordered Violent Offenders*, 8 SOC. PSYCHIATRY 220, 220 (1973) (finding that "crimes of violence committed by 533 mentally ill and mentally retarded offenders were quantitatively proportional to the number of crimes of violence committed by the total population.")).

111. John Monahan, *The Prediction of Violent Behavior: Toward a Second Generation of Theory and Policy*, 141 AM. J. PSYCHIATRY 1, 10 (1984).

violence, might show that mental health professional had at least *some* ability to make predictions.[112] Indeed, some studies from the late 1980s led researchers to conclude that, in contrast to long-term predictions, short-term predictions—those covering a period of a few days, a time period relevant, for example, to civil commitment—had a "high degree of . . . predictive validity."[113]

By the late 1990s, however, mental health professionals had developed new views about the accuracy of violence predictions, which led Professor Monahan to conclude that, contrary to what had seemed true a decade before, "clinicians are able to distinguish violent from nonviolent patients with a modest, better-than-chance level of accuracy."[114] The scientific results that supported this new position did not involve new empirical findings, but a reinterpretation of available data. This reinterpretation involved application of techniques for quantifying diagnostic accuracy that had been used in radiologic studies since the 1970s[115] and that started appearing in mental health professionals' publications in the late 1980s.[116] Underlying these techniques is the recognition that one can have varying levels of confidence in whether an either/or event will occur, and that a proper

112. *Id*. at 10.

113. Dale E. McNeil & Renée L. Binder, *Predictive Validity of Judgments of Dangerousness in Emergency Civil Commitment,* 144 AM. J. PSYCHIATRY 197, 197 (1987).

114. John Monahan, *Clinical and Actuarial Predictions of Violence, in* 1 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (D. Faigman et al. eds., 1997) § 7-2.2[2], at 317. *See also* John Monahan, *The Scientific Status of Research on Clinical and Actuarial Predictions of Violence*, SCIENCE IN THE LAW: SOCIAL AND BEHAVIORAL SCIENCE ISSUES § 2-2.2.2, at 110–11 (D. Faigman et al. eds., 2002) [hereinafter Monahan, *Scientific Status*], providing "[f]our brief quotations [that] are representative of a growing consensus among researchers on risk assessment of violence" supporting the view that violence predictions are accurate.

115. *See, e.g.*, D. J. Goodenough et al., *Radiographic Applications of Receiver Operating Characteristic (ROC) Curves*, 110 RADIOLOGY 89 (1974) (discussing the role of ROC analysis in quantifying radiologists' accuracy); and Barbara J. McNeil et al., *Measures of Clinical Efficacy. Cost-Effectiveness Calculations in the Diagnosis and Treatment of Hypertensive Renovascular Disease*, 293 NEW ENG. J. MED. 216 (1975) (discussing plotting of renogram data as ROC curves).

116. Research psychologists outside the mental health field had recognized the potential value of ROC analysis. *See, e.g.,* Richard C. Atkinson, *A Variable Sensitivity Theory of Signal Detection*, 70 PSYCHOL. REV. 91 (1963) (discussing theoretical prediction of ROC curves in detecting signals in psychophysiological experiments); and*, especially,* DAVID M. GREEN & JOHN A. SWETS, SIGNAL DETECTION THEORY AND PSYCHOPHYSICS (1966) (a classic text in the field). It was much later, however, that mental health clinicians recognized the value of ROC methods for describing diagnostic accuracy. Three of the earliest such reports are Harold P. Erdman et al., *Suicide Risk Prediction by Computer Interview: A Prospective Study*, 48 J. CLINICAL PSYCHIATRY 464 (1987); Jane M. Murphy et al., *Performance of Screening and Diagnostic Tests: Application of Receiver Operating Characteristic Analysis*, 44 ARCHIVES GEN. PSYCHIATRY 550 (1987); and Douglas Mossman & Eugene Somoza, *Maximizing Diagnostic Information from the Dexamethasone Suppression Test: An Approach to Criterion Selection Using Receiver Operating Characteristic Analysis*, 46 ARCHIVES GEN. PSYCHIATRY 653 (1989).

description of detection accuracy will reflect these varying levels.

a. An Imaginary Study

To illustrate how these techniques work, consider the following dilemma face by the imaginary Dr. Jones, the unfortunate superintendent of the imaginary, 500-bed Farblundget State Psychiatric Hospital (FSPH). Dr. Jones receives instructions telling him that, due to budgetary cutbacks, he must downsize the hospital by quickly releasing a substantial portion of its inpatients. I describe Dr. Jones as "unfortunate" for two reasons. First, the downsizing directive comes with the additional command not to release any "dangerous" patients. Faced with no alternative, Dr. Jones and his clinical staff set about to identify those patients who, they hope, have the best chance of not doing anything violent if they leave the hospital.

After a week's effort, the FSPH staff have identified 163 patients who, they believe, have a low likelihood of doing violence if released to the community. Then, Dr. Jones learns of the second reason why he is unfortunate: budgetary constraints will require that FSPH close, and all 500 of its patients will be released.

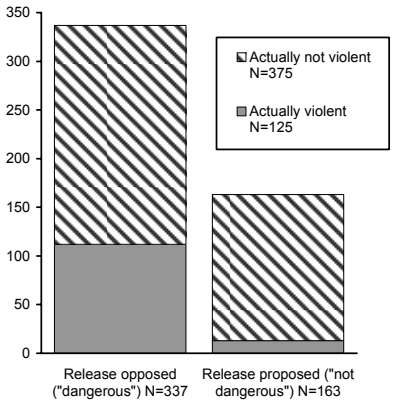
Although this announcement is bad news for the FSPH staff who will lose their jobs (and, if one believes that psychiatric hospitalization is helpful, for the patients of FSPH), it is good news for social scientists, who now have "an excellent opportunity for naturalistic research"[117] that will help them learn how accurately clinicians can predict violence. They decide to follow the former FSPH patients for the first year after their release, meeting with the former patients, talking with family members, and checking their police records to see whether they commit any violent acts.[118] In this study, any confirmable report of striking, physically fighting with, or doing physical harm to another person will identify a former patient as having been "violent."

Twelve months after release, 125 former FSPH patients—one-fourth of those released—have committed acts that define them as "violent." The first set of results from the study appear in Figure 1. Using these data, one can ask, "What is the likelihood that a patient whom the clinicians thought was too dangerous to release actually became violent?"

117. MONAHAN, *supra* note 103, at 46. The quoted phrase is language Professor Monahan used to describe research on the transfer, following *Baxtrom v. Herold*, 383 U.S. 107 (1966), of residents from former hospitals for the criminally insane to civil psychiatric hospitals.

118. The MacArthur Violence Risk Assessment Study used a similar method for following former inpatients in the community and detecting whether they had acted violently. *See* Steadman et al., *supra* note 86, at 394 (contact every ten weeks with former inpatients and collateral informants).

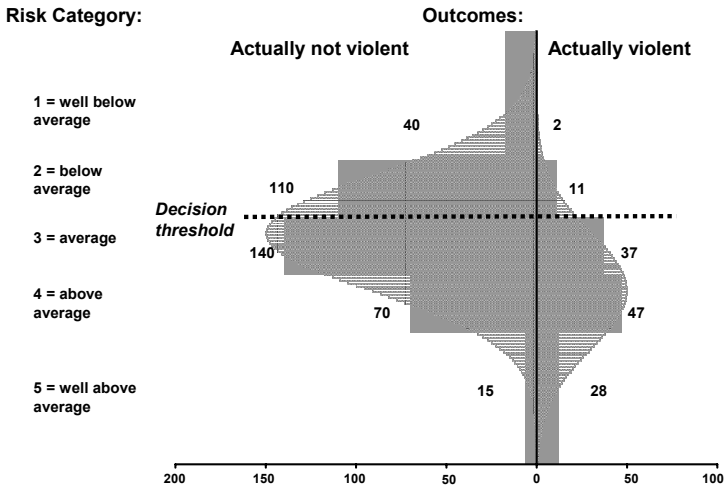
Figure 1. — Results of the imaginary study of former FSPH patients, showing that about two-thirds of the patients who were not recommended for release were not violent in the year following the hospital's closure.



The answer (consistent with the study by Kozol and colleagues and the overall gist of Monahan's findings) was 112 out of 337 cases, or "about one-third of the time," which might lead to the interpretation that these clinicians "overpredicted" violence. Suppose one next asks how often the clinicians were correct. The data in Figure 1 suggest that whey were "right" about the 112 actually violent former patients whose release they opposed (the "true positives") and the 150 actually nonviolent former patients whose release they recommended (the "true negatives"). The FSPH clinicians thus correctly categorized 262 (about 52%) of the patients. Notice, though, that analyzed in this way, simply recommending that all patients be released would have yielded a higher correct percentage—75%. It thus seems as though the clinicians made more errors than they would have by simply accepting the base rate. Indeed, it often is the case that, when base rates of a phenomenon are low, one often can get more answers right by simply guessing that the phenomenon will not occur than by trying to figure out whether the phenomenon will occur.

But the FSPH study has access to some additional data about the FSPH clinicians' decision-making, and Figure 2 summarizes these data. In making their judgments about dangerousness, FSPH clinicians conducted a risk assessment that ultimately placed patients in one of five levels of dangerousness. This meant that they could rank patients in five categories according to their perceived likelihood of being violent, from 1 = "well below average" risk to 5 = "well above average" risk.

Figure 2. — Outcomes for imaginary FSPH patients by risk category. Notice that those patients who actually were violent tended to have had higher risk ratings than patients who were not violent following release. The Gaussian (normal, or "bell-shaped") distributions overlying the data suggest that violent patients fell about one standard deviation higher on the FSPH clinicians' latent decision axis.



In other words, the clinicians had five levels of belief about patients' likelihood of post-hospitalization violence. The clinicians' discharge recommendations, which reflected their mandate not to release patients who were "dangerous," translated into proposals to discharge only those who had below-average levels of risk. The result was a discharge policy that was highly sensitive to violence at the expense of specificity in making recommendations about who should leave. Numerically, this sensitive-and-cautious policy is expressed in the finding that nearly 90% of the violent patients would not have been released had FSPH downsized but not closed, but only 40% of the nonviolent patients would have returned to the community. Put another way, the clinicians' perceived mandate made release of a dangerous (violent) patient—a "false negative" error—much worse than retention of a nondangerous (nonviolent) patient—a "false positive" error, and the clinicians responded accordingly.

b. ROC Analysis

The statistical advance of the 1990s involved a recognition that clinicians' decision thresholds could and should be separated from their judgments about levels of dangerousness, and that the accuracy of violence predictions should be judged using statistical methods that separate effects of base rates and decision thresholds from intrinsic detection capabilities. The statistical tool for accomplishing this is receiver operating characteristic (ROC) analysis.[119] Two articles[120] used this method to quantity prediction accuracy shortly after my recommendation, in 1994, that studies of violence prediction do so.[121] As a way of demonstrating the value of ROC methods, I had reanalyzed a representative sample of existing data from both "first-" and "second-generation" studies on violence prediction. The results led me to conclude that, in contrast to what courts and mental health publications had claimed, "clinicians are able to distinguish violent from nonviolent patients with a modest, better-than-chance level of accuracy," and that short-term predictions covering a week's time were not more accurate than predictions that covered periods of a year or more.[122] By the middle of the 21st century's first decade, ROC indices had become the standard way that investigators reported prediction accuracy in studies of violence prediction and of tools for predicting recidivism by sex offenders.

Figure 2 depicts the distribution of patients in the clinicians' five-category ratings as rectangular areas extending to the left (for nonviolent patients) or right (for violent patients) of the vertical axis. Superimposed on the rectangular distributions are two Gaussian ("normal" or "bell-shaped") distributions. ROC analyses of prediction data often use a plot such as the one shown in Figure 3 to depict key features of these types of data.

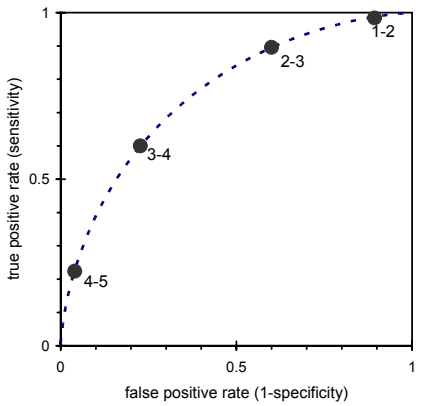
119. "Receiver operating characteristic" analysis reflects an early use of this statistical method—describing the accuracy of radar during World War II. The notion is that radar detection capabilities (the ratio of "hits" to "false alarms") are characterized by the decision threshold at which the receiver operates. Larry B. Lusted, *ROC Recollected*, 4 MED. DECISION MAKING 131 (1984).

120. Marnie E. Rice & Grant T. Harris, *Violent Recidivism: Assessing Predictive Validity,* 63 J. CONSULTING CLINICAL PSYCHOLOGY 737 (1995); William Gardner et al., *Clinical Versus Actuarial Predictions of Violence in Patients with Mental Illness*, 64 J. CONSULTING CLINICAL PSYCHOLOGY 602 (1996).

121. Mossman, *Accuracy*, *supra* note 86; Douglas Mossman, *Further Comments on Portraying the Accuracy of Violence Predictions*, 18 LAW HUM. BEHAV. 587 (1994).

122. Mossman, *Accuracy*, *supra* note 86, at 790.

Figure 3. — ROC graph depicting the accuracy of the FSPH clinicians. Large circles are cut-offs that would be formed by placing decision thresholds at the boundaries between the five risk categories in Figure 2. The curve (dashed line) linking the cut-offs uses the "binormal assumption" of ROC analysis.



From looking at Figure 2, one sees that the five risk categories used by the FSPH clinicians actually gave them four potential decision thresholds (in addition to keeping everyone or discharging everyone) for making release recommendations. The threshold that the clinicians chose put them close to the desired goal of reducing the hospital population by one-third, but it also meant that they set their apparent sensitivity at 112/125 = 0.896 and their specificity at 150/375 = 0.400. Had they proposed discharging only patients whose dangerousness was "well below average," the effect of this threshold would be to increase sensitivity to 123/125 = 0.984, but to reduce specificity to 40/375 = 0.107. Choosing other decision thresholds would have increased specificity at the cost of sensitivity.

Figure 3 contains a ROC graph, which depicts the accuracy of the FSPH clinicians as a set of trade-offs between sensitivity and specificity as a detection method's decision threshold is moved throughout the entire range of possibilities. It is customary to do this by plotting the true positive rate of a detection method (equal to sensitivity) as a function of the false positive rate of the detection method (equal to 1 – specificity). The large circles in Figure 3 represent the four cut-offs that would be formed by placing decision thresholds at the four boundaries between the FSPH clinicians' five categories of risk shown in Figure 2. The smooth ROC curve (in Figure 3, the dashed line) linking these four cut-offs derives from the "binormal assumption" of ROC analysis.

In the context of violence prediction, the binormal assumption suggests that discrimination capacity can be succinctly explained by assuming that the violence prediction method partially separates violent and nonviolent individuals along a continuous, latent decision axis.[123] Along this axis, the violent and nonviolent populations form two overlapping, "normal" distributions with different means (or average values) and standard deviations (a statistical term that describes how "spread out" a distribution is). For our imaginary FSPH study, the binormal assumption states that we can think about the clinicians as potentially having the ability to rate each patient's violence risk along a continuum of risk—or, at least, along a latent decision axis with many, many gradations in risk levels.

The binormal assumption helps us to recognize that, although intrinsic discriminatory power was limited by the clinicians' predictive abilities, the five categories of risk that the FSPH clinicians chose to use, and the boundaries (thresholds) between these categories, were somewhat arbitrary. The clinicians might have used three- or seven-category classifications of risk, and if they had done so, the number and location of the boundaries (thresholds) between categories would have had different locations along the underlying decision axis.

ROC analysis gives investigators several ways of summarizing prediction accuracy, but for our purposes, we shall focus on three of these. First, the area under the curve (AUC) is a simple summary index of accuracy[124] that, in the present context, has this practical interpretation: AUC equals the probability that the prediction method would give a randomly selected, actually violent person a higher score than a randomly selected, nonviolent person. A perfect violence prediction method—one that always would give a randomly chosen violent person a higher score than a randomly chosen nonviolent person—would have an AUC of 1.0. A violence prediction method that gave no information about future behavior—that is, a method that did no better than a coin toss at distinguishing a violent person from a nonviolent person—would have an AUC of 0.5. For the clinicians at

123. Concerning the notion of a latent decision axis and the bi-normal model discussed here, *see* Charles E. Metz et al., *Maximum Likelihood Estimation of Receiver Operating Characteristic (Roc) Curves from Continuously-distributed Data,* 17 STAT. MED. 1033, 1037 (1998). A highly technical discussion of problems related to this concept appears in Donald D. Dorfman & Kevin S. Berbaum, *A Contaminated Binormal Model for ROC Data: Part II. A Formal Model*, 7 ACAD. RADIOLOGY 427 (2000). The classic article on this topic is Donald D. Dorfman & Edward Alf Jr., *Maximum Likelihood Estimation of Parameters of Signal Detection Theory and Determination of Confidence Intervals—Rating Method Data*, 6 J. MATHEMATICAL PSYCHOL. 487 (1969).

124. James A. Hanley & Barbara J. McNeil, *The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve*, 143 RADIOLOGY 29 (1982).

FSPH, AUC = 0.76, which is a fairly typical result in recent studies of methods used to predict or detect violence.[125]

A second way to characterize prediction accuracy makes reference to the *effect size* of the prediction method. Looking at the bell-shaped distributions in Figure 2, we see that the distribution for violent patients is displaced downward compared to the distribution of nonviolent patients; in this case, it turns out that the downward displacement equals one standard deviation. To put this another way: the size of the effect of the FSPH clinicians' assessment is equivalent to separating the distributions of violent and nonviolent patients by one standard deviation,[126] or more simply, the effect size equals 1.

A third way of describing accuracy utilizes the binormal assumption and refers to the locations of the normal distributions along the latent decision axis. If we assign the mean and standard deviation of the nonviolent population the values of 0 and 1, respectively, then we can express the accuracy of a detection system (in this case, a system for detecting violent patients) using the following linear equation:

[1] $$Z_{TPR} = A + B \cdot Z_{FPR}$$

In Equation 1, Z_{TPR} and Z_{FPR} are the normal deviates, or *z*-transforms, of the true and false positive rates, respectively; A equals the distance between the means of the violent and nonviolent populations, measured in units of the standard deviation of the nonviolent population; and B equals the ratio of the standard deviations (*SD*) of the nonviolent and violent populations (*i.e.*, SD_{V-}/SD_{V+}). When B=1 (as is the case for the

125. *See, e.g.*, Marnie E. Rice & Grant T. Harris, *Violent Recidivism: Assessing Predictive Validity*, 63 J. CONSULTING CLINICAL PSYCHOL. 737 (1995) (AUC = 0.76 for re-offending); Kevin S. Douglas et al., *Assessing Risk for Violence Among Psychiatric Patients: The HCR-20 Violence Risk Assessment Scheme and the Psychopathy Checklist: Screening Version*, 67 J. CONSULTING CLINICAL PSYCHOL. 917 (1999) (AUC = 0.76 for any physical violence); Michael J. Furlong & Michael P. Bates, *Predicting School Weapon Possession: A Secondary Analysis of the Youth Risk Behavior Surveillance Survey*, 38 PSYCHOL. SCH. 127 (2001) (AUC = 0.75 for weapons possession at school); Kevin S. Douglas & James R. Ogloff, *Violence by Psychiatric Patients: The Impact of Archival Measurement Source on Violence Base Rates and Risk Assessment Accuracy*, 48 CAN. J. PSYCHIATRY 734, 738 (2003) (AUCs of 0.72 to 0.80 for community violence, variously defined).

126. Often, studies refer to the effect size as "Cohen's *d*," acknowledging a frequently cited book concerning this statistic. *See* Jacob Cohen, STATISTICAL POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES (1969); Jacob Cohen, *Quantitative Methods in Psychology: A Power Primer*, 112 PSYCHOL. BULL. 155 (1992). One can convert Cohen's *d* to another statistic, the common language effect size, or *CL*. K. O. McGraw & S. P. Wong, *A Common Language Effect Size Statistic*, 111 PSYCHOL. BULL. 361 (1992). Under most circumstances, *CL* is roughly equal to AUC. For further discussion of these relationships, *see* Marnie E. Rice & Grant T. Harris, *Comparing Effect Sizes in Follow-up Studies: ROC Area, Cohen's d, and r*, 29 L. HUM. BEHAV. 615 (2005).

FSPH data[127]), then *A* equals the effect size.

2. The Association Between Violence and Mental Disorder

As was noted earlier, the limited evidence available in the 1970s and 1980s suggested no association between mental illness and violence, once one factored in sociodemographic factors that were known to raise violence risk. However, the 1990s witnessed the accumulation of scientific studies that showed that mental illness was a definite risk factor for acting violently, though a relatively minor one.

The first of these study appeared in July 1990. Swanson and colleagues[128] used data originally gathered in the 1980s for the Epidemiologic Catchment Area studies to calculate the rates of various mental disorders in the United States.[129] While gathering these data, researchers had also asked participants about whether they had committed various types of violent acts in the year before the interview. Reanalyzing these data allowed Swanson and colleagues to show that the presence of substance use problems or serious mental illnesses

127. I created data such that *B*=1 for this Article so as to simplify, without loss of generality, some subsequent calculations. For further discussion of the bi-normal assumption in this context, including values of *B*, *see* Mossman, *Accuracy*, *supra* note 86, at 785, 788; and Douglas Mossman, *Dangerousness Decisions: An Essay on the Mathematics of Clinical Violence Predictions and Involuntary Hospitalization*, 2 U. CHI. L. SCH. ROUNDTABLE 95, 116–17, nn. 67, 68 (1995) [hereinafter Mossman, *Dangerousness Decisions*].

128. Jeffrey W. Swanson et al., *Violence and Psychiatric Disorder in the Community: Evidence from the Epidemiologic Catchment Area Surveys*, 41 HOSP. & COMMUNITY PSYCHIATRY 761, (1990).

129. The Epidemiologic Catchment Area (ECA) collected data in 1980–85 on the prevalence and incidence of mental disorders in five U.S. areas: New Haven, Baltimore, St. Louis, Durham (North Carolina), and Los Angeles. At each site, data gatherers interviewed more than 3,000 community residents and 500 residents of institutions, yielding more than 20,000 respondents overall. Interviewers used the Diagnostic Interview Schedule (DIS), which had been developed by the National Institute of Mental Health to permit trained nonprofessionals to obtain information about several psychiatric disorders, including mania, depression, schizophrenia, alcohol dependence, and antisocial personality. For a fuller description, *see* Daniel A. Regier et al., *The NIMH Epidemiologic Catchment Area Program: Historical Context, Major Objectives, and Study Population Characteristics,* 41 ARCHIVES GEN. PSYCHIATRY 934 (1984).

Among the items contained in the DIS were the following five questions:

- Did you ever hit or throw things at your wife/ husband/partner? [If so] were you ever the one who threw things first, regardless of who started the argument? Did you hit or throw things first on more than one occasion?
- Have you ever spanked or hit a child (yours or anyone else's) hard enough so that he or she had bruises or had to stay in bed or see a doctor?
- Since age 18, have you been in more than one fight that came to swapping blows, other than fights with your husband/wipe/partner?
- Have you ever used a weapon like a stick, knife, or gun in a fight since you were 18?
- Have you ever gotten into physical fights while drinking?

Monahan, *Scientific Status*, *supra* note 114, at 94 n.11.

increased violence rates among adults, even after one controlled statistically for a person's sex, age, and income level.

In 1992, Link and colleagues[130] published findings from another study that took advantage of previously gathered, archival data. The nature of these data allowed Link and colleagues to use statistical methods to control for several demographic variables, even including the neighborhood where an individual lived. After factoring in this more extensive set of variables, Link and colleagues still could show that having been a "mental patient"[131] statistically increased a person's risk of violence.[132]

Several similar studies followed, with the result being that, by the end of the 1990s, mental health professionals could no longer deny that psychiatric disorders were an independent risk factor for acting violently. One could certainly point to groups of people with higher rates of violence than persons with psychiatric disorders,[133] and one could certainly point to other common social factors (e.g., violence on television and other visual media) with solid causative links to violence.[134] Nonetheless, the link between mental disorder and violence that formed the implicit rationale of *Tarasoff* and its progeny had achieved a statistical vindication.

3. Changes in Risk Assessment "Technology"

The 1990s also witnessed the beginnings and dissemination of a new approach to, or "technology" for, making violence risk assessments.

---

130. Bruce G. Link et al., *The Violent and Illegal Behavior of Mental Patients Reconsidered*, 57 AM. SOC. REV. 275 (1992).

131. The data on this study's "mental patients" came from inpatient and outpatient psychiatric patients receiving treatment at the Columbia-Presbyterian Medical Center. *Id.* at 279.

132. *Id.* at 285–288 (table shows predictive value of having been a patient, despite controlling statistically for other variables).

133. Professor Monahan summarizes:

> The policy implications of mental disorder as a risk factor for violent behavior can be understood only in relative terms. Compared to the magnitude of risk associated with the combination of male gender, young age, and lower socioeconomic status, for example, the risk of violence presented by mental disorder is modest. Compared to the magnitude of risk associated with alcoholism and other drug abuse, the risk associated with "major" mental disorders . . . is modest indeed. Clearly, mental health status makes at best a trivial contribution to the overall level of violence in society.

Monahan, *Scientific Status*, *supra* note 114, at 108–09.

134. *See, e.g.*, Craig A. Anderson et al., *The Influence of Media Violence on Youth*, 4 PSYCHOL. SCI. PUB. INT. 81 (2003).

Before 1990, most studies of psychiatric violence prediction examined efforts in which mental health professionals had used their "clinical judgment" to gauge violence and make decisions relevant to patients. In the context of violence prediction, the exercise of "clinical judgment" refers to a process in which a decision-maker combines data mentally ("in his head") when assessing a person's risk of becoming violent. In theory at least, clinical judgments about dangerousness incorporate a psychiatrist's or psychologist's professional knowledge, personal experience, "gut" feelings, intuitions about the evaluee, and whatever other information about the situation that seems relevant to the assessment. Clinical judgment is the method physicians usually use to evaluate and treat patients: doctors listen to patients and examine them, think about and decide what probably is wrong, then recommend treatments.[135]

Life insurance actuaries do not use clinical judgment when making assessments of an individual's life-span. Instead, they use formulae, tables, algorithms, or other pre-specified ways of combining information to estimate life expectancy. For this reason, psychologists have adopted the term "actuarial" to describe methods of prediction that resemble the processes an insurance actuary uses to formulate a judgment about the risk of a future event.[136] Actuarial prediction techniques are established, and their claims to accuracy rest upon, empirically determined relationships between specific types of data and the event to be predicted. Because actuarial methods of judgment use fixed, predetermined, empirically based techniques to combine data and render assessments of future events, they are also called "statistical," "mechanical," or "formal" methods for making judgments.[137]

In the 1990s, mental health literature witnessed the publication of several studies that described the capacities of actuarial "technology" in assessing the risk of future violence. This technology typically requires clinicians to gather information about ten to twenty factors concerning the individuals undergoing evaluation. The clinicians then score that information about each factor using an instruction manual or some other

135. Although we do not usually think about our doctors' actions as predictions, it is reasonable to do so. When they select treatments, our doctors are making judgments that the proposed treatment will (or is very likely to) remedy our ailments.

136. Robyn M. Dawes et al., *Clinical Versus Actuarial Judgment*, 243 SCI. 1668 (1989). For a non-insurance example, *see* Sei J. Lee et al., *Development and Validation of a Prognostic Index for 4-Year Mortality in Older Adults*, 295 JAMA 801 (2006).

137. William M. Grove & Paul E. Meehl, *Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy*, 2 PSYCHOL. PUB. POL'Y & L. 293 (1996).

pre-specified method. This process generates a numerical value or category that describes the evaluees' risk of violence.

By now, the social science literature contains scores of studies of actuarial prediction methods yielding results that imply well-above-chance levels of predictive accuracy. Some writers, partly on the basis of research findings demonstrating the superiority of actuarial judgments in a host of other types of predictions,[138] interpret available violence prediction research as showing that actuarial measures are clearly superior to clinical judgments in predicting future violence.[139] Indeed, one research group advocates for "the complete replacement of existing practice [*i.e.*, using clinical judgment] with actuarial methods."[140] This recommendation probably is "premature"[141] and ignores the limitations of currently available tools to implement actuarial judgment.[142] Yet even skeptical commentators recognize that available evidence shows that actuarial judgment "can enhance a variety of dangerousness risk assessments," and clinicians who perform "risk assessments have a professional responsibility to be aware of the advantages and limitations of using such risk assessment tools."[143]

The available research also shows that actuarial measures are far from perfect at distinguishing those individuals who will be violent from those who will not. As was the case with the imaginary FSPH study, real-life studies of actuarial violence prediction report AUCs of 0.70 to 0.80. That is, studies of actuarial judgment show that the score distributions of violent and nonviolent individuals have considerable overlap. True, the scores of violent subjects are, on average, higher than the scores of nonviolent subjects, so that, as the score increases, the probability of violence increases. But this means that a user of a

138. William M. Grove et al., *Clinical Versus Mechanical Prediction: A Meta-Analysis*, 12 PSYCHOL. ASSESSMENT 19 (2000).

139. *E.g.*, Mossman, *Accuracy*, *supra* note 86, at 789 (predictions using past behavior and discriminant functions appear better than clinical judgment); William Gardner et al., *Clinical Versus Actuarial Predictions of Violence in Patients With Mental Illnesses,* 64 J. CONSULTING CLINICAL PSYCHOL. 602, 602 (1996) ("Actuarial predictions based only on patients' histories of violence were more accurate than clinical predictions, as were actuarial predictions that did not use information about histories."); and Marnie E. Rice, et al., *The Appraisal of Violence Risk*, 15 CURRENT OPINION PSYCHIATRY 589, 589 (2002) ("evidence favoring actuarial methods for appraising the risk of violence is increasing").

140. VERNON L. QUINSEY ET AL., VIOLENT OFFENDERS: APPRAISING AND MANAGING RISK 171 (1998).

141. Thomas R. Litwack, *Actuarial Versus Clinical Assessments of Dangerousness*, 7 PSYCHOL. PUB. POL'Y & L. 409, 410 (2001).

142. I discuss some of these limitations in Douglas Mossman, *Another Look at Interpreting Risk Categories,* SEXUAL ABUSE: J. RES. TREATMENT (forthcoming 2006).

143. Litwack, *supra* note 141, at 438.

violence risk assessment must recognize that a decision to take some course of action comes not from the risk assessment itself, but from a judgment about what level of risk should trigger that course of action.

A patient either will or will not act violently, and a therapist treating that patient can either take or not take a course of action to intervene. But, as is the case with our knowledge about most future events, a therapist's predictive knowledge about future violence is really an ability to make risk estimates or to assess the likelihood of violence. Probabilities, however, cannot tell us what level of violence risk justifies a course of preventive action. Whether to act, given a particular probability of violence, is something the therapist must decide independently.

V. IMPLICATIONS OF A DECISION THRESHOLD

A. Implications of 1990s Advances

The notion that a therapist's foreknowledge of a patient's dangerousness is represented by degrees or levels of violence risk, rather than a yes-or-no judgment about whether violence will occur, has unavoidable consequences for the implementation of risk assessments. The *Tarasoff* decision recognizes that therapists exercising their best professional judgment may still make prediction mistakes,[144] which could include false positive errors that either wrongfully attribute "serious danger" to persons who will not become violent,[145] and false negative errors that "miss" patients who later act violently. Yet, in telling therapists neither to "routinely . . . reveal such threats" nor to "disclose a confidence unless such disclosure is necessary to avert danger to others,"[146] *Tarasoff* fails to recognize that therapists do not have yes-or-no advance knowledge about whether a threat (or other behavior) implies that a disclosure is necessary because of future violence. At best, therapists know about probabilities or (more often) degrees of relative risk. This means that what must trigger a therapist's

144. *See Tarasoff II*, 551 P.2d 334, 345 (Cal. 1976) (recognizing "the difficulty that a therapist encounters in attempting to forecast whether a patient presents a serious danger of violence. . . . [P]roof, aided by hindsight, that he or she judged wrongly is insufficient to establish negligence.").

145. In his dissent, Justice Clark notes that, "under existing psychiatric procedures only a relatively few receiving treatment will ever present a risk of violence, the number making threats is huge, and it is the latter group—not just the former—whose treatment will be impaired and whose risk of commitment will be increased" by the majority's ruling. *Tarasoff II*, 551 P.2d at 360 (Clark, J., dissenting).

146. *Id*. at 347.

decision to take protective action can only be the therapist's perception of a sufficient likelihood that violence will occur—not a yes-or-no determination by the therapist that a "serious danger of violence" has presented itself.[147] As Dr. Paul Appelbaum has noted, *Tarasoff* (like the decisions that have followed it) implies that the duty to protect arises "only when a threshold of probability is crossed."[148]

What, then, constitutes a "threshold of probability" that tells a therapist what likelihood of violence is "sufficient" to warrant protective action? Here is Dr. Appelbaum's answer: "the terms used to define that threshold have varied, and never has it been specified with any precision."[149]

B. Balancing in Law and in Dangerousness Decisions

Figure 2 provides some clues about the type of information one might use in an attempt to find a decision threshold. As we saw earlier, each of the four boundaries between the FSPH clinicians' five risk categories could function as a threshold (or cut-off point) for making a release decision. Associated with each threshold is a specific set of correct prediction rates (that is, true positive and true negative rates) and incorrect prediction rates (false positive and false negative rates). These prediction rates, in turn, reflect three factors: (1) the number of rating categories the clinicians used (in this case, five), (2) the location, along the implicit decision axis, of the boundaries between those categories, and (3) the intrinsic accuracy of the clinicians' underlying decision technique, represented by the two normal distributions that appear in Figure 2, and summarized by Equation 1, with A=1 and B=1.

At least in theory, the FSPH clinicians could have used a larger number of categories in their risk classification,[150] or (again, in theory)

147. One can argue that I am intentionally misinterpreting, or at least being unfair to, the majority's opinion. One could say that, in requiring protective action when a "patient presents a serious danger of violence to another," *Tarasoff II,* 551 P.2d at 340, the majority implicitly acknowledges that some patients' statements or behaviors may represent what might be called "nonserious" or "less-than-serious" dangers. But this begs the question by assuming that a therapist's advance perception of risk cleanly distinguishes those risks that are "serious" and those that are not. Perception of risk is not a yes-or-no phenomenon, but a matter of degree, and no perceptual boundary defines those risks that are "serious."

148. Paul S. Appelbaum, *Ask the Experts*, 17 AM. ACAD. PSYCHIATRY L. NEWSL. 19, 19 (1992).

149. *Id.*

150. This is possible using various available assessment instruments. For example, the *HCR-20* has forty-one potential scores, because evaluees can obtain scores from 0 (lowest risk) to 40 (highest risk). For a review of the *HCR-20*, *see* Douglas Mossman, *Evaluating Violence Risk "By the Book": A Review of HCR-20: Assessing Risk for Violence, Version 2 and the Manual for the Sexual Violence Risk-20,* 18 BEHAV. SCI. & L. 781 (2000).

they could have used a different number of rating categories with potential boundaries different from those shown in Figure 2. In other words, the number of boundaries (or decision thresholds) along the axis is as large as the number of measurement gradations in the clinicians' risk assessment. Within the limits of these gradations, the FSPH clinicians would be free to fine-tune a decision threshold for recommending release so as to achieve whatever they believed was the optimum balance of correct predictions versus incorrect predictions. Here, the trade-offs involve monetary savings from discharging nonviolent patients and public safety gained by retaining violent patients (both good things), versus needless retention of nonviolent patients and release of violent persons (both bad things).

The notion that balancing errors should serve as a guide to imperfectly accurate legal decisions is well established in criminal law, where numerous cases discuss the appropriate ratio of wrongful acquittals to wrongful convictions.[151] Around the time of the *Tarasoff* decisions, a few commentators recognized that the duty to warn or protect the public raised a similar problem of balancing errors—for example, a balancing of wrongful decisions to involuntarily hospitalize patients out of a misplaced fear of future violence with wrongful decisions not to hospitalize patients who go on to do violence. Here are two examples of such considerations:

> Assume that one person out of a thousand will kill. Assume also that an exceptionally accurate test is created which differentiates with 95% effectiveness those who will kill from those who will not. If 100,000 people were tested, out of the 100 who would kill 95 would be isolated. Unfortunately, out of the 99,900 who would not kill, 4,995 people would also be isolated as potential killers. In these circumstances, it is clear that we could not justify incarcerating all 5,090 people. If, in the criminal law, it is better that ten guilty men go free than that one innocent man suffer, how can we say in the civil commitment area that it is better that fifty-four harmless people be incarcerated lest one dangerous man be

151. The best known ratio is Blackstone's (10:1), derived from the great commentator's oft-quoted view that "it is better that ten guilty persons escape, than that one innocent suffer." WILLIAM BLACKSTONE, 4 COMMENTARIES *352. Other historical authorities have suggested other ratios for wrongful acquittals and convictions. *See, e.g.*, MATTHEW HALE, PLEAS OF THE CROWN (reprint of 1678 ed., Oxford, Professional Books, 1972) (suggesting a 5:1 ratio); and J. FORTESCUE, A LEARNED COMMUNICATION OF THE LAWS OF ENGLAND (reprint of 1567 ed., New York, W. J. Johnson, 1969) (suggesting a 20:1 ratio). An amusing and thorough review of this issue in criminal law is provided by Alexander Volokh, *N Guilty Men*, 146 U. PA. L. REV. 173 (1997). Two examples of mathematical analyses of this issue are David H. Kaye, *Clarifying the Burden of Persuasion: What Bayesian Decision Rules Do and Do Not Do*, 3 INT'L J. EVIDENCE PROOF 1 (1999); and Stephen J. Ceci & Richard D. Friedman, *The Suggestibility of Children: Scientific Research and Legal Implications*, 86 CORNELL L. REV. 33 (2000).

free?[152]

> [I]t may be possible ethically to justify short-term commitment even if the predictions of imminent violence on which it is based are less accurate than the long-term research indicates. Paraphrasing Blackstone, it may be better that ten "false positives" suffer commitment for three days than that one "false negative" go free to kill someone during that period.[153]

Similar reasoning (though without any numerical specification) appears to underlie the "clear and convincing" standard of proof in civil commitment hearings established in *Addington v. Texas*:

> One who is suffering from a debilitating mental illness and in need of treatment is neither wholly at liberty nor free of stigma. . . . It cannot be said, therefore, that it is much better for a mentally ill person to "go free" than for a mentally normal person to be committed.[154]

Previous sections of this Article have explained the types of evidence showing that when violence "prediction" is redefined as an ability to rank relative risk of future violence, clinicians do better-than-chance at ranking individuals. As we have seen, though, that evidence also shows that rankings are imperfect, and that decisions based on such rankings inevitably involve errors. Can we find a way to decide upon an optimal balancing of those errors? For example, we might ask, as Professor Monahan and colleagues have, "How many safe people should be hospitalized as 'dangerous' to prevent discharging one patient who turns out to be violent?" In contrast to what we find in the criminal law, "No court has ever answered that question with a number. Judges are notoriously reluctant to set decision thresholds that depend on overt cost-

152. Dennis W. Daley, Comment, Tarasoff *and the Psychotherapist's Duty to Warn*, 12 SAN DIEGO L. REV. 932, 942–943 n.75 (1975).

153. John Monahan, *Strategies for the Empirical Analysis of the Prediction of Violence in Civil Commitment*, 1 L. HUM. BEHAV. 363, 370 (1977).

Monahan's approach is commonly applied to medical decisions (and other types of judgments), under terms such as the "preferred marginal tradeoff" or "number needed to treat." In medical contexts, the preferred marginal tradeoff usually

> is *the number of treatment errors* [*i.e.*, treatments of persons without disease] *that are acceptable in order to treat correctly one additional person with the disease.* In the framework of utility theory, the preferred marginal tradeoff is equivalent to the ratio of the net benefit of treating a diseased person to the net harm of treating a well person, so it is independent of disease prevalence.

Peter DeNeef & Daniel L. Kent, *Using Treatment-tradeoff Preferences to Select Diagnostic Strategies: Linking the ROC Curve to Threshold Analysis*, 13 MED. DECISION MAKING 126, 126 (1993) (emphasis in the original). An application of the number needed to treat metric appears in Alec Buchanan & Morven Leese *Detention of People with Dangerous Severe Personality Disorders: a Systematic Review*, 358 LANCET 1955, 1957 (2001).

154. Addington v. Texas, 441 U.S. 418, 429 (1979) (internal citations omitted).

benefit consideration[s], as are many other professionals and officials."[155]

C. What Is the Threshold?

1. Previously Published Research[156]

I believe that judges will never set a rational decision threshold because there is no social agreement about such a probability, and I have done two empirical studies to show that this is so. Both studies used the following reasoning.

Often, potentially violent persons with apparent mental problems are transported to psychiatric emergency rooms such as the one at University Hospital in Cincinnati. In the ER, it is difficult for patients to walk out, and therefore relatively easy for clinicians to fulfill the psychiatric duty to protect by arranging for involuntary hospitalization. Although this practice may benefit the public by confining persons who would otherwise commit violence in the community, it deprives persons who are hospitalized of their liberty. Because violence psychiatric predictions are not perfectly accurate, involuntary hospitalization would confine some patients who would not have been violent if had they been released.

Suppose that ER clinicians made decisions about hospitalization using a Future Violence Test (FVT) such as the one shown in Figure 4, which had the same (typical) accuracy as the prediction method used by the FSPH clinicians.[157] Suppose, for purposes of clarity in exposition, that persons evaluated with the FVT could score from 0 to 100, with higher scores implying higher likelihoods of violence in the near future. To use the FVT, ER clinicians would need to pick a decision threshold, or cut-off score about which a patient would be hospitalized.[158]

---

155. John A. Swets et al., *Psychological Science Can Improve Diagnostic Decisions,* 1 PSYCHOLOGICAL SCI. PUB. INT. 1, 22 (2000). However, we should note that the Comment quoted *supra*, text at note 152, is cited in its entirety in *Thompson v. County of Alameda*, 614 P.2d 728, 735 (Cal. 1980), which seemingly implies endorsement—without a definite numerical answer—for some kinds of cost-benefit balancing.

156. This section is adapted from Douglas Mossman & Kathleen J. Hart, *How Bad Is Civil Commitment? A Study of Attitudes Toward Violence and Involuntary Hospitalization*, 21 BULL. AM. ACAD. PSYCHIATRY L. 181, 182–90 (1993); and Mossman, *Rabbi's Sermon*, *supra* note 28, at 360–61.

157. That is, the effect size is 1. On the FVT scale depicted in Figure 4, the mean of the nonviolent persons is 45, and the mean of the violent persons is 55; both distributions have standard deviations of 10.

158. What I present here is a simplified discussion of how one would operationalize a detection method. For a more extensive discussion, see Mossman, *Dangerousness Decisions*, *supra* note 127, at 106–125.

Figure 4. — Score distributions for a "Future Violence Test" (FVT), with accuracy typical of actuarial methods of risk assessment. Dashed lines represent possible cut-offs or decision thresholds. Arrows are cut-offs that corresponds to the central 80 percent of responses depicted in Figure 5.

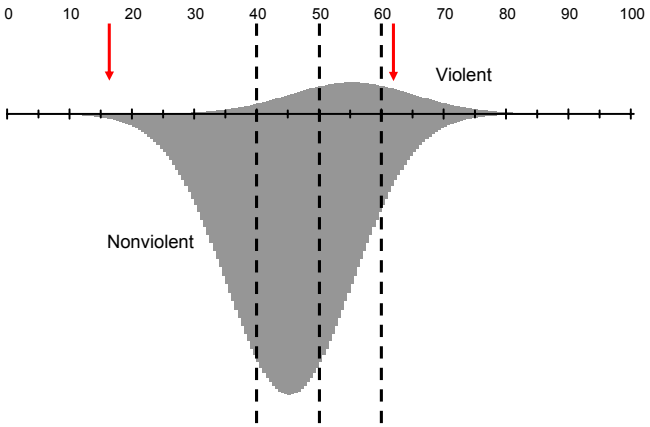


Figure 4 includes three vertical lines representing three possible FVT decision thresholds (at 40, 50, and 60). At each threshold, patients with scores greater than the cut-off are deemed "positive," that is, their risk of violence is deemed high enough to make them subject to involuntary hospitalization; patients with scores below the threshold are "negative" and are not subject to hospitalization. Moving the decision threshold changes the outcomes. For example, as the threshold decreases, more of the actually violent persons are hospitalized (because lower thresholds increase the sensitivity of the FVT), but more nonviolent patients are also hospitalized, because the probability of correctly identifying nonviolent patients person (the FVT's specificity) decreases.

Given that errors are inevitable, the rational way to choose a decision threshold would be to find the value of the FVT that strikes the right balance of correct and incorrect decisions about patients, given the values that we assign to incorrect and correct decisions. That is, we would like to find an FVT score that produces the highest *expected utility*; that point, by definition, would be the best threshold. Though not all the decisions made using this cut-off score would produce correct predictions about violence, using this score would produce the best balance of erroneous and correct predictions.[159]

159. "[S]ome kinds of errors may be much more important than other kinds, and the 'best' strategy [for predicting violence] should take into account the relative 'weights' or 'costs' of different

Our task, then, is to find a way to assign values to the possible outcomes of a decision about hospitalization, making use of some formal methods for quantifying utilities that have appeared in the decision analysis literature.[160] This will allow us to make mathematical calculations and discover the optimal decision threshold.[161] In the remarks quoted above, Monahan suggests that, when it comes to civil commitment, false negative mistakes (releasing dangerous patients) are much worse than false positive mistakes (hospitalizing nondangerous patients). But exactly how much worse? Do people agree, at least roughly, about how bad violence is compared to involuntary hospitalization? How could we find a way to answer these questions, that is, to get people to compare experiencing violence with experiencing involuntary hospitalization?

When most people (for example, legal decision-makers including legislators and judges) think about decisions to keep someone psychiatrically hospitalized, they think about the potential consequences for the public at large. From this perspective, a mental health professional's predictions about violence can lead either to (1) no one's being harmed, which occurs when the professional makes correct predictions that violence will or will not occur (that is, true positive or true negative predictions), or (2) a violent attack following a professional's incorrect (false negative) prediction of nonviolence. But notice that this perspective assumes one is not a patient: it does not take into account the experience of a nonviolent patient who, because of a prediction of violence, undergoes an involuntary hospitalization. The studies that I have conducted require individuals to consider the perspective of the patient as well as the perspective of the "public" whose safety is to be protected. Thus, the studies require subjects to think about releasing or involuntarily hospitalizing someone as possible outcomes that could happen to *them*.[162]

In a study I conducted with Professor Kathleen Hart, we told our study subjects (young adults attending Xavier University and the University of Cincinnati's medical school) to imagine they were helping

kinds of mistakes." MONAHAN, *supra* note 103, at 46–47.

160. *See, e.g.,* SIMON FRENCH, DECISION THEORY: AN INTRODUCTION TO THE MATHEMATICS OF RATIONALITY (1988); and the classic text, R. DUNCAN LUCE & HOWARD RAIFFA, GAMES AND DECISIONS: INTRODUCTION AND CRITICAL SURVEY (1957).

161. For a short discussion, *see* Douglas Mossman & Eugene Somoza, *Balancing Risks and Benefits: Another Approach to Optimizing Diagnostic Tests*, 4 J. NEUROPSYCHIATRY CLINICAL NEUROSCIENCES 331 (1992).

162. Although a full discussion of this assumption lies beyond the scope of this article, my intent was to have subjects consider outcomes of involuntary hospitalization from a frame of reference analogous to John Rawls's "initial position." *See* JOHN RAWLS, A THEORY OF JUSTICE 11–22 (1971).

to calibrate a Future Violence Test, specifically, to provide information that would allow mental health professionals balance incorrect predictions of violence and nonviolence. The key set of questions asked subjects to compare having to spend various lengths of time as a patient in state hospital to being attacked by a man wielding a knife. We asked them to consider how they would feel about experiencing each alternative "right now," and to tell us which alternative they would prefer.[163]

The answers we received spanned the entire range of time periods about which we asked, with (on one end of the spectrum) many subjects preferring several years of confinement and (on the other end) many subjects who would not want be willing to spend a day in a psychiatric ward to avoid a being attacked by a man wielding a knife.[164] In contrast to what Professor Monahan had suggested, many of our subjects gave answers implying that, when they considered the consequences of a false positive decision (a decision to hospitalize a nonviolent person for a few days) as happening to them, that decision seemed far *worse* than a false negative decision (a decision to release a violent person). The five-orders-of-magnitude[165] span of responses that we obtained suggested that, even in a fairly homogeneous group of persons, there was *no* social agreement on the right balance of correct and incorrect decisions about future violence.

2. Mental Health Professionals—Unpublished Research

Wondering whether mental health professionals might express more agreement[166] than did our set of young adult subjects, I conducted a similar study in October 2001.

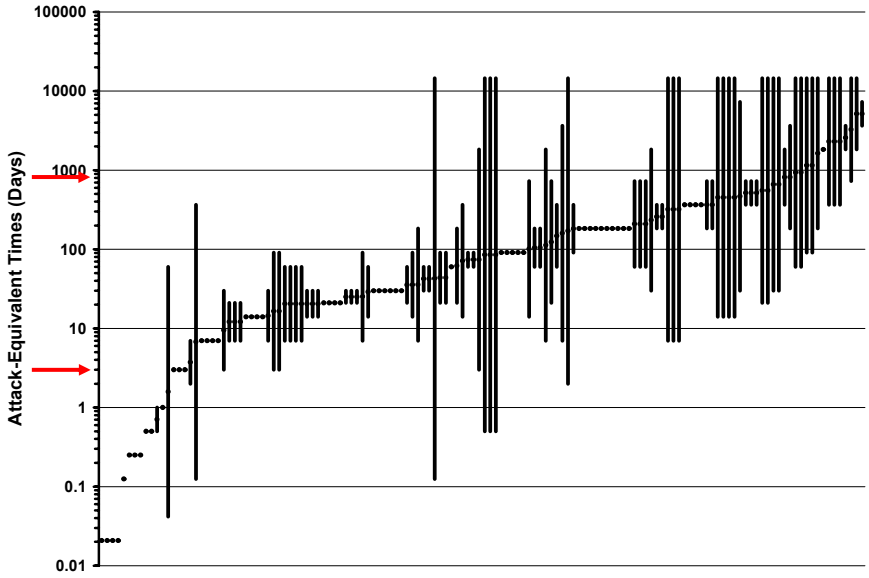
163. We asked subjects to consider periods of hospitalization ranging from three hours to ten years. The subjects could respond that they preferred being attacked or undergoing hospitalization for the period of time, or they could say they felt "about the same" regarding the alternatives. Details of the data collection procedures appear in Mossman & Hart, *supra* note 156, at 185–87.

164. The shortest period about which we inquired was three hours because, before conducting the study, we guessed that anyone would be willing to spend three hours in a psychiatric ward to avoid a knife attack. But, to our surprise, we found that several subjects preferred a knife attack to spending even three hours as an inpatient.

165. Ten years equals 87,600 hours. Assuming that the individuals who would not undergo three hours of hospitalization would agree to one hour, the ratio of highest to lowest acceptable time periods was nearly 10^5.

166. I thought there might be more agreement because mental health professionals would be less averse to hospitalization, and that few if any of them would prefer a knife attack to being hospitalized for a few days. I was wrong.

Figure 5. — Mental health professionals' responses to beings asked whether they would prefer being attacked by a man wielding a knife or spending a certain time period in the hospital. Vertical lines represent ranges of time periods about which a subject could not express a preference. Arrows designate time periods marking the middle 80 percent of the response distribution.



This time, the subjects, who were attending a presentation on violence prediction, included 141 mental health professionals[167] ranging in age from 22 to 77 years (mean and median age = 45 years).

This second study's key finding—which has not been previously published—is that the range of responses given by mental health professionals was similar to the range of answers from the first study's participants. As with the previous study, I asked subjects whether they would prefer being attacked by a man wielding a knife or spending a certain time period in the hospital. Subjects could respond "I don't know" if they were not sure whether they would prefer being attacked or the time period about which they were asked. For purposes of data interpretation, I calculated, for each subject, the geometric of the shortest time period for which the subject chose hospitalization and the

167. Among those who identified their sex, 31% were women. Those who identified their professions included twenty-eight counselors, twenty-eight psychologists and psychology trainees, twelve nurses, seventeen psychiatrists and psychiatry residents, and fifty-two social workers.

longest time for which the subject chose the attack.[168] In Figure 5, subjects' responses are ordered left to right according to this geometric mean; vertical lines represent ranges of time periods about which subjects could not express a preference. As Figure 5 shows, more than a tenth of the subjects thought that it might be better to spend two years or more in a hospital than be attacked by a man wielding a knife, and a similarly fraction preferred being attacked to spending seven days as a psychiatric inpatient.

Suppose that, to operationalize the FVT, we choose to ignore the feelings of the ten percent of subjects at the lowest end and the ten percent at the highest end of the distribution. That is, we take only the middle 80% to decide what range of cut-offs or decision threshold might be acceptable. The arrows in Figure 4 indicate the result: the range of thresholds lies between 17 and 62, that is, a range that includes decision thresholds implying hospitalization of virtually everyone (violent and nonviolent alike) to a threshold where 76% of the violent patients are released.[169]

D. Conclusion: No Agreement, and an Insoluble Dilemma

The results of these two studies show that if, as fairness requires, people are asked to consider the effects of involuntary hospitalization on those who are confined along with the benefits that confinement may give to society, people express very diverse views. This is true even in a situation where the threat (being attacked at a specific moment by an assailant whose choice of weapon is known) and its alternative (psychiatric hospitalization) are specified to a much greater degree than happens in real-life evaluations conducted in mental health professionals' offices or psychiatric emergency rooms. If it is the case, as Professor Monahan and colleagues suggest, that courts are reluctant to provide overt, numerical rules about balancing public safety and individual liberty, these studies' findings show that courts have a good reason to avoid doing so: even homogeneous groups of people cannot

168. For example, suppose a subject preferred being hospitalized for one month to being attacked, preferred being attacked to spending one year in the hospital, and said "I don't know" concerning intermediate time periods (*i.e.*, two, three, and six months). The geometric mean (in days) of one month and one year is $(30 \times 365)^{1/2} = 104.6$, which was the value used to order that subject's responses.

169. The derivation of these results appears in Appendix I of this Article.

One might conclude that this at least shows agreement that the individuals with the highest probability of violence should be hospitalized. But this is not true. The calculations do not take into account uncertainty in the estimated base rate of violence, which I have arbitrarily set a 10% for sample calculation. Factoring in this uncertainty would lead to an even wider span of acceptable cut-offs. *See* Mossman, *Dangerousness Decisions*, *supra* note 127, at 120–125.

agree on how much of their own liberty they would sacrifice to prevent violence.

Though violence prediction tools might be accurate, there would be no social agreement about using these tools, which would require establishing what probability of risk necessitated taking steps to protect third parties. In other words, even though mental health professionals can accurately rank individuals' risk of violence, society cannot agree upon what level of risk is "serious" enough to trigger a *Tarasoff*-type response to future danger.

Moreover, it appears that there is no agreement about risk levels even among judges. A study conducted by Professor Monahan and his colleague Eric Silver asked judges what minimum level of risk of violence would justify authorizing civil commitment. The judges gave answers that ranged from 1% to 56%.[170] This provides yet another demonstration that, even within homogeneous groups, there is a wide divergence of beliefs about what probability of violence justifies a specific action to avert danger.

Thus, thirty years after *Tarasoff*, scientific findings suggest that the decision contains an inescapable contradiction. In a situation such as the one presented to Poddar's therapists, a clinician has an obligation to apply standards of his profession to determine whether his patient represents a "serious danger of violence." The therapist then may—depending on whether he believes the patient represents a "serious danger"—have an obligation to respond protectively through some course of action. Yet the therapist also knows that there is and can be no rationally established, broadly accepted criterion for what probability of risk constitutes the level of "serious danger" that should trigger a protective response. As established by the California Supreme Court, the *Tarasoff* rule requires a therapist to recognize an apparently quantifiable entity—"a serious danger of violence"—when the requisite quantity cannot be specified. The impossibility of rationally implementing the *Tarasoff* obligation suggests something basically wrong with the major premise the California Supreme Court formulated so that Poddar's therapists might be found negligent.

170. John Monahan & Eric Silver, *Judicial Decision Thresholds for Violence Risk Management*, 2 INT'L J. FORENSIC MENTAL HEALTH 1, 4 (2003). In Figure 4, 1% and 56% probabilities of violence correspond to scores of 26 and 74 respectively. The computation of these results appears in Appendix II of this Article.

VI. A KANTIAN PERSPECTIVE

A. Problems with Predicting Consequences

Whether or not one thinks that Prosenjit Poddar's psychologist did the right thing when he called the police, the statement that Poddar made concerning his intent to kill Tatiana Tarasoff does strike me—and I think would strike most mental health professionals—as an event that should have led a clinician to *at least consider* taking some action beyond simply continuing outpatient therapy sessions. Yet, as *Tarasoff* and subsequent cases have framed the matter, a psychotherapist's undertaking a potentially protective response to a patient's possible future violence becomes a matter of trade-offs, in which, for obscure "policy" reasons, the value of protecting the public is paramount. What the patient loses—be it confidentiality (if the therapist issues a warning) or freedom (if the therapist initiates hospitalization)—is justified by balancing the consequences for the patient of taking action against the consequences for society of failing to do so. What is good for the individual must be sacrificed for the (greater) good of society.

I have stated this result baldly, perhaps tendentiously, but nonetheless (I think) fairly to highlight what should be a source of ethical discomfort about the characterization of ethical decision-making that we find in *Tarasoff*. In traditional medical ethics, doctors serve individual patients and have fiduciary obligations to them, not to those around them. Yet the medical literature affords repeated examples that consider physicians' clinical decisions in light of the broader social implications of those decisions, such that doctors are urged to make choices different from ones that they would be dictated by considering only the welfare of an individual patient.[171] In such cases, it seems, ethical principles are in conflict with each other, and the problem facing doctors therefore is to achieve some sort of resolution among the conflicting principles.

171. For example, "Although antibiotics have little or no benefit for colds, upper respiratory tract infections, or bronchitis, these conditions account for a sizable proportion of total antibiotic prescriptions for adults by office-based physicians in the United States." R. Gonzales et al., *Antibiotic Prescribing for Adults with Colds, Upper Respiratory Tract Infections, and Bronchitis by Ambulatory Care Physicians,* 278 JAMA 901, 901 (1997). The needless prescription of an antibiotic may not harm the individual who receives the drug (and may placate a patient's desire to get something tangible from an office visit), but the widespread practice of unnecessarily distributing antibiotics conduces to the evolution of drug-resistant organisms. Richard Colgan & John H. Powers, *Appropriate Antimicrobial Prescribing: Approaches That Limit Antibiotic Resistance,* 64 AM. FAM. PHYSICIAN 999 (2001). For this reason, physicians have called upon their colleagues to desist from this practice. B. Schwartz et al., *Preventing the Emergence of Antimicrobial Resistance. A Call for Action by Clinicians, Public Health Officials, and Patients*, 278 JAMA 944 (1997).

The presence of such apparent conflicts is exemplified by the Preamble of American Medical Association's *Principles of Medical Ethics*, which tells doctors that their "profession has long subscribed to a body of ethical statements developed primarily for the benefit of the patient. As a member of this profession, a physician must recognize responsibility to patients first and foremost, as well as to society . . . ."[172] Elsewhere in the *Principles,* the physician is told that "while caring for a patient," the physician must "regard responsibility to the patient as paramount."[173] Yet the same *Principles* also advise the physician to "respect the law and. . .seek changes in those requirements which are contrary to the best interests of the patient"[174] and to "safeguard patient confidences and privacy," though only "within the constraints of the law."[175]

Such statements reflect the problems inherent in attempting to create guidelines for ethical choice absent an overall theory that provides the background for resolving or reconceptualizing conflicts between rules that customarily inform conduct. In a series of writings,[176] I have suggested that Kant's theoretical approach has much to offer psychiatrists (and implicitly, other mental health professionals) in contexts where their duties to patients seem to conflict with the expectations or demands of society. Kant's theories have especial appeal in the current context because they express eschew consequentialism, that is, the view that normative pronouncements

172. AMERICAN MEDICAL ASSOCIATION'S PRINCIPLES OF MEDICAL ETHICS [hereinafter AMA ETHICS], http://www.ama-assn.org/ama/pub/category/ 2512.html (last visited Feb. 8, 2006).

173. *Id*. at Principle VIII.

174. *Id*. Principle III.

175. *Id*. at Principle IV. An earlier version of this Principle is cited in *Tarasoff II*, 551 P.2d 334, 347 (Cal. 1976). In December 1983, following *Tarasoff* and subsequent related decisions, AMA's Council on Ethical and Judicial Affairs developed an ethics policy to address psychiatrists' new, legally imposed obligations:

> The obligation to safeguard patient confidences is subject to certain exceptions which are ethically and legally justified because of overriding social considerations. Where a patient threatens to inflict serious bodily harm to another person or to him or herself and there is a reasonable probability that the patient may carry out the threat, the physician should take reasonable precautions for the protection of the intended victim, including notification of law enforcement authorities.

AMA Code of Medical Ethics Opinion E-5.05, http://www.ama-assn.org/ama/ pub/category/8353.html (last visited Feb. 8, 2006).

176. Douglas Mossman, *The Psychiatrist and Execution Competency: Fording Murky Ethical Waters*, 43 CASE W. RES. L. REV. 1, 53–88 (1992); Douglas Mossman, *Is Forensic Testimony Fundamentally Immoral?*, 17 INT'L J. L. & PSYCHIATRY 347, 357–68 (1994); Douglas Mossman, *Is Prosecution "Medically Appropriate"?* 31 NEW ENG. J. ON CRIM. & CIV. CONFINEMENT 15, 60–78 (2005) [hereinafter Mossman, *Medically Appropriate*].

should reflect calculations about future consequences.[177] Having just shown, in previous sections, the futility of making decisions based on calculations about the consequences of violence predictions, I now address how a Kantian approach to the facts in *Tarasoff* can lead to a superior lesson from the case and a more practicable legal rule.

B. Kant's Approach to Ethics

1. Two Formulations of the "Categorical Imperative"

Kant's ethical theory is premised on (among other things) the idea that morality addresses itself to rational beings. This makes moral obligation something that sensible people cannot avoid considering, because we cannot rationally claim that we are not rational beings. If, then, we discover some fundamental law or rule that prescribes what a rational agent must do under certain circumstances, we have come upon a duty that we cannot logically avoid. Acting out of moral duty constrains us to follow this fundamental law, which describes how any rational creature should act. We cannot ignore or avoid this duty; because of its universal content, it retains its validity under all possible circumstances, and is therefore "categorical."[178]

This leads to Kant's first formulation of his "categorical imperative," which tells each person, "Act only on that maxim through which you can at the same time will that it should become a universal law."[179] On this formulation, "the moral worth of an action does not depend on the result expected from it," says Kant; "nothing but the *idea of the law* itself . . . can constitute that pre-eminent good which we call moral, a good which is already present in the person acting on this idea and has not to be awaited merely from the result."[180]

For example, suppose I wish to find out whether it is permissible to promise deceitfully. The categorical imperative directs me to consider not whether it might be prudent to do so (which would involve an inquiry about the expected short- and long-term gains and losses from breaking a promise), but whether such conduct could be right. To find this out, I need to ask whether it could be a universal law that *everyone*

177. For a nice explanation of consequentialist moral theories (and their limitations), *see* Walter Sinnott-Armstrong, *Consequentialism*, *in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta, ed., Summer 2003), *available at* http://plato.stanford.edu/entries/consequentialism/; for a defense, *see* Walter Sinnott-Armstrong, *An Argument for Consequentialism*, 6 ETHICS 399 (1992).

178. GMS, *supra* note 27, at 87–88 [420–21].

179. *Id*. at 88 [421].

180. *Id*. at 69 [401].

might make a false promise when one could not otherwise extricate oneself from some difficult situation. Simply stating this allows me to see that lying could not be a universal law of conduct: if lying were universal, "there could properly [speaking] be no promises at all," because no one would believe anyone else's promise. "[C]onsequently my maxim, as soon as it was made a universal law, would be bound to annul itself."[181]

An important point, for our present purposes, is that I do not have to be able to predict the future (or, to use Kant's phrase, "I need no far-reading ingenuity"[182]) to discover what I must do to behave morally. Though I cannot know and am "incapable of being prepared for all the chances that happen in [the world], I ask myself only: 'Can you also will that your maxim should become a universal law?' Where you cannot, it is to be rejected, . . . not because of a prospective loss to you or even to others, but because it cannot fit as a principle into a possible enactment of universal law."[183]

The preceding paragraphs' concern motives for action that can address themselves only to rational beings, because only rational beings can attempt to guide themselves in conformity to laws. If I regard myself as a rational being guided by laws, I must also regard other rational beings as equal, in this respect, to myself. When I attempt to guide myself by categorical imperatives, my will directs itself to an end that has noncontingent, absolute worth—an end, that is, in itself. But now I have discovered "something *whose existence* has *in itself* an absolute value, something which *as an end in itself* could be a ground of determinate laws."[184] If I recognize my own rational nature as an end in itself, I must also acknowledge that "every other rational being conceives his existence"[185] similarly, which yields "an *objective* principle, from which, as a supreme practical ground, it must be possible to derive all laws for the will."[186]

Realizing this allows Kant to conclude that, as a practical matter, the first formulation of the categorical imperative is equivalent to the following rule: "*Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end.*"[187] Returning to the

181. *Id*. at 71 [403].
182. *Id*.
183. *Id*.
184. *Id*. at 95 [428].
185. *Id*. at 96 [429].
186. *Id*. at 96 [428–29].
187. *Id*. at 96 [429].

example of the deceitful promise, we gain from this second formulation an additional understanding of why such an act is wrong. Kant says that someone who contemplates deceiving someone else should "see at once that he is intending to make use of another man *merely as a means*"[188] without regarding the other as an end in himself, "[f]or the man whom I seek to use for my own purposes by such a promise cannot possibly agree with my way of behaving to[ward] him, and so cannot himself share the end of the action."[189]

2. An Aside

Although this derivation of moral principles is abstruse, a key force behind the attractiveness of these two principles is their agreement with some of our everyday intuitions about moral thinking. Universalization, for example, underlies the homely question, "How would you feel if someone did that to you?" that we use to get people to think about their behavior less selfishly.

Kant pointedly states, however, that the categorical imperative is not the same as the command, "That which you would not want done to you, do not do unto others." The latter is only a consequence of the former. Also, the latter does not provide a basis for duties to oneself, for duties to treat others benevolently (because many people might consent to be left alone in order to be excused from being good to others), or for the justness of criminal punishments (against which the criminal might argue to the judge who would sentence him).[190] My point here is only that Kant's ideas accord with how we really make moral arguments to each other. Kant makes a similar point in the *Critique of Practical Reason*:

> Ask yourself whether, if the action which you propose should take place by a law of a nature of which you yourself were a part, you could regard it as possible through your will. Everyone does, in fact, decide by this rule whether actions are morally good or bad. . . . If the maxim of action is not so constituted as to stand the test of being made the form of a natural law in general, it is morally impossible. Even common sense judges in this way, for its most ordinary judgments, even those of experience, are always based on natural law.[191]

188. *Id*. at 97 [429].

189. *Id*.

190. *Id.*

191. KpR, *supra* note 27, at 72–73 [69].

Treating all rational beings as ends in themselves, and not as means to some other end, is what lies behind our notion that morality requires us to restrict criminal punishment only to those who are guilty. That this must be an absolute bound on conduct accounts for our intuitive grasp of Abraham's argument against destroying Sodom and Gomorrah in their entirety if those cities contained righteous men. Abraham's challenge to G–d—"It would be sacrilege to You to do such a thing as this, to kill the righteous with the wicked, and have the righteous and the wicked fare alike. It would be sacrilege to You! Shall the Judge of the whole world not do justice?"[192]—rests upon "the audacious claim"[193] that even the Almighty's acts must be judged against a moral standard that does not tolerate the death of some humans as a means for attaining some larger goal.

3. Conclusions from the Categorical Imperative

The universalization and end-in-itself formulations of the categorical imperative have direct behavioral implications, says Kant, for our development as human beings and for out treatment of each other. First, these principles tell us that we have a duty to develop our natural talents. It is true that a world could exist in which we were lazy and content with how we are. But we could not will this course of conduct as a universal law of nature: as rational beings, we must realize and will that our talents be developed, because our talents are useful to us. Moreover, to neglect our talents, though possibly consistent with the "*maintenance*" of humanity as an end in itself, is not consistent with the *"promotion"* of our humanity.[194]

Second, these formulations tell us that we have a duty to care about the welfare of others and to promote their improvement. True, we can conceive of a world in which no one cared about or helped others, but we cannot will that ignoring the needs of others should be a universal practice. Such a maxim must contradict itself, for there are many times in which we would need and want aid, love, or sympathy from others, but our universalization of not caring about others would deprive us of any reason to hope of obtaining those things from anyone. Also, if our regard for others as ends in themselves is to receive its full meaning and have its full effect on us, then we must see the promotion of others'

192. *Genesis* 18:25 (my translation).

193. RABBINICAL ASSEMBLY, ETZ HAYIM: TORAH AND COMMENTARY 103 (2001).

194. GMS, *supra* note 27, at 97–98 [430].

faculties as the realization of their being treated as ends.[195]

4. Two Other Formulations[196]

Kant offers a third (in his mind, inter-derivable) formulation of the categorical imperative, one that Kant summarizes as the principle of autonomy.[197] When one wills from duty, one renounces self-interest and regards oneself as the source of a universal legislative will, rejecting maxims that are inconsistent with this posture. Applied to everyone, we obtain the"Idea *of the will of every rational being as a will which makes universal law*,"[198] which is a candidate for a categorical imperative because it is not based on self-interest and can therefore be universal. This formulation focuses on our status as law-giving actors and is therefore a source of dignity. We renounce selfish motives and are guided by those principles that express our independence from self-interest. We strive, that is, to adhere to those principles that express the autonomy of a rational will, one that can be the source of universal laws and can obligate itself through them.[199] Our autonomy, dignity, and intrinsic value as human beings inheres in our being rational agents who can create and use binding moral laws as guides for conduct.[200]

This leads to Kant's fourth formulation of the categorical imperative, which envisions rational beings united in their commitment to "[a]ct on the maxims of a member who makes universal laws for a merely possible kingdom of ends."[201] That is, we conform our actions to maxims laid down by a legislator of universal laws that could bind all rational wills, including his own; at the same time, we see everyone as a legislator who must be treated as an end in himself. In other words, we recognize that we have a fundamental obligation to act on principles that would be acceptable to a community of rational agents, each of whom had an equal share in creating the laws that governed the community.[202]

195. *Id*. at 98 [430].

196. Kant believed his four formulations of the categorical imperative were equivalent, at least practically—they all led to the same results as guides for conduct. For a good discussion of this issue, and a useful summary of Kant's moral philosophy, *see* Robert Johnson, *Kant's Moral Philosophy*, *in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta, ed., Spring 2004), *available at* http://plato.stanford.edu/entries/kant-moral/.

197. GMS, *supra* note 27, at 100 [433].

198. *Id*. at 98 [431].

199. *Id*. at 99–100 [431–33].

200. *Id*. at 102–03 [435–36].

201. *Id*. at 106 [438–39].

202. *Id*. at 101 [433–34].

C. Implications of Kant's Ethical Philosophy

If we adopt Kant's principles as valid ethical constraints for Poddar's clinicians, then any action required of clinicians by a tort rule must recognize that those clinicians operate under those ethical constraints when they choose how to respond to Poddar. Those constraints include the notion, first, that the response required of clinicians must be "universalizable"; that is, the response must apply to all situations such as the one that arose in August 1969, when Poddar made known his intent to kill Tatiana. The required response of clinicians should not depend on any ability to predict violence, but should be based on the information they have at hand. The clinicians must respond to what Poddar has done, not to what he might do.

Second, whatever practice the law would require of clinicians must allow them to treat Poddar as an end in himself. The required practice should not use him merely as a means to achieve some social goal (however attractive that goal may seem). Third, any required practice must respect Poddar's rationality. This is not meant as a comment on whether each (or any) of Poddar's plans or wishes actually are rational. Rather, the clinicians' response to Poddar must respect his humanity and his logical aspiration, as a reasoning being and member of the kingdom of ends, to act rationally in accordance with motives that his fellow rational creatures could adopt.

VII. A REVISION OF THE *TARASOFF* RULE

A. Review

In Section II.B., I posited that after looking at the facts leading to Tatiana Tarasoff's death, the majority of judges hearing the *Tarasoff* case drew the conclusion that Poddar's actions merited some response by his clinicians beyond simply continuing his psychotherapy. I then characterized the court's job as construing a major premise such that, taking the facts of the case as a minor premise, liability on the part of Poddar's clinicians (and other clinicians in similar circumstances) was a logical conclusion. The general form of this major premise was:

- If a patient does A and the therapist does not do B and the patient later harms C, the therapist will be liable for the harm to C.

Filled in by the *Tarasoff* court, that major premise became:

- If a patient displays behavior that a therapist should recognize as indicating that the patient presents a serious danger of violence, and

> if the therapist does not both (1) apply the standards of the therapist's profession to determine that the patient presents a serious danger of violence and (2) use reasonable care to protect the intended victim from the danger, and if the patient later harms C, then the therapist will be liable for the harm to C.

As previous sections have explained, this major premise permitted post-*Tarasoff* courts to expand the duty to protect well beyond those clinical junctures in which patients made threats about specific people whom they intended to harm, to encompass any clinical encounter in which a patient might pose a "serious risk" of harming anyone who might conceivably cross his path. As we have also seen, application of this major premise, in the context of the imperfect ability of clinicians to distinguish persons who will not do violence from those who will, necessarily requires clinicians to confine (or otherwise restrict the freedom of) patients who may not have threatened anyone and who have not yet done anything harmful, simply because they pose a risk of doing so. The level of probability that should trigger this protective action cannot be defined, however. Moreover, by countenancing needless confinement as a statistical consequence of clinicians' imperfect predictions, this legal scheme explicitly treats nonviolent patients' loss of freedom as a means toward the end of protecting other members of society.

Is it possible to construct a major premise that, given a situation like the one that Poddar's statements created, would hold a therapist to some responsibility to act protectively without treating psychiatric patients as mere objects whose freedom can be disregarded if doing so will further society's interests? In fact, the job of creating such a premise has been accomplished in the form of legislation that more than twenty state legislatures have enacted.[203] Often, this legislation has responded to concerns of mental health professionals who, faced with the expanding and uncertain prospects of court-imposed liability for patients' violent acts, sought help from state legislatures to provide "well-defined limitations of their protective duties."[204]

---

203. Claudia Kachigian & Alan R. Felthous, *Court Responses to* Tarasoff *Statutes*, 32 J. AM. ACAD. PSYCHIATRY L. 263, 264 (2004) (at the publication date, "23 such statutes pertaining to psychiatrists have been enacted.").

204. *Id*. at 272. "The creation of such a duty to protect has aroused much controversy because of . . . ambiguity about what actions the clinician must undertake to discharge the duty. To clarify clinicians' responsibilities, many states have enacted laws that limit therapists' potential liability if they take specified actions when a patient makes a serious threat against an identifiable victim." Dale E. McNeil et al., *Management of Threats of Violence Under California's Duty-to-Protect Statute*, 155 AM. J. PSYCHIATRY 1097, 1097 (1998) (citations omitted). *See, e.g.*, Jenks v. Brown, 557 N.W.2d 114, 116

B. Ohio's Major Premise

An example of such a major premise is the 1999 statute enacted in Ohio in response to the state's first clear duty-to-protect decision, *Estates of Morgan v. Fairfield Family Counseling Center*.[205] The decision arose from a lawsuit brought against a mental health center and its professionals following a July 1991 episode in which Matt Morgan shot his parents to death and wounded his sister. Morgan was charged with murder, but a jury found him not guilty by reason of insanity.[206] Morgan had experienced mental problems for a few years and had taken antipsychotic drugs in Philadelphia before returning to Ohio with instructions to continue care. A psychiatrist treating Morgan at the Fairfield Family Counseling Center (Fairfield) stopped Morgan's medication. When Morgan's condition deteriorated months later, other clinicians decided they could not force medication or hospitalization upon him.

A 4–3 majority of the Ohio Supreme Court stated that the plaintiffs had valid grounds to sue. Relying (as it in had several other duty-to-protect decisions) on Restatement (Second) of Torts §§ 315 and 319, the court found that a "special relationship" exists between a psychotherapist and a patient, such that the therapist has "a duty to exercise his or her best professional judgment to prevent such harm from occurring."[207]

The *Morgan* majority expressly assigned mental health care the social role of controlling violence through treatment, especially medication. The court reasoned that antipsychotic "medication controls symptoms of schizophrenia in approximately seventy percent of schizophrenics," that while taking medication Morgan had been "a medication-controlled . . . patient," and that if Morgan had continued to take "medication, he would not have had the overt psychotic symptoms that led him to kill his parents and injure his sister."[208] Also, the clinicians at Fairfield had the power to initiate civil commitment procedures. "Thus," said the *Morgan* majority, "we conclude that the psychotherapist-outpatient relationship embodies sufficient elements of

(1996) (commenting on the legislative history of Michigan's statute: "[e]nacted in 1989, the duty to warn statute was created to limit the liability of mental health practitioners.").

205. 673 N.E.2d 1311 (Ohio 1997).

206. For a fuller description of Matt Morgan's background and current living situation, *see* Encarnacion Pyle, *In Shadow of Tragedy, Normal Life Takes Root: Man Who Killed Parents Optimistic Despite Illness,* COLUMBUS DISPATCH, Nov. 27, 2005, at A1.

207. *Morgan*, 673 N.E.2d at 1328–1329.

208. *Id.* at 1323–1324.

control to warrant a corresponding duty to control."[209] Moreover, said the *Morgan* court:

> Society has a strong interest in protecting itself from those mentally ill patients who pose a substantial risk of harm. (citation omitted). To this end, society looks to the mental health profession to play a significant role in identifying and containing such risks. (citation omitted). The mental health community, therefore, has a broadly based responsibility to protect the community against danger associated with mental illness.[210]

Ohio mental health professionals responded to *Morgan* just as had clinicians in other states where courts had imposed duties to protect the public—by appealing to the state's legislature for a more reasonable rule about when liability could be imposed. The result was a 1999 statute passed expressly to supercede *Morgan*,[211] with provisions similar to a model statute proposed in the late 1980s by the American Psychiatric Association (APA).[212] As a result, an Ohio mental health professional may be held liable for harm done by his patient to a third party, but only if the "patient or a knowledgeable person has communicated to the professional . . . an explicit threat of inflicting imminent and serious physical harm to or causing the death of one or more clearly identifiable potential victims, [and] the professional . . . has reason to believe that the . . . patient has the intent and ability to carry out the threat."[213] If such a threat is made, taking one of several actions—arranging for the emergency, voluntary, or involuntary hospitalization of the patient; establish and undertake a form of treatment that is calculated to eliminate the risk of violence; or warning the police and the potential target of violence—immunizes the clinician from liability.[214] Ohio's law directs that the clinician, in selecting a course of action, pick an alternative that, while eliminating the danger, "would least abridge the

209. *Id*. at 1324.

210. *Id*. at 1324 (citations omitted). I am not sure whether this assertion is more insulting to mental health professionals or to the patients whom we treat. As I have observed elsewhere, "I had entered psychiatry to help patients become more autonomous and to fulfill their human potential. In *Morgan*, however, Ohio's supreme court said my job was to contain risks posed by dangerous people whose willful acts would otherwise spread like deadly germs." Here was (to use Phil Resnick's phrase) "the zoo-keeper theory of psychiatry" officially endorsed in a legal opinion. Mossman, *Rabbi's Sermon*, *supra* note28, at 362.

211. OHIO REV. CODE § 2305.51 (West 1999). Ohio House Bill 71 also reworded portions of OHIO REV. CODE § 5122.34. Michigan is one of several states with an equivalent law, although it is organized differently and worded much more clearly. *See* MICH. COMP. LAWS § 330.1946 (West 2006).

212. *See* Appelbaum et al., *supra* note 14, at 827–28.

213. OHIO REV. CODE § 2305.51(B) (West 2006).

214. *Id*.

rights" of the patient.[215]

In the analytical framework used in this Article, the "major premise" created by the Ohio statute is

- If a patient or a knowledgeable person communicates to a mental health professional an explicit threat to do serious physical harm to one or more clearly identifiable potential victims, and if the professional has reason to believe that the patient intends and has the ability to carry out the threat, and if the therapist does not take a specified action—arrange for hospitalization of the patient, effectuate a treatment plan calculated to eliminate the risk of violence, or warn both the police and the potential victim—,and if the patient later harms C, then the therapist may be liable for the harm to C.

Let us now return to the problem faced by the majority in *Tarasoff*, who, viewing the facts presented by the plaintiffs, felt that the appropriate response was to construct a rule under which the defendants would have answer in tort. Does the just-articulated major premise, coupled with the facts in Poddar's treatment, lead to the conclusion that Poddar's clinicians could be liable for Tatiana's death? It certainly does. Poddar uttered a specific threat with a specific target, and his psychologist apparently believed Poddar could and intended to act upon his threat. The psychologist, in consultation with other clinicians, attempted to arrange hospitalization but did not do so effectively; the clinicians did not alter Poddar's treatment to address his homicidal thoughts (and may actually have compromised his treatment); the clinicians alerted the police but not Tatiana.

C. Is Ohio's Duty to Protect Acceptable?

1. Fairness to Mental Health Professionals

Ohio's statutory response to *Morgan* would thus satisfy the *Tarasoff* majority's needs for a liability-creating major premise as well as did the rule that the majority actually articulated. But is the Ohio statutory rule any better than the *Tarasoff* rule? For selfish reasons, almost all mental health professionals would probably prefer Ohio's rule because it clearly demarcates those events that create a duty to act protectively and what a mental health professional must do to fulfill the duty. Mental health professionals also might argue that the Ohio rule is objectively fairer to

215. OHIO REV. CODE § 2305.51(C)(2) (West 2006).

them. After all, among the reasons that traditional Anglo-American common law "has persistently refused to impose on a stranger the moral obligation of common humanity to go to the aid of another human being who is in danger," are "the difficulties of setting any standards of unselfish service to fellow men. . . ."[216] Yet duties to protect or rescue others exist in Jewish religious law,[217] French criminal law,[218] and laws in some U.S. states,[219] which place limited legal requirements on individuals to provide assistance to others facing danger, requirements that accord with our moral sense of what "common humanity" requires. Mental health professionals might with some justification argue that assigning them a protective obligation to intervene is fair if they are given clear boundaries on when the legal duty would apply and clear instructions on what actions discharge the duty.[220]

---

216. W. PAGE KEETON ET AL., PROSSER AND KEETON ON THE LAW OF TORTS, § 56 at 375–76 (5th ed. 1984). The latter quote appears in *Tarasoff II*, 551 P.2d 334, 343 n. 5. (Cal. 1976).

217. *See Leviticus* 19:16, which contains the Hebrew commandment "*lo ta'amod al-dam rei'echa.*" These words literally say, "[d]o not stand on the blood of your neighbor," but their meaning is better captured in the Jewish Publication Society's 1917 translation, "neither shalt thou stand idly by the blood of thy neighbor." *Leviticus* 19:16 (Jewish Pub. Soc. 1917). I provide a more detailed discussion of this commandment in the context of the *Tarasoff* obligation in *Rabbi's Sermon, supra* note 28, at 362.

218. *See* C. PÉN, ART. 223-6:

> *Quiconque pouvant empêcher par son action immédiate, sans risque pour lui ou pour les tiers, soit un crime, soit un délit contre l'intégrité corporelle de la personne s'abstient volontairement de le faire est puni de cinq ans d'emprisonnement et de 500 000 F d'amende.* (Anyone who, being able to prevent by his immediate action, without risk to himself or to third parties, a felony or a misdemeanor against the bodily integrity of a person, wilfully abstains from doing so, is punished by five years' imprisonment and a fine of 500,000 francs [$75,000].)
>
> *Sera puni des mêmes peines quiconque s'abstient volontairement de porter à un personne en péril l'assistance que, sans risque pour lui ou pour les tiers, il pouvait lui prêter soit par son action personnelle, soit en provoquant un secours.* ("The same penalties apply to anyone who wilfully abstains from rendering help to a person in danger that he could render, without risk to himself or to third parties, either by his personal action or by initiating rescue operations.").

Id., *available at* http://admi.net/code/index-CPENALLL.html (my translation). The history of this obligation in French criminal law is reviewed in Peter M. Agulnick & Heidi V. Rivkin, *Criminal Liability for Failure to Rescue: A Brief Survey of French and American Law,* 8 TOURO INT'L L. REV. 93, 106–110 (1998).

219. VT. ST. ANN., tit. 12 § 519 (a) (2005); R.I. GEN. LAWS § 11-56-1 (2005); MASS. GEN. LAWS ANN. ch. 268, § 40 (2006); WASH. REV. CODE ANN. § 9.69.100 (2006); WIS. STAT. ANN. § 940.34 (2005); FLA. STAT. ANN. § 794.027 (2006).

220. Over two decades ago, Dr. Appelbaum observed, with his customary sagacity:

> Even before the original *Tarasoff* decision, therapists often felt a responsibility to protect potential victims of their patients and acted in that regard. It is, indeed, difficult to formulate a moral argument *against* the position that therapists should act to protect those whom they believe to be endangered, as should all human beings. Recent data suggest that the majority of therapists would support this position. The overlay of legal liability

A key feature of Ohio's duty-to-protect statute is that the circumstances that trigger the duty do not involve predictions of violence or calculations of whether the probability of violence has reached some threshold level. Instead, mental health professionals need only to take the threats of patients seriously, and to decide, using their common sense, whether their patients can and really intend to carry out those threats. Under such circumstances, the statute in effect tells the professional, "The law expects you to try to intervene somehow, just like anyone should." The statute expects mental health professionals to consider actions—arranging for hospitalization or changing treatment plans—that ordinary citizens cannot do. Yet these special capacities and the requirement to consider them have parallels in other legally bestowed powers or responsibilities exercised by other citizens, including school officials, school crossing guards, lifeguards, fire-fighters, and police.The Ohio statute recognizes the special status of mental health professionals but does not single them out unfairly.

2. The Kantian Perspective

As previous sections have shown, the protective duty enunciated in *Tarasoff* leads to two objectionable results: impracticability and less-than-full regard for the humanity of psychiatric patients. Ohio's duty-to-protect statute solves the impracticability problem by referring to known events rather than predictions, probabilities, and undefinable thresholds. But does Ohio's statute address the moral problems described in Section VI?Judged from a Kantian perspective, how does Ohio's statute measure up?

a. The Patient as Human Actor

On initial inspection, the Ohio statute has the clear moral advantage of treating the patient as a human being, capable of rationally carrying out goals for potentially rational ends. In looking to a patient's threats as the trigger for action, the Ohio duty-to-protect statute treats the patient as a planners and initiators of actions undertaken for reasons.

> has served to distort this moral core of the *Tarasoff* doctrine. Further, the requirement that therapists protect victims not only when they know of potential dangerousness but when, according to professional standards, they *should know* of it is probably too stringent, given the limits of current abilities to predict dangerousness and the absence of professional standards for this task.

Paul S. Appelbaum, Tarasoff *and the Clinician: Problems in Fulfilling the Duty to Protect*, 142 AM. J. PSYCHIATRY 425, 429 (1985) (citations omitted).

The statute directs mental health professionals to attend to a patient's expressed intent as an indicator of what the patient will do, rather than to regard the patient as a statistical source of future harm, the probability of which the professional must estimate. Though it is not the job of mental health professionals (*qua* mental health professionals) to provide moral evaluations of their patients, the Ohio statute lets professionals regard their patients as actors who think, make choices, and may perform stupid or blameworthy actions—the only attitude that is consistent with regarding patients as human, moral agents.

b. The Problem of Coercion

A problem arises, however, when we put portions of Ohio's statute to the tests of morality specified in Kant's ethics. Directing professionals to make changes in a treatment plan seems fully consistent with what Kant would require. Although the word "psychiatrist" did not exist when Kant was alive, Kant knew that people could be irrational and make bad judgments, and suggests that a form of self-administered cognitive therapy is essential:

> Reason must in all its undertakings subject itself to criticism; should it limit freedom of criticism by any prohibitions, it must harm itself, drawing upon itself a damaging suspicion. Nothing is so important through its usefulness, nothing so sacred, that it may be exempted from this searching examination, which knows no respect for persons.[221]

At the same time, Kant construed the obligation to respect others' humanity as placing limits on how we may respond to others' ill-advised plans. We may not treat another person as though he were not capable of using reason to control of himself. "Even in a case where someone evidently *is* wrong or mistaken," writes Christine Korsgaard, Kant believed that "we ought to suppose he must have what he takes to be good reasons for what he believes or what he does. . . . [T]his attitude is something that we *owe* to him, something that is his right."[222] In responding to someone's "errors," states Kant, we should not call them "absurdities, poor judgment, and so forth, but rather [should] suppose that his judgment must yet contain some truth and to seek this out" while at the same time "explaining to him the possibility of his having erred, to preserve his respect for his own understanding."[223] While Kant may

221. KrV, *supra* note 27, at 596 [A739/B767].

222. Chiristine M. Korsgaard, *The Right to Lie: Kant on Dealing with Evil*, 15 PHIL. PUB. AFF. 325, 335 (1986).

223. MS, *supra* note 27, at 210 [463].

encourage self-examination and self-improvement, what we can legitimately require of others is limited by the obligation to address them as rational beings and ends in themselves. "Reason depends on this freedom for its very existence," states Kant. "For reason has no dictatorial authority; its verdict is always simply the agreement of free citizens, of whom each one must be permitted to express, without let or hindrance, his objections or even his veto."[224]

We therefore may reasonably take Kant to say that, if I know someone is about to do something wrong, I may try to convince him to do otherwise, perhaps through argument or (as his psychiatrist) psychotherapy. But Kant also says I may not do things to subvert his rational choice, even if I believe doings these things will thwart his plans to carry out his evil intent.

In the most (in)famous example of this tenet, Kant insists that it would be wrong to lie to a would-be murderer who comes to my door and asks whether the intended victim, my friend whom I am sheltering within, is in the house. Some of Kant's reasons for this seem contrived, wrong, and consequentialist. He states, for example, that if I answered "no," but was unaware that my friend was leaving my home, so that my lie resulted in the killer's finding my friend when he otherwise would not have, I am responsible for the consequences; whereas if I had merely told the truth as I knew it, perhaps the neighbors would have come and apprehended the killer.[225] But Kant's more Kantian reason is that the duty to tell the truth "makes no distinction between persons toward whom we have this duty, and toward whom we may be free from it; but it is an *unconditional duty* which holds in all circumstances."[226] Telling the truth to a would-be murderer may well become one link in a chain of events in which harm occurs, but only "accidentally."[227] Lying, however, is always wrong, because when I lie I am thereby attempting to manipulate another human being for my own purpose—in the case of Kant's murderer-at-the-door example, the purpose of saving my friend. Because the murderer cannot possibly agree to be lied to, he "cannot himself share [in] the end of the action."[228]

224. KrV, *supra* note 27, at 595 [A739/B767].

225. SRTL, *supra* note 27, at 362–63.

226. *Id*. at 364.

227. *Id*. at 365. Kant's point here is that harm to my friend is a kind of casualty of my truth-telling, rather than the direct result of a wrongful action. SRTL, *supra* note 27, at 364.

228. GMS, *supra* note 27, at 97 [429]. Kant goes on to state,

> This incompatibility with the principle of duty to others leaps to the eye more obviously when we bring in examples of attempts on the freedom and property of others. For then it is manifest that a violator of the rights of man intends to use the person of others merely

Kant's position thus seems to preclude issuing a warning to a potential victim over a potentially violent patient's objection, because doing so might involve breaking a promise (implicit or explicit) not to breach therapeutic confidentiality, and because the goal (the end) of the warning would not be an end shared by the patient. Even if the patient grudgingly acquiesced to the warning, the goal of the warning is still to interfere with the patient's desire, and thus is something to which he cannot logically give his assent. The same problems would apply, *a fortiori*, to warnings given to police (who would presumably act to interfere with the patient's intent) and to overt restrictions on freedom carried out through involuntary hospitalization.

c. Coercive Punishment[229]

Kant recognized, though, that society needs coercive punishments as a condition of freedom. It therefore is instructive to look to how he solves this analogous problem—the moral acceptability of punishment in criminal law—for some clues about when other coercive social practices might be tolerable. Kant's moral position requires him to justify criminal sanctions within a framework that centralizes the humanity of the criminal himself, that is, to show how political coercion can be justified within a moral context that requires individuals always to be treated as ends in themselves. Kant accomplishes this through application of his critical technique, arriving at a "transcendental deduction" of a regulative idea of reason. A regulative idea is a necessary goal of reason's efforts to organize experience[230] into a "systematic unity."[231] Ideas that allow reason to arrive at a "systematic

> as a means without taking into consideration that, as rational beings, they ought always at the same time to be rated as ends—that is, only as beings who must themselves be able to share in the end of the very same action.

Id. at 97 [430].

229. This section is adapted from Mossman, *Medically Appropriate, supra* note 176, at 65–72.

230. John Ladd, *Translator's Introduction* to KANT, THE METAPHYSICAL ELEMENTS OF JUSTICE xviii (New York, Macmillan 1965).

231. In the CRITIQUE OF PURE REASON, Kant explains that "transcendental" knowledge refers to knowledge "by which we know that—and how—certain representations (intuitions or concepts) can be employed or are possible purely *a priori*. KrV *supra* note 27, at 96 [A56=B80]. The term 'transcendental,' that is to say, signifies such knowledge as concerns the *a priori* possibility of knowledge, or its *a priori* employment." *Id*. at 96 [A56=B80]. The "transcendental deduction of all ideas of . . . reason" involves showing that these ideas are "rules of the empirical employment of reason [that] lead us to a systematic unity, under the presupposition of such an *object in the idea*; and that they thus contribute to the extension of empirical knowledge, without ever being in a position to run counter to it." *Id*. From such a deduction, we conclude that reason must "proceed always in accordance with such ideas." *Id*. at 550 [A671=B699].

unity" and achieve coherence are necessary rules that must govern the area of reason under consideration.[232]

The law is concerned with the effect that one person's choices and ensuing actions have on others. Because laws apply to everyone equally, they must be consonant with the freedom-maximizing "Universal Principle of Justice," which requires that "my action or my condition in general can coexist with the freedom of everyone in accordance with a universal law."[233] My having a right to act freely within the bounds of this restriction entails a right to prevent others from unjust hindrances of my freedom. Using coercion to counteract unjust hindrances to freedom "is consistent with freedom according to universal laws," and thus, any right I have to act in a permissible way "is united with the authorization to use coercion against anyone" that interferes with my right.[234] Having rights only means that use of coercion to enforce the right "is entirely compatible with everyone's freedom," including the freedom of the person against whom coercion is used, "in accordance with universal laws. Thus 'right' . . . and 'authorization to use coercion' mean the same thing."[235]

When, for example, I legally acquire a possession, I claim entitlement to use coercive force to defend my ownership right, at the same time acknowledging the legitimacy of all other persons' legitimate claims of ownership against me.[236] Such claims are possible, however, only in a society where laws protect ownership through public legislation, backed by coercive power, in civil society.[237] Kant therefore concludes that, as an "a priori" Idea of reason, people should participate in a legal system "if they ever could (even involuntarily) come into a relationship with one another that involves mutual rights,"[238] because it is only within such a system that one's ownership can "be established lawfully and secured . . . by an effective power" that is more than one's own mere physical capacity.[239] Living in civil society creates a better form of freedom than we would have if we relied only on our own personal strength to protect ourselves and our belongings by giving us freedom

232. Kevin Thompson, *Kant's Transcendental Deduction of Political Authority*, 92 KANT-STUDIEN 62, 66 (2001).

233. MAR, *supra* note 27, at 35 [230].

234. *Id*. at 36 [231].

235. *Id*. at 37 [232].

236. Thompson, *supra* note 232, at 74–75.

237. MAR, *supra* note 27, at 65 [256].

238. *Id*. at 70 [306].

239. *Id*. at 76 [312].

under laws that allow for assertion and protection of our rights.[240] This, in brief, is Kant's "deduction" of civil society, laws, and their coercive power as the condition of the possibility of the freedom to act in a world where people cannot avoid having contact with one another.[241]

Though this deduction has established the legitimacy of state coercion, it leaves open the problem of what types of coercion the state may impose on wrongdoers. Because systems of punishment must be consistent with dictates of interpersonal morality, they must recognize the humanity of the criminal and the victim equally. Punishment therefore must inflict on the criminal only the equivalent of what the criminal's unlawful act has inflicted on another person,[242] and punishments must be strictly retributive. Though punishment serves a coercive purpose (*i.e.*, deterring would-be criminals from violating laws that protect freedom-promoting relationships[243]), a criminal's guilt is the necessary and sufficient condition for a court's imposing a sentence. This assures that those criminals who undergo punishment merely experience the logical consequence of their decisions to break the law. Kant believes that if a legal system were to coerce someone for a reason other than his actually having committed a criminal offense (for example, to "punish"[244] him in order "to promote some other good for the criminal himself or for civil society"), it would be manipulating him "merely as a means to the purposes of someone else . . . ." The innate personhood of any individual, even an accused criminal, "protects him against such treatment . . . . He must first be found to be deserving of punishment before any consideration is given to the utility of this punishment for himself or for his fellow citizens."[245]

Kant thus accepts punishment as a rationally necessary consequence of our enjoying rights in civil society. Punishment is administered as a coercive response to the criminal's actions, but does not lessen the criminal's moral status because it confirms his humanity and freedom along with the humanity and freedom of all other members of civil society. Because Kant's theory precludes future-oriented goals such as deterrence and rehabilitation from entering into consideration of whether to punish someone, the legal system should not pretend to make

240. *Id*. at 80–81 [316].

241. Thompson, *supra* note 232, at 76–77.

242. MAR, *supra* note 27, at 133 [363].

243. *See* Thomas E. Hill, Jr., *Kant on Wrongdoing, Desert, and Punishment*, 18 LAW & PHIL. 407, 430 (1999).

244. The scare-quotes are placed here to signify that such an activity would not actually be punishment because it would not have been imposed in response to a crime.

245. MAR, *supra* note 27, at 100 [331].

punishment decisions based on the criminal's potential for reform or his likelihood of future misbehavior. Rather, the legal denunciation that occurs during conviction and punishment expresses our respect for the criminal's worthiness as a rational being by affirming his moral status as a responsible person who has acted for a reason.[246]

d. Application to Threats in Therapy

i. Threats as Hostile Acts

The considerations underlying Kant's justification of punishment provide guidance to evaluating laws like Ohio's duty-to-protect statute. Making credible threats is a criminal offense, and laws in many jurisdictions provide for punishment of individuals who do so.[247] From a Kantian standpoint, laws making such an action criminal and punishable appear legitimate because uttering a credible threat goes beyond being statistically dangerous and therefore poses a risk to society. One who threatens simultaneously performs a conscious action that really harms another person through fear-induced restriction in that other person's freedom. Threats, in other words, can be actions with adverse consequences and can therefore occasion a legitimate coercive response (*i.e.*, punishment) to the threatener.

Ohio's duty-to-protect statute deals with a circumstance not covered by criminal statutes on threats—the utterance of a threat to someone other than the target of the violent action. A response within the context of the criminal justice system therefore is not appropriate. Instead, the more logical response to a threat not yet carried out might be to counteract it somehow. As we have seen earlier, in jurisdictions where failure to help others is a criminal offense, the therapist who fails to do

246. Michael S. Moore, *The Moral Worth of Retribution*, *in* RESPONSIBILITY, CHARACTER, AND THE EMOTIONS 179, 198–217 (Ferdinand Schoeman ed., 1987).

247. *See, e.g.*, CAL. PENAL CODE § 422 (West 2006) (prescribing punishment of up to one year for threatening to commit a crime that will cause "death or great bodily injury to another person," that conveys "an immediate prospect of execution of the threat," and thereby causes reasonable fear for the target's safety); OHIO REV. CODE § 2903.21 (West 2006) (defining the misdemeanor or felony of "aggravated menacing" as "knowingly caus[ing] another [person] to believe that the offender will cause serious physical harm to" another person, the person's property, or person's fetus or family member). *See also, e.g.*, WYO. STAT. ANN. § 6-2-505 (Lexis 1999) ("terroristic threat," defined as threat "to commit any violent felony with the intent to cause evacuation . . . or otherwise to cause serious public inconvenience" punishable by three years' imprisonment); MINN. STAT. ANN. § 609.713 (West 2006) (five years' imprisonment for similar activity); MODEL PENAL CODE § 211.3 (2005) (third-degree felony to threaten "to commit any crime of violence with purpose to terrorize another or to cause evacuation of a building . . ." or other major inconvenience).

something to avert danger might be criminally liable.

ii. Implications of Not Responding

Even in jurisdictions that follow the common law tradition of not requiring individuals to rescue others, the therapist who hears a patient's credible threat would seem to be faced with only one of two choices: make some response that would address the threat, or not do so.

Whatever the Kantian objections might be to taking protective steps beyond verbal efforts at persuasion, not trying to do anything creates ethical problems, too. First, it is conceivable that failure to do more, when one has a legal requirement such as a statutory or court-created duty to protect, is tantamount to complicity if the patient carries out the threat. Typically, complicity in a criminal offense requires that an individual solicit, conspire with, or cause another person to commit the offense, or aid or abet another person in committing the offense, or in some way ally oneself with the aim of the person committing the offense[248]—which clearly is not the therapist's intent by maintaining silence. From a therapeutic standpoint, however, a therapist's failure to take a stance against potential violence when it is possible to intervene has the psychological effect of involving therapist in the patient's violent fantasies and colluding with them.

This leads to the second ethical problem, which is that in an important sense, the therapist who allows a patient to harm someone is failing to live up to the commitment the therapist has made to promoting the patient's health. As Dr. Gutheil explains:

---

248. *See, e.g.,* OHIO REV. CODE § 2923.03 (Anderson, 2005). "[T]he majority view—which . . . has been adopted under the Model Penal Code—holds that 'traditional definitions of accomplice liability demand that [the accessory] in some sort [of way] associate himself with the venture, that he participate in it as in something that he wishes to bring about, that he seek by his action to make it succeed.' " Kyong Suk Lee v. Anchorage, 70 P.3d 1110, 1112 (Alaska Ct. App. 2003) (citing, *inter alia*, MODEL PENAL CODE § 2.06).

In the article that he has prepared for this symposium, Professor Slobogin construes the Model Penal Code and case law as potentially making a therapist criminally liable for harm caused by a patient. *See* Christopher Slobogin, Tarasoff *as a Duty to Treat: Insights from Criminal Law*, 75 U. Cin. L. Rev. (forthcoming Winter 2006). My reading of the Model Penal Code, however, is consistent with the reading in *Lee,* above. That is, A can be guilty of an offense committed by B, another individual person for whom A is legally accountable (MODEL PENAL CODE § 2.06(1). But A is legally accountable for the conduct of B only if (1) A has caused B, an innocent or irresponsible person to engage in such conduct, or (2) A is made criminally accountable for B's conduct by the Code or by the law defining the offense, or (3) A is an accomplice of B (MODEL PENAL CODE § 2.06(2)). Someone who refrains from carrying out a legal duty can be deemed an accomplice only if his purpose in refraining was to promote or facilitate the offense (MODEL PENAL CODE § 2.06(2)(a)(iii).

> Because the clinician works, not for the patient, but for healthy side of the patient, the use of a *Tarasoff* warning may be seen to take place in service to that side of the patient that wishes not to harm another person. This posture supports a moral justification for a *Tarasoff*-type warning: The therapist acts at the unexpressed "behest" of the patient's healthy side.
>
> * * *
>
> [P]reventing the patient from harming a victim, though clearly beneficial to the victim, is also beneficial to the patient him or herself and fulfills a duty; the patient is spared the emotional, legal, and social consequences of having harmed another, perhaps while mentally impaired.[249]

Implicit in Dr. Gutheil's reference to a patient's "healthy side" is the idea that a person's full (and morally relevant) set of desires or intentions may be more extensive and complex than what the person happens to express at a particular (and perhaps a particularly weak) moment. From a Kantian viewpoint, a therapist (along with everyone else) is required to respect and promote the humanity of his patient (along with everyone else).[250] Therapists take on patients not to just make them feel better (for in this case, prescribing euphoria-inducing drugs would be a perfect treatment), but in the belief that treatment can allow patients to achieve legitimate goals such as to function more autonomously.[251] One could argue, then, that a therapist who fails to take available steps to offset a patient's violent plans has failed to fulfill the duty to help the patient preserve his own autonomy.

This last point leads to a final set of considerations. Even if a therapist is convinced that it is proper to take one of the extratherapeutic courses of action set out in the Ohio duty-to-protect statute, something about doing this "feels" wrong. The reason may be that, in an ideal therapeutic situation, such actions *would be* wrong. Kant's ethics tell us why—the therapist is interfering with the patient's free action, and ideally, this should not occur. But as Professor Korsgaard points out, the problem here may be "that morality itself sometimes allows or even requires us to do something that from an ideal perspective is wrong."

249. Gutheil, *supra* note 2, at 349.

250. In fact, Kant holds that the duty to treat "*humanity as an end in itself*" entails the positive obligation that "every one endeavours also, so far as in him lies, to further the ends of others. For the ends of a subject who is an end in himself must, if this conception is to have its *full* effect in me, be also, as far as possible, *my* ends." GMS, *supra* note 27, at 98 [430].

251. In Kantian terms, the therapist seeks to advance the patient's "capacities for greater perfection which form part of nature's purpose for humanity in our person." *Id*. at 97–98 [430]. The therapist does not make the patient feel better, but promotes the patient's development because, "as a rational being," the patient "necessarily wills that all his powers should be developed, since they serve him, and are given him, for all sorts of possible ends." *Id*. at 90 [423].

Though Kant's ethics describe right action in "an *ideal* system, . . . we need special principles for dealing with evil."[252]

iii. Life in the Real World

Kant's kingdom of ends informs us about what sorts of interactions we may have with others in an ideal realm where everyone acts justly. But to figure out how to act in the real world, we must contend with the fact that not everyone will comply with rules that promote mutual freedom. Professor Korsgaard notes, "Certain ongoing natural conditions . . . prevent the full realization of the ideal state of affairs . . . the problems of dealing with the seriously ill or mentally disturbed, for instance, belong to this category."[253] Under such circumstances, we tolerate and endorse behavior that falls short of what the Kantian ideal would require if such behavior will foster conditions that bring matters closer to that ideal.

In the present context, an extratherapeutic protective action such as an involuntary hospitalization temporarily deprives a patient of freedom. Yet this action is justified because by averting violence, the action brings the world closer to an ideal in which no one improperly impinges (through violence or other means) on the freedom of others. The point is not merely that hospitalization averts violence, which is something that is generically bad,[254] but that hospitalization as a liberty-restricting interaction is acceptable because if the world had no way to respond to threats, threats would create a far greater restriction of liberty because of the fear they would introduce.[255] Involuntary hospitalization is not undertaken lightly because confining someone involuntarily represents a departure from ideal conduct and from how we ordinarily think about a person's autonomy. As Professor Korsgaard observes, "[r]egret for an action we would not do under ideal circumstances seems appropriate even if we have done what is clearly the right thing."[256]

The notion that we should treat the patient as an end in himself can still serve as the therapist's goal even if the therapist takes action intended to thwart the patient's efforts. If one of the versions of the categorical imperative requires us to regard everyone as a universal

252. Korsgaard, *supra* note 222, at 327.

253. *Id.* at 342.

254. This would be a consequentialist argument for involuntary hospitalization, and asserting this alone would be arguing for treating the patient as a means (*i.e.*, restricting his freedom to benefit others).

255. *Cf.* Korsgaard, *supra* note 222, at 343–46 (discussing the role of a double-level theory in dealing with the non-ideal world).

256. Korsgaard, *supra* note 222, at 346.

legislator, then the therapist must respect the need for the patient to so regard others. Letting the patient harm someone else would allow him to deny the humanity of another individual and would require ignoring that other individual's humanity as well. It seems incoherent to argue that a therapist should do nothing to stop a patient's plan to eradicate a source of human value out of respect for the patient's humanity.[257]

VIII. CONCLUSION

This Article has derived a "major premise" under which the facts giving rise to the *Tarasoff* decision could result in tort liability if a therapist failed to take some action to intervene and harm later ensued. The major premise has two main parts: first, the patient commits the type of action that Poddar committed, namely, make a direct, credible threat to harm an identified person; second, the therapist fails to select one of several specific, pre-specified responses to the patient's statement. Using this major premise to characterize why situations such as those of Prosenjit Poddar and Tatiana Tarasoff demand therapist intervention has the advantage of being relatively uncontroversial, in that the expectations placed upon mental health clinicians conform to what several states' statutes and the American Psychiatric Association's model statute[258] recommend. But if this Article accomplished nothing more than justify already existing laws and recommendations, it might be little but an empty academic exercise.

I think, however, that this Article's argument, if correct, has practical benefits for how courts and therapists think about what mental health professionals do, and potentially, for certain legal and scientific problems that remain despite the passage of three decades since *Tarasoff*.

First, the wording of the *Tarasoff* ruling[259] amounts to a requirement that a clinician assess whether a patient will act violently and then decide, based on that calculation, whether some further protective action is warranted. This requirement, which went beyond what the facts in *Tarasoff* demanded, paved the way for subsequent courts to expect mental health professionals to gauge the risk of and take responsibility for any (retrospectively) foreseeable harm caused by a patient. The problem with this requirement is not, as was once believed, that mental

257. *Cf.* Korsgaard, *supra* note 222, at 347 (discussing how the formula of humanity creates a clear argument against suicide).

258. Appelbaum et al., *supra* note 14, at 827–28.

259. I here refer to the California Supreme Court's ultimate majority ruling, *Tarasoff II,* 551 P.2d 334, 340 (Cal. 1976), as well as the portions of the decision discussed *supra* Section II.C.1.b.

health professionals cannot "predict dangerousness." The problem with the *Tarasoff* rule is that it presupposes that assessments of dangerousness are yes-or-no predictions, whereas what mental health clinicians have is the ability to assign persons to different levels of risk. To take action based on a level of risk requires, in turn, a judgment about what level of risk is sufficient to justify the action.

But no court has provided guidance as to what this level of risk is, and, as a few empirical studies have now shown, homogeneous groups of people express irreconcilably broad ranges of opinion about what level of risk justifies even a very specific intervention such as involuntary hospitalization. This disagreement (along with other statistical uncertainties) prevents a clinician from implementing decisions by referring to a threshold that designates the probability of violence that should justify an intervention—no accepted threshold exists. The *Tarasoff* ruling, in requiring therapists to assess future risk and to act upon perceptions about that risk, implies use of a decision-making scheme about which there can be no agreement. What is wrong with *Tarasoff* scientifically is not that therapists cannot make accurate judgments about the future, but that they cannot know when to act protectively based on their probabilistic judgments about the future.

Second, *Tarasoff* rests upon a troubling rationale, which involves consequentialist policy decisions articulated by courts. Even if one accepts consequentialism as a basis for legal decisions, *Tarasoff* and the decisions that have followed it simply insist, rather than prove, that the benefits to society of added safety outweigh the costs borne by patients in terms of lost privacy, embarrassment, disruption of therapy, and involuntary confinement. As a practical matter, three decades have shown that *Tarasoff* has not been (as some had feared) a "disaster"[260] for mental health treatment, whether or not society has been made safer by making therapists potentially liable for their patients' violence. As a moral matter, however, *Tarasoff* is troubling in its willingness to sacrifice the interests of patients for the sake of society. The same notions of fairness and justice that prevent us from imposing confinement on people because they might commit future crimes also tell us that undeserving patients should not suffer adverse consequences for things they only have a probability of doing.

Third, the reasons why the APA proposed its legislative solution and the rationale for its adoption in several states have centered on the practical (and understandable) concerns of mental health professionals.

260. STONE, *supra* note 13, at 181.

As this Article has explained,[261] clinicians have felt that, as defined by courts, *Tarasoff*-type duties were amorphous and overly burdensome. Clinicians have therefore sought statutory boundaries on the duty to protect, boundaries that tell them when the duty arises (usually, following explicit threats toward specific targets) and that define specific ways of discharging the duty. Even if legislators take the concerns of mental health professionals to heart, however, there is no reason that courts will. Across the United States, state supreme courts have struck down as unconstitutional numerous legislative efforts at "tort reform."[262] In Arizona, a law restricting therapist liability to situations in which the patient has communicated an explicit threat of imminent action and the clinician fails to take reasonable precautions[263] has been held to violate the state constitution because it abrogated the common law right to recover for negligence.[264] Although some courts in other jurisdictions have held that duty-to-protect statutes superseded any common law rule,[265] the Arizona experience shows that the simple existence of a law is no guarantee that courts will accede to the hopes or needs of mental health professionals.

All three of these problems are addressed by this Article's argument for specifying the duty to protect as the Ohio statute does. Because the duty arises only when patients utter credible threats that they can carry

261. *See supra* Part VII.

262. *E.g.,* State *ex rel*. Oh. Acad. of Trial Lawyers v. Sheward, 715 N.E.2d 1062 (Ohio 1999) (caps on non-economic and punitive damages violate separation of powers); Knowles v. U.S., 544 N.W.2d 183 (S.D. 1996) ($1 million medical malpractice compensatory damage cap violates substantive due process); Arneson v. Olson, 270 N.W.2d 125, 135–36 (N.D. 1979) ($300,000 limit on damages recoverable in medical malpractice actions violates state and federal equal protection guarantees).

263. ARIZ. REV. STAT. ANN. § 36-517.02 (2006).

264. Little v. All Phoenix S. Cmty. Mental Health Ctr., Inc., 919 P.2d 1368 (Ariz. Ct. App. 1995). The court reasoned that common law recognizes a general negligence cause of action, and, under *Hamman v. County of Maricopa*, 775 P.2d 1122 (Ariz. 1989), general negligence includes injury to persons who are in the "reasonably foreseeable area of danger," and not just to those whom someone has threatened. Under Article 18, section 6 of the Arizona Constitution, a statute may not eliminate a common law cause of action. Thus, the court held that the portion of Arizona's duty-to-protect statute requiring a threat was unconstitutional.

265. Tabor v. Veteran's Admin. *ex rel*. U.S., 198 F.3d 247 (6th Cir. 1999) (unpublished table decision) (no liability because Tennessee statute required an actual threat); Riley v. United Health Care of Hardin, Inc., 165 F.3d 28 (6th Cir. 1998) (unpublished table decision) (no specific threat: statute superceded previous common law duty). In other cases, courts have ruled in such a way that duty-to-protect statutes have afforded therapists some protection, without ruling on whether those statutes supercede common law. *See, e.g.*, Jenks v. Brown, 557 N.W.2d 114 (Mich. Ct. App. 1996) (no need to use a common-law theory of negligence by defendant, because the patient had not "communicated [any] . . . threat of physical violence against plaintiff"), *and* Swan v. Wedgwood Christian Youth & Family Serv., Inc, 583 N.W.2d 719, 724–25 (Mich. Ct. App. 1998) (holding that the court had no need to decide whether a common-law duty survived the enactment of Michigan's statute because the defendant never had reason to foresee danger to the plaintiff's decedent).

out, the statute takes therapists out of the violence prediction business. Although the purpose of doing this was to relieve clinicians of an onerous burden, the statute also eliminates the need for clinicians to make a level-of-risk decision about which there is and can be no practical guidance. By requiring that a patient's action be the trigger for the protective duty and by specifying response to the patient's action, the statute treats patients as human beings. The statute tells therapists to regard their patients as making choices that have consequences, rather than requiring therapists to regard patients as sources of statistical risk. The statute respects patients as free actors whose expressed intentions are taken seriously and can trigger responses from other free actors. Under the statute, these responses include actions by therapists to counteract patients' expressed intent to violate other human beings' freedom.

Courts in some jurisdictions may construe common law or state constitutional provisions as encompassing a duty to protect that must survive legislative complete abrogation. However, jurisdictions need not repeat the scientific flaws and moral problems inherent in the liability-creating major premise created by *Tarasoff*. That is, jurisdictions can construe a common-law duty to protect as arising in the fact situation presented in *Tarasoff*, where a patient uttered a credible threat to harm a specific individual, without going further and stating that therapists must make predictions about every patient's likelihood of violence.

In this Article, I have emphasized the benefits of Ohio's duty-to-protect statute in a way that I hope will appeal to legislators, courts, and those who affect the thinking of these decision-makers. But mental health clinicians have something to gain from statutes such as the one enacted in Ohio that goes beyond having a clearly delineated duty and liability protection: a better understanding of the role of studies of violence prediction. As I have argued here and elsewhere,[266] in most cases, violence prediction techniques are not accurate enough to affect decisions in clinical management because the differences between patients with "low" and "high" risk usually are not big enough to justify treating them differently.

This does not mean that mental health professionals should not conduct studies of violence prediction or be interested in such studies' results, nor does it mean that clinicians should not make interventions that reduce violence. For example, studies examining post-

266. Douglas Mossman, *Commentary: Assessing the Risk of Violence – Are "Accurate" Predictions Useful?,* 28 J. AM. ACAD. PSYCHIATRY L. 272, 280 (2000).

hospitalization violence suggest that nonadherence to medication and (especially) substance abuse are risk factors for aggressive behavior during the months after hospital discharge.[267] Studies have also confirmed that—as was true in many of the cases described in this article—friends and family members are those most likely to become victims of violence by psychiatric patients.[268] It has been difficult to demonstrate whether treatment or other risk management strategies actually reduce violence. However, studies have shown that outpatient commitment and assiduous community follow-up may improve outpatient outcomes and increase the chances that patients will continue their treatment after hospitalization,[269] and accumulating research is indicating that the newer, "atypical" antipsychotic drugs may reduce aggression in persons with schizophrenia.[270] These and similar findings over the past decade support the belief that continuing to study violence committed by individuals with mental problems may "augment our understanding of the risk factors for violent behavior" and "improve the ability of clinicians, courts, and criminal justice staff to make informed decisions about treatment."[271]

But hopefully, all clinicians believe that better treatment for patients is desirable whether or not such improvements reduce the risk of violence. Effective treatment and helping to assure that patients receive it are good things because they enhance patients' autonomy. A clinician's well-founded belief that a patient needs and deserves certain

267. Marvin S. Swartz et al., *Violence and Severe Mental Illness: The Effects of Substance Abuse and Nonadherence to Medication*, 155 AM. J. PSYCHIATRY 155 (1998); Henry J. Steadman et al., *supra* note 86; *see also* Gerald Melnick et al., *Use of the COVR in Violence Risk Assessment,* 57 PSYCHIATRIC SERVICES 142 (2006) (reanalysis of MacArthur data shows correlation between violence and intensity of substance use).

268. Steadman et al., *supra* note 86; Kenneth Tardiff et al., *A Prospective Study of Violence by Psychiatric Patients after Hospital Discharge*, 48 PSYCHIATRIC SERVICES 48 (1997); Sue E. Estroff et al., *Risk Reconsidered: Targets of Violence in the Social Networks of People with Serious Psychiatric Disorders*, 33 SOC. PSYCHIATRY PSYCHIATRIC EPIDEMIOLOGY 95 (1998, Suppl. 1).

269. Virginia A. Hiday & T. L. Scheid-Cook, *A Follow-up of Chronic Patients Committed to Outpatient Treatment*, 40 HOSP. COMMUNITY PSYCHIATRY 52 (1989); Marvin S. Swartz et al., *Can Involuntary Outpatient Commitment Reduce Hospital Recidivism?: Findings from a Randomized Trial with Severely Mentally Ill Individuals*, 156 AM. J. PSYCHIATRY 1968 (1999).

270. Jeffrey W. Swanson et al., *Effectiveness of Atypical Antipsychotic Medications in Reducing Violent Behavior among Persons with Schizophrenia in Community-based Treatment*, 30 SCHIZOPHRENIA BULL. 3 (2004); Jeffrey W. Swanson et al., *Reducing Violence Risk in Persons with Schizophrenia: Olanzapine Versus Risperidone*, 65 J. CLINICAL PSYCHIATRY 1666 (2004) (in "real world" conditions, olanzapine is superior to risperidone in reducing violence risk, in part because of better adherence); Eric Elbogen et al., *Violence Risk Management and Adherence to Treatment With Atypical Antipsychotics in Schizophrenia*, 24 BEHAV. HEALTH MGMT. S1, S2–S3 (Nov./Dec. 2004) (explaining studies and pharmacological theory of drugs' anti-aggressive action).

271. Melnick et al., *supra* note 267, at 142.

treatments to function better is sufficient grounds by itself to motivate the clinician to see that a patient gets those treatments and to justify making arrangements for them, and the potential of such treatments to reduce the patient's violence risk should be an at-most-small factor in a clinician's decision-making.

Studies of violence prediction may alert clinicians to specific risk factors that can become the focus of treatment.[272] In the world of clinical mental health care, however, the purpose of learning such risk factors is that they matter to patients' well-being, and not because they will allow clinicians to make better predictions. Sound clinical interventions may be turn out to be socially useful because they reduce violence potential. But for mental health professionals, protecting the public should be an incidental result of the autonomy-enhancing effects of effective psychiatric treatment. For clinicians, the lesson of this article should be that a proper framing of the duty to protect will allow them to refocus their attention on providing effective treatment rather than on making predictions about violence.

272. As has resulted from research on factors increasing the risk of suicide. *See, e.g.,* Richard C. W. Hall et al., *Suicide Risk Assessment: A Review of Risk Factors for Suicide in 100 Patients Who Made Severe Suicide Attempts; Evaluation of Suicide Risk in a Time of Managed Care*, 40 PSYCHOSOMATICS 18, 18 (1999) ("Severe anxiety, panic attacks, a depressed mood, . . . were excellent predictors of suicidal behavior"); Katie A. Busch et al., *Clinical Correlates of Inpatient Suicide*, 64 J. CLINICAL PSYCHIATRY 14, 14 (2003) (severity of anxiety and agitation may help identify patients at acute risk for suicide and indicate possible treatment interventions).

APPENDIX I

The expected utility, *EU*, of a decision is:

[A1] $$EU = (BR)(TPR)(U_{TP}) + (BR)(1–TPR)(U_{FN}) + (1–BR)(FPR)(U_{FP}) + (1\text{-}BR)(1–FPR)(U_{TN}),$$

where *BR*, *TPR*, and *FPR* have the same meanings as in Table 1, and U_{TP}, U_{FN}, U_{FP}, and U_{TN} are the utilities of true positive, false negative, false positive, and true negative outcomes, respectively.

From Equation 1 in the text, setting $B = 1$,

[A2] $$Z_{TPR} = Z_{FPR} + A.$$

Because Z_{TPR} and Z_{FPR} are the normal deviates of *TPR* and *FPR*,

[A3] $$TPR = \Phi(Z_{TPR}) = \Phi(Z_{FPR} + A) \text{ and } FPR = \Phi(Z_{FPR}),$$

where $\Phi(\cdot)$ is the cumulative normal distribution function:

[A4] $$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{t^2}{2}} dt.$$

It is customary to assign utilities values between 0 (for the worst) and 1 (for the best) outcomes. Let us assume that TP and TN outcomes—correct identifications of violent and nonviolent persons—are equally good. Setting $U_{TP} = U_{TN} = 1$, Equation A1 becomes

[A5] $$EU = (BR)(TPR) + (BR)(1–TPR)(U_{FN}) + (1–BR)(FPR)(U_{FP}) + (1–BR)(1–FPR).$$

We now differentiate Equation A5 with respect to Z_{FPR}:

[A6] $$\frac{\partial EU}{\partial Z_{FPR}} = \frac{1}{\sqrt{2\pi}} \left(\begin{array}{l} (1 - U_{FN})(BR) e^{-\frac{(Z_{FPR}+A)^2}{2}} + \\ U_{FP}(1 - BR) e^{-\frac{Z_{FPR}^2}{2}} - (1 - BR) e^{-\frac{Z_{FPR}^2}{2}} \end{array} \right).$$

To find the value of Z_{FPR} that maximizes *EU*, we set this derivative equal to 0, then rearrange terms:

$$[A7] \qquad \frac{e^{-\frac{(Z_{FPR}+A)^2}{2}}}{e^{-\frac{Z_{FPR}^2}{2}}} = \frac{(1-BR)}{BR}\frac{(1-U_{FP})}{(1-U_{FN})}.$$

Expanding the left side of Equation A7, and taking the natural logarithm of both sides,

$$[A8] \qquad -AZ_{FPR} - \frac{A^2}{2} = \ln\left(\frac{(1-BR)}{BR}\frac{(1-U_{FP})}{(1-U_{FN})}\right).$$

When Equation A8 is solved for Z_{FPR} and $A = 1$, we obtain

$$[A9] \qquad Z_{FPR} = \frac{1}{2} - \ln\left(\frac{(1-BR)}{BR}\frac{(1-U_{FP})}{(1-U_{FN})}\right).$$

Figure 5 shows that the central 80 percent of the time periods endorsed by mental health professionals lie between three and 816 days. Suppose an individual is indifferent between being attacked and undergoing a three-day hospitalization, and assume further that the result of an emergency commitment is three days. This means that U_{FP} (the utility of a false-positive diagnosis of "violent") is the same as the U_{FN} (the utility of a false-negative diagnosis of "nonviolent"), so that $U_{FP} = U_{FN} = 0$. Suppose further, as Figure 4 indicates, that the base rate of violence (BR) is 10 percent, or 0.1. Plugging these values into Equation A9, we obtain the result $Z_{FPR} = -1.7$. For the violence prediction instrument depicted in Figure 4, where the mean score of nonviolent persons is 45 and the standard deviation of their scores is 10, this is equivalent to a score of 62.

Suppose that an individual is indifferent between being attacked and undergoing a 816-day hospitalization. For this individual, the worst outcome clearly is a false-negative diagnosis of "nonviolent," so we assign U_{FN} the value of 0. How should we assign a value of U_{FP} for this individual? Because an 816-day hospitalization is 272 times as long as a three-day emergency hospitalization, we can estimate roughly that this individual is indifferent between (a) a lottery in which he has a $^1/_{272} = 0.00368$ chance of being hospitalized for 816 days, and a $(1 - {}^1/_{272}) = 0.99632$ chance of no hospitalization, and (b) undergoing a three-day hospitalization. For this individual, therefore, $U_{FP} = 0.99632$. Plugging these values in Equation A9 gives us $Z_{FPR} = 3.9$; for the violence prediction instrument depicted in Figure 4, this is equivalent to a score of 16.

APPENDIX II

In Figure 4, the scores of violent and nonviolent subjects are depicted as "bell-shaped" or normal distributions. The usual way to express the general form of normal probability distribution function is

$$[A10] \qquad y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}},$$

where X is the value at which the function is evaluated, and μ and σ are the mean and standard deviation of the distribution. In Figure 4, $\mu = 55$ for the violent subjects, $\mu = 45$ for the nonviolent subjects, and for both groups, $\sigma = 10$.

Let BR equal the base rate of violence in the population (which, for this example, is set at 0.1), and let p equal the probability of violence selected by the judge as justifying hospitalization. Then the cut-off or value of FVT corresponding to p will be that value of the FVT where the ratio of the height of the distributions will be $p/(1-p)$. That is,

$$[A11] \qquad \frac{p}{1-p} = \frac{BR}{(1-BR)} \frac{\frac{1}{\sqrt{2\pi(10^2)}} e^{-\frac{(FVT-55)^2}{2(10)^2}}}{\frac{1}{\sqrt{2\pi(10^2)}} e^{-\frac{(FVT-45)^2}{2(10)^2}}}.$$

Canceling out common factors, taking the natural logarithm of both sides, and rearranging terms in Equation A11, one finds:

$$[A12] \qquad \ln \frac{p}{(1-p)} \frac{(1-BR)}{BR} = \frac{(FVT-45)^2}{2(10)^2} - \frac{(FVT-55)^2}{2(10)^2}.$$

Expanding and further simplifying:

$$[A13] \qquad 200 \ln \frac{p}{(1-p)} \frac{(1-BR)}{BR} = 20FVT - 100.$$

Solving for FVT, one obtains:

$$[A14] \qquad FVT = 50 + 10 \ln \frac{p}{(1-p)} \frac{(1-BR)}{BR}.$$

To find the FVT score that corresponds to p when the base rate is BR, one simply plugs the appropriate values into Equation A14. When $BR = 0.1$, $p = 0.01$ corresponds to a FVT score of 26. If $p = 0.56$, the FVT score will be 74.