NOTE

A RESPONSE-DEPENDENT PERSPECTIVE ON THE THEORY OF INSANITY

Bruno Patrick Babij†

INTRODUCTION

Since at least the thirteenth century, when the English jurist Henry de Bracton excused persons who "lack their sense and reason" from criminal responsibility, judges and jurisprudents have attempted to describe the qualities of mind that make a person criminally irresponsible.[1] The effort to formulate a satisfactory insanity defense was especially vigorous in the second half of the twentieth century, when both judges and professional legal academics produced long and thoughtful studies of the merits and demerits of the various available for-

† B.A., Stanford University, 2018; M.Phil., University of Cambridge, 2019; J.D., Cornell Law School, 2022.

[1] Anthony M. Platt, *The Origins and Development of the "Wild Beast" Concept of Mental Illness and Its Relation to Theories of Criminal Responsibility*, 1 ISSUES IN CRIMINOLOGY 1, 5 (1965) (quoting Henry de Bracton).

mulations.[2] That effort continues today.[3] And yet—despite centuries of thought about the question, and at least six decades of sustained attention to it by leading judges, academics, and professional associations—there remains fundamental disagreement about the nature of legal insanity. Thus, in 2020, the Supreme Court in *Kahler v. Kansas* declined to find that due process guaranteed a defendant the right to present an affirmative insanity defense, finding that, given "continuing division over the proper scope of the insanity defense," no single rule of insanity was settled enough to have constitutional status.[4] As the Court observed, the proper formulation of the insanity defense was "replete with uncertainty"[5] and marked by "flux and disagreement."[6]

The Supreme Court's observations in *Kahler* confirm what is unmistakable upon even a brief study of the insanity literature: the efforts to specify the qualities of mind that make a person criminally irresponsible have not succeeded. The problem, in particular, is that every existing theory or test of insanity gives rise to unacceptable false positives or false negatives. A fruitless dialectic of insanity theory ensues when one scholar presents a *reductio ad absurdum* of all existing

---

2 *See, e.g.*, ROYAL COMMISSION ON CAPITAL PUNISHMENT 1949–1953 REPORT 112–16, 129 (Greenwood Press ed., 1953) (proposing that the *M'Naghten* rule be abrogated or enlarged); Durham v. United States, 214 F.2d 862, 874–75 (D.C. Cir. 1954) (repudiating the *M'Naghten* test and replacing it with a rule that "an accused is not criminally responsible if his unlawful act was the product of mental disease or mental defect"); Joseph M. Livermore & Paul E. Meehl, *The Virtues of M'Naghten*, 51 MINN. L. REV. 789 (1967) (defending the *M'Naghten* rule as the most adequate test of insanity); *United States v. Brawner*, 471 F.2d 969, 1032 (D.C. Cir. 1972) (en banc) (Bazelon, C. J., dissenting) (proposing that the *Durham* rule be replaced by a rule "that a defendant is not responsible if at the time of his unlawful conduct his mental or emotional processes or behavior controls were impaired to such an extent that he cannot justly be held responsible for his act" (emphasis omitted)); Walter Sinnott-Armstrong, *Insanity v. Irrationality*, PUB. AFFS. Q. 1, 1 (July 1987) (arguing that "what provides the excuse [of insanity] is not . . . irrationality but . . . incapacity to conform to law"); Stephen J. Morse, *Culpability and Control*, 142 U. PA. L. REV. 1587 (1994) (arguing that the fundamental excusing condition is irrationality, not a defect of control or volition).

3 *See, e.g.*, Michael Corrado, *The Case for a Purely Volitional Insanity Defense*, 42 TEX. TECH L. REV. 481, 482 (2009) (arguing that insanity is "a defect of will or a lack of control"); Michael S. Moore, *The Neuroscience of Volitional Excuse*, *in* PHILOSOPHICAL FOUNDATIONS OF LAW AND NEUROSCIENCE 179, 179 (Dennis Patterson, Michael S. Pardo eds., 2016) (attempting to lay foundation for a research program in neuroscience that might "help in the conceptualization of, and/or in the verification of the existence of, volitional excuse"); Stephen P. Garvey, *Agency and Insanity*, 66 BUFF. L. REV. 123, 126 (2018) (proposing that insanity be understood as a "*defect of consciousness*, and in particular . . . a *lost sense of agency*").

4 Kahler v. Kansas, 140 S. Ct. 1021, 1037 (2020).

5 *Id.*

6 *Id.* (quoting Clark v. Arizona, 548 U.S. 735, 752 (2006)).

theories and sets forth a purportedly superior one—except that theory, in its turn, becomes the subject of another scholar's *reductio*. As this dialectic proceeds, there is more writing, but no progress in understanding.

The purpose of this Note is to intervene in the apparently fruitless dialectic of insanity theory by suggesting the value of a new perspective on what makes a person criminally irresponsible—a perspective that has already proved enormously valuable in the philosophical theory of responsibility, but which, so far, has been neglected by students of legal insanity. This Note proposes, in particular, that the theory of insanity should take heed of the "unfamiliar but incredibly intriguing conceptual reversal" suggested by the philosopher P.F. Strawson in his 1962 essay "Freedom and Resentment."[7] Strawson's proposal, as it has been formulated by a later interpreter, is "that being responsible is a function of being held responsible."[8] On the Strawsonian view, to be held responsible is to be an object of moral sentimental responses such as anger, resentment, or indignation.[9] The received idea of responsibility, which Strawson and his followers sought to replace, is that we treat someone as morally responsible—we respond to that person as if that person were responsible—because of certain intrinsic facts about that person's being responsible.[10] In that sense, our treatment of her answers to, or is determined by, facts independent of our responses to her. According to the revision Strawson and his followers urge, however, facts about a person's being responsible are not intrinsic and independent of our moral sentimental responses but instead depend on, or are determined by, our responses themselves, or lack of them: a person is responsible if we treat her as responsible.[11]

If this conceptual reversal were adopted in the theory of insanity, we would cease to understand our treatment of some-

---

7 Neal A. Tognazzini, *Blameworthiness and the Affective Account of Blame*, 41 PHILOSOPHIA 1299, 1300 (2013); *see* Peter Strawson, *Freedom and Resentment*, *in* FREE WILL 72 (Gary Watson ed. 2003).

8 David Shoemaker, *Response-Dependent Responsibility; or, a Funny Thing Happened on the Way to Blame*, 126 PHIL. REV. 481, 481–82 (2017).

9 *See id.* at 493–94.

10 *Id.* at 483 ("The much more popular alternative view . . . is a response-*independent* view of responsibility, according to which there are antecedent properties of being responsible that our practices of holding responsible must respect and respond *to*.").

11 *See id.* at 481–82; Gary Watson, *Responsibility and the Limits of Evil: Variations on a Strawsonian Theme*, *in* AGENCY AND ANSWERABILITY: SELECTED ESSAYS 219, 227 ("In a Strawsonian view . . . nonpropositional responses are constitutive of the practice of holding responsible.").

one as criminally irresponsible as properly determined by certain facts about that person. (Traditionally, these facts have been supposed to be the incapacity to know right from wrong or the inability to have conformed one's conduct to the law.) Instead, we would view a person's being insane as determined by our moral sentimental responses or reactions to that person: a person is insane if we treat her as insane. This conceptual reversal promises to resolve enduring dissatisfaction with traditional "tests" for insanity, each of which, by calling upon a jury to treat a person as insane only when certain facts are true of that person, suspend the natural operation of the moral sentimental reactions or responses that determine who is insane. Because the traditional tests suspend the natural operation of the moral sentiments, either when a real jury applies them to the facts in the record or when insanity theorists apply them to the facts in a thought experiment, these tests readily and inevitably yield both false positives—judgments that a sane person is insane—and false negatives—judgments that an insane person is sane. This Note proposes that, instead of further prosecuting the fruitless inquiry into the essence of insanity, insanity theorists should recognize the virtue of a theoretical quietism. A person is insane if she is not an object of the moral sentimental responses, which are *causa sui* and explanatorily basic, and no more than that can be said.

This Note has three parts. The first introduces the idea of response-dependent responsibility in more detail. The second part argues that the traditional tests for insanity assume the view of responsibility the response-dependent account sought to correct and that the failure of these tests to provide satisfactory results follows from that assumption. The third part proposes that a quietism be adopted in the theory of insanity and considers what effect, if any, the quietist conclusion should have on the administration of the insanity defense in practice.

I
THE THEORY OF RESPONSE-DEPENDENT RESPONSIBILITY

Strawson's proposed conceptual reversal was an intervention in the philosophical controversy about what, if anything, could remain of our moral practices if causal determinism were true. In general, the thesis of causal determinism is that, given a state of the world at a time *t*, the state of the world at a later time *t'* is entailed by the conjunction of *t* and the laws of phys-

ics.[12] The truth of determinism might be inconsistent with our traditional understanding of human agency in at least two ways. First, if determinism were true, it might be that humans would not be free, in any meaningful way, to do otherwise. Second, if determinism were true, it might be that humans would not be, in any meaningful way, the sources or originators of their actions.[13] Those who accept either or both of these theses are generally called *incompatibilists* because they believe that the truth of determinism is incompatible with the existence of free will (whether defined as freedom to do otherwise or as the power to be an ultimate originator of action).[14] In the literature most relevant to this Note, incompatibilists are also understood to affirm a further thesis: that because free will is a necessary condition of moral responsibility, it follows from the truth of determinism that humans are not morally responsible agents.[15] *Compatibilists*, in contrast to incompatibilists, are traditionally understood to affirm that the truth of determinism is consistent or compatible with the existence of free will in some meaningful sense.[16] In the literature most relevant to this Note, compatibilists are also understood to affirm a distinct but related thesis: that the truth of determinism is compatible with the status of humans as morally responsible agents.[17] A compatibilist in this latter sense need not necessarily be a compatibilist in the former sense. In principle, one could accept that determinism is incompatible with free will but maintain nonetheless that determinism is compatible with moral responsibility. The compatibilist of this kind would deny that free will is a necessary condition of moral responsibility.

---

12 *See* Peter Van Inwagen, *The Incompatibility of Free Will and Determinism*, 27 PHIL. STUD. 185, 186 (1975).

13 *See* Michael McKenna & D. Justin Coates, *Compatibilism*, *in* STAN. ENCYC. PHIL. §§ 1.1-.2 (2019), https://plato.stanford.edu/entries/compatibilism/ [https://perma.cc/AXK6-U4A9].

14 *See id.* (rehearsing the *classical incompatibilist* argument, the premise of which is that alternative possibilities are necessary for free will, and the *source incompatibilist* argument, the premise of which is that free will requires that a person be the ultimate source of her actions).

15 R. JAY WALLACE, RESPONSIBILITY AND THE MORAL SENTIMENTS 3 (1994).

16 *See* McKenna & Coates, *supra* note 13, § 1.3 ("[C]ompatibilists have denied that freedom requires the ability to do otherwise; that causal determinism precludes the ability to do otherwise; and that freedom or control require sourcehood.").

17 This is the sense of "compatibilism" adopted by Wallace, who "call[s] an account of moral responsibility 'incompatibilist' if it affirms that moral responsibility requires that agents be strongly free, and 'compatibilist' if it denies this." WALLACE, *supra* note 15.

It is a compatibilism of this kind—a compatibilism that denies that determinism is incompatible with moral responsibility—that Strawson defends in "Freedom and Resentment." Strawson chooses to call compatibilism of this kind *optimism*. According to this optimism, our moral practices would remain vital even if the thesis of determinism were true. According to *pessimism*, however, our moral practices would lose all justification if determinism were true.[18] In defending optimism, Strawson disclaims the utilitarian thesis that moral practices are justified simply by their value in social regulation.[19] Instead, Strawson defends optimism by attending to the operation of what he calls the *reactive attitudes*.[20] Some reactive attitudes, such as gratitude, resentment, forgiveness, love, and hurt feelings, are proper to direct interactions with second parties.[21] In some circumstances, it is natural that these reactive attitudes be suspended and an objective attitude assumed in their place, as when "the agent is seen as excluded from ordinary adult human relationships by deep-rooted psychological abnormality—or simply by being a child."[22] Strawson then asks: "[C]ould, or should, the acceptance of the determinist thesis lead us always to look on everyone exclusively in this [objective] way?"[23]

Strawson provides two reasons to think that the answer is that it could not. The first is that a global suspension of the reactive attitudes on the basis of a "general theoretical conviction" such as the truth of determinism is "practically inconceivable."[24] On Strawson's view, "[a] sustained objectivity of interpersonal attitude, and the human isolation which that would entail, does not seem to be something of which human beings would be capable, even if some general truth were a theoretical ground for it."[25] The second reason is that when the reactive attitudes are suspended in a particular case, it is not as "the consequence of a theoretical conviction [that] might be expressed as '[d]eterminism in this case.'"[26] Strawson then proposes to apply these lessons about the reactive attitudes proper to direct interactions with second parties to the reactive atti-

18 Strawson, *supra* note 7, at 72.
19 *Id.* at 91–92.
20 *Id.* at 75.
21 *Id.*
22 *Id.* at 81.
23 *Id.*
24 *Id.*
25 *Id.*
26 *Id.* at 82.

tude felt with respect to a third party: moral disapprobation. When the attitude of moral disapprobation is suspended with respect to a person, it is not because of a theoretical conviction that determinism is true of that person. It is not the case that "some general metaphysical proposition is repeatedly verified . . . in all cases where it is appropriate to attribute moral responsibility."[27]

That is Strawson's negative contribution to the controversy about determinism and moral responsibility: the attribution of moral responsibility to a person does not depend on a theoretical conviction that some metaphysical proposition is true of that person. Implicit in this negative claim about what the attribution of moral responsibility does *not* depend on, however, is a positive claim about what the attribution of moral responsibility *does* depend on. Strawson's counterintuitive suggestion in this regard is that "reactive attitudes and practices themselves . . . are constitutive of responsibility."[28] Or, as David Shoemaker has formulated the claim: "Strawson maintains that being responsible is a function of being held responsible, that is, it is somehow a function of being a target of" the reactive attitudes.[29] This claim is that facts about being responsible are *response-dependent*.[30] Both Shoemaker and another prominent Strawsonian, R. Jay Wallace, have recognized that one (unsatisfactory) way of formulating the response-dependence thesis is in dispositional terms: that a person is morally responsible if and only if we are disposed to hold that person morally responsible.[31] The main defect of this formulation is that it cannot account for why a disposition to hold responsible a certain class of people—the intellectually disabled, for example—may be typical of a community but nonetheless *wrong*. The critic of the dispositional analysis will necessarily ask: "Does the mere fact that certain attitudes are taken toward an agent establish that he is an appropriate candidate for this treatment?"[32] The dispositional account provides no ready answer.

Wallace therefore proposes that a standard of *appropriateness* be incorporated in the analysis of the response-dependence thesis: a person is morally responsible if it would be

---

27 *Id.* at 92.

28 PERSPECTIVES ON MORAL RESPONSIBILITY 16 (John Martin Fischer & Mark Ravizza eds.,1993).

29 Shoemaker, *supra* note 8, at 481–82.

30 *See id.* at 482.

31 WALLACE, *supra* note 15, at 89; Shoemaker, *supra* note 8, at 496.

32 PERSPECTIVES ON MORAL RESPONSIBILITY, *supra* note 28, at 18.

appropriate to hold that person responsible.[33] Wallace hopes that the standard of appropriateness will allow a principled way of assessing the correctness or incorrectness of holding someone responsible, but "without postulating a prior and independent realm of moral responsibility facts."[34] Because the norm of appropriateness is general, Wallace proposes to analyze appropriateness in terms of the purportedly more definite norm of *fairness*.[35] On Wallace's account, then, it is appropriate to hold someone responsible when it is *fair* to hold someone responsible.[36] This formulation is not subject to the objection raised above to the dispositional analysis because it does provide grounds to say that holding certain people responsible is wrong. But neither is it properly a formulation of the *response-dependence* thesis about moral responsibility. Facts about what is fair are not in any meaningful way determined by our reactive attitudes. They are, rather, fixed by normative theory.[37] By subordinating the reactive attitudes to facts ascertained by normative theory, Wallace does not adhere to the Strawsonian conceptual reversal, which, roughly put, explains facts about responsibility-conditions—whether normative or theoretical—by the reactive attitudes, and not the other way around. The pessimist Strawson had sought to discredit wished for "some general metaphysical proposition . . . verified in all cases where it is appropriate to attribute moral responsibility."[38] So too does Wallace, in asking for satisfaction of the proposition "fair" in all cases of appropriate attribution of moral responsibility.[39] It is not therefore in Wallace's account that the Strawsonian conceptual reversal has its true expression.

---

33 WALLACE, *supra* note 15, at 91.

34 *Id.* at 92.

35 *Id.* at 93–95.

36 *Id.* at 103–09.

37 *See, e.g.*, Leif Wenar, *John Rawls*, *in* STAN. ENCYC. PHIL. § 4.6 (2021), https://plato.stanford.edu/archives/sum2021/entries/rawls/ [https://perma.cc/7HEQ-6C86] (describing Rawls' notion of the original position, a thought experiment intended to specify the "fair terms of social cooperation for free and equal citizens").

38 Strawson, *supra* note 7, at 92.

39 Wallace frankly acknowledges that defining the conditions of moral responsibility as those in which it *fair* to hold someone responsible is a gambit foreign to the Strawsonian perspective. WALLACE, *supra* note 15, at 103. At least one other commentator has recognized that "Wallace does not articulate a 'reversal' thesis." Patrick Todd, *Strawson, Moral Responsibility, and the "Order of Explanation": An Intervention*, 127 ETHICS 208, 210 (2016); *cf.* David Shoemaker, *Response-Dependent Theories of Responsibility*, *in* THE OXFORD HANDBOOK OF MORAL RESPONSIBILITY (Dana Kay Nelkin & Derk Pereboom eds., 2022) ("[M]any claim[ ] to be Strawsonians without actually embracing Strawson's fundamental program, namely, its

Shoemaker provides a more properly Strawsonian response to the objection to the dispositional analysis raised above.[40] Shoemaker's account has two components. The first is a *negative* argument that an account of blameworthiness as a function of response-independent properties yields both false negatives—judgments that a blameworthy person is not blameworthy—and false positives—judgments that a blameless person is blameworthy.[41] By attending to the ways in which the response-independent account produces false negatives and false positives, Shoemaker hopes to show that "[i]t is difficult to see what response-independent natural features might impose unity on the wide array of activities and attitudes we deem blameworthy."[42] This negative claim is preliminary to Shoemaker's *positive* argument: that "the better account—simpler, more plausible, and with greater explanatory value—is that our *emotional responses themselves* are what impose unity on" the variety of acts and people we deem blameworthy.[43] Shoemaker attempts to meet the skeptical objection to the dispositional account above by the notion of *angerworthiness*, disagreements about which are settled not by reference to facts about the object of anger (the person whose blameworthiness is at issue) but by reference to facts about the subject of anger (the person who may feel anger and then assess blame).[44] Because disagreements about the propriety of a reactive attitude such as anger are settled by reference to the subject rather than the object of anger, the wrongness of certain reactive attitudes need not be explained by response-independent properties.[45] Shoemaker's account of response-dependent responsibility benefits from an analogy with humor. He observes that the class of things that are funny cannot be accurately defined by reference to response-independent properties of the candidates for humor.[46] Instead, this class is accurately defined only by our responses themselves.[47] So too, he would like to conclude, the class of responsible people is accurately defined, not by

---

response-dependent understanding of being responsible.") (citation and internal quotation marks omitted).

40 *See* Shoemaker, *supra* note 8, at 496–508.

41 *See id.* at 498–508.

42 *Id.* at 508.

43 *Id.*

44 *See id.* at 508–12.

45 *See id.*

46 *See id.* at 486–87.

47 *Id.* at 487–90.

response-independent facts about them, but by our responses themselves.[48]

Shoemaker begins his negative argument about responsibility by providing a list of acts that tend to arouse anger and that we are also likely to call blameworthy.[49] In keeping with the Strawsonian perspective, Shoemaker assumes that the moral sentimental response *anger* tends to pick out those people or acts that are blameworthy (and therefore responsible). The relevant question, then, is how to understand the relationship between the moral sentimental response *anger* and the quality of blameworthiness.[50] The response-independent theorist advances this thesis:

> The blameworthy consists in a property (or properties) of agents that makes anger at them appropriate, a property (or properties) whose value-making is ultimately independent of our angry responses. Anger at someone for X is appropriate if and only if, and in virtue of the fact that, she is antecedently blameworthy (and so accountable) for X.[51]

According to the response-independent theorist, then, "anger is, at most, a reliable epistemic tracker of the blameworthy."[52] But anger does not determine the conditions in which a person is blameworthy. It is then incumbent upon the response-independent theorist to specify the response-independent properties in virtue of which someone is antecedently blameworthy. As reported by Shoemaker, response-independent theorists have settled on three properties constitutive of blameworthiness: voluntariness, control, and knowledge.[53] The prevailing formulation of the response-independent conditions of blameworthiness is therefore that "[a]n agent is blameworthy—and *so as a result* merits anger—for something bad if and only if, and

---

48 *Id.* at 508.

49 Among the blameworthy acts Shoemaker lists are: "A Yankees fan at a Bronx bar punches you in the face because you are wearing a Red Sox cap"; "An employee of yours does everything you ask, but with obvious condescension"; "A workman tosses pieces of heavy slate off a roof to the ground below without checking to see if anyone is there, and he just misses hitting you as you walk by"; "A friend comes to visit you in the hospital, but when thanked, she responds, 'I only came because it is my Christian duty.'" *Id.* at 495.

50 *Id.* at 496.

51 *Id.* at 498 (emphasis omitted).

52 *Id.*

53 *Id.* at 499. Shoemaker defines a *voluntary* action as one *chosen* by the agent. An agent who lacks *control* "lack[s] a key modal property, namely, the ability to refrain from [certain] choices or activities, or to be responsive to alternative reasons." *Id.* at 498. Knowledge "is standardly thought to involve both knowledge of what one is doing and knowledge that what one is doing is wrong/bad." *Id.* at 499.

because, the agent generated it knowingly, voluntarily, and under control."[54] Shoemaker then observes that this formulation produces false negatives and false positives.[55] For example, someone who is nonculpably ignorant of the wrongness of racism is nonetheless blameworthy, notwithstanding lack of knowledge (false negative).[56] Someone who drives recklessly and kills someone is more blameworthy than someone who drives recklessly without causing harm, even though both are by hypothesis equivalent in knowledge, volition, and control (a false positive as to the driver who did not cause harm).[57]

If, as Shoemaker argues, the necessary and sufficient conditions of blameworthiness specified by the response-independent theorist do not accurately define the class of blameworthy acts or people, then there is reason to seek some other principle of definition.[58] Shoemaker proposes that this is the principle of "fitting response-dependence":

> The blameworthy (in the realm of accountability) just is whatever merits anger (the angerworthy); that is, someone is blameworthy (and so accountable) for X if and only if, and in virtue of the fact that, she merits anger for X.[59]

In simple terms, the fitting response-dependent theorist argues that the blameworthy is the anger-worthy. Now it is obvious that a response-independent theorist need not disagree that the blameworthy is the anger-worthy, if the anger-worthy is defined by reference to response-independent properties of the possible object of anger. It is in refusing this definition of anger-worthiness that the fitting response-dependent thesis acquires its distinctiveness. Anger-worthiness is not defined by reference to response-independent properties of the possible object of anger, but by reference to the proper operation of the "anger sensibility" possessed by the *subject* of anger.[60] Shoe-

---

54 *Id.* at 499 (emphasis omitted).

55 *Id.*

56 *See id.* at 501.

57 *Id.* at 507–08. Doubtless one can quarrel with Shoemaker about whether these and the other examples he provides are really the false negatives or false positives they purport to be or, if they are, about whether the response-independent account cannot be adjusted to accommodate them. I prescind from these questions in this Note, important as they are, because the purpose of the Note is not to defend the merits of Shoemaker's account but to suggest what might be true of the theory of insanity, if Shoemaker's account is substantively correct. A more complete discussion of the question of false negatives and false positives can be found in Shoemaker's article itself. *See id.* at 498–508.

58 *Id.* at 508.

59 *Id.* (emphasis omitted).

60 *Id.* at 511.

maker argues that it is not true of this analysis, as it was of the dispositional analysis, that it provides no principled way of explaining how in some cases an anger response can be *incorrect*. He observes that a party to a disagreement about what is anger-worthy will attempt to resolve this disagreement by "asking the other person to defend the refinement and development of his or her anger sensibility."[61] Because "we know that there are defective senses of anger . . . generated by dysfunctional human machinery, the product perhaps of coddled or brutalized youth," it is natural to ask, "What makes you a good judge in these matters? Why should I defer to your normative authority?"[62] To settle a disagreement about what is anger-worthy—and to come therefore to a conclusion that a certain anger response is inapt—is simply to evaluate the fitness of another person's anger sensibility. That this is so does not preclude appeal to response-independent facts about the putative object of anger. Appeal to these response-independent facts, however, serves only to ensure that a person's anger sensibility operates with full information. It remains that the aptness of anger is ultimately determined not by facts about the candidate object of anger, but by the (properly operating) anger sensibility of the candidate subject of anger.

Shoemaker's response-dependent thesis forecloses substantive theorizing about responsibility (or insanity). Reduced to its essence, Shoemaker's proposal is that *causa sui* (explanatorily basic) moral sentimental responses are, on any given occasion, the ultimate determinants of whether someone is responsible. This is a thesis about *in virtue of what* someone is responsible.[63] It is perfectly consistent with this thesis that more or less accurate generalizations be made about the circumstances under which the moral sentimental responses are likely to be suspended. (The conventional idea that a person is irresponsible if the person acted without knowledge, volition, or control can be understood as one such generalization.) But these generalizations—to the extent that they are consistent with the response-dependent thesis—are merely descriptive, because they do not purport to *explain why*, or *say in virtue of*

---

61 *Id.*

62 *Id.*

63 A claim that something exists *in virtue of* something else is a claim about metaphysical grounding. The grounding relation is thought to be one kind of *explanatory* relationship: to say that A exists *in virtue of* B else is to say that A is *explained by* B. *See* Ricki Bliss & Kelly Trogdon, *Metaphysical Grounding*, *in* STAN. ENCYC. PHIL. (2021), https://plato.stanford.edu/entries/grounding/ [https://perma.cc/9ZGN-FE69].

what, someone is responsible. If it is the office of a theory to give explanations, no theory can be given of responsibility, because responsibility is explained only by moral sentimental responses that are themselves *causa sui* (explanatorily basic).

II
PERPLEXITY ABOUT WHO IS INSANE HAS INEVITABLY ARISEN BECAUSE THE EXISTING TESTS OR THEORIES OF INSANITY HAVE ASSUMED THAT INSANITY IS A RESPONSE-INDEPENDENT CONDITION

Insanity theorists have so far assumed the correctness of the response-independent thesis: they have attempted "to understand in what insanity consists" and to formulate legal rules that "capture or define it."[64] To this end, insanity theorists have followed a common method, which is to test candidate theories against uncontroversial intuitions about whether a defendant in a particular case is responsible.[65] If a theory gives a result consistent with intuition, it is *prima facie* adequate; if it gives a result contrary to intuition, it is *prima facie* inadequate. Despite this common method, basic disagreement among insanity theorists persists, even after more than half a century of thoughtful and thorough commentary.[66] This disagreement arises because every existing theory or test of insanity readily yields false positives—judgments that a sane person is insane—or false negatives—judgments that an insane person is sane. An apparently fruitless dialectic ensues as one insanity theorist, having identified embarrassing consequences of all existing tests or theories, proposes a new and apparently more adequate account. But this new account is in its turn refuted by another theorist, and one more new account proposed in its place. One reasonable response to the unsatisfactory dialectic is to remain committed to the possibility of consensus: eventually, a theorist will more or less adequately state what insanity consists in, and the dialectic will be arrested. But the re-

64 Garvey, *supra* note 3, at 126 n.18.

65 *See, e.g.*, Corrado, *supra* note 3, at 510 ("[T]he question ought to be whether [the candidate test] tracks our moral intuitions and the facts as we know them."); Garvey, *supra* note 3, at 130 ("Ignore for a moment what you think the verdict on M'Naghten's sanity would be under this or that legal test, and consult your intuitions.").

66 *See, e.g.*, MICHAEL MOORE, PLACING BLAME: A GENERAL THEORY OF THE CRIMINAL LAW 598–603 (1997) (arguing that academic accounts of insanity are inadequate and defending an irrationality theory); Corrado, *supra* note 3 (defending a "purely volitional" account); Garvey, *supra* note 3 (arguing that existing accounts of insanity, including Moore's irrationality theory, are inadequate and proposing a theory of insanity as lost agency).

sponse-dependent intervention in the philosophical study of responsibility suggests the virtues of another response: to present principled grounds on which to stop practicing insanity theory. According to this response, traditional insanity theory has assumed, incorrectly, that insanity is determined by facts antecedent to or independent of our moral sentimental responses and has sought vainly to specify these facts in a form of words. On a revised view, however, insanity is determined by the moral sentimental responses themselves. On this view, the false negatives and false positives endemic to traditional insanity theory arise because insanity theory attempts to define the class of insane people according to response-independent facts. To accept instead that the class of insane people is defined according to our moral sentimental responses promises some measure of theoretical repose, or at least a plausible explanation of why that repose has been so far unachieved by traditional insanity theory.

The benefit of adopting a response-dependent reversal in the theory of insanity can be illustrated by a presentation and then resolution of a stylized dialectic in the theory of insanity. The first moment in this stylized dialectic is adherence to the *M'Naghten* test.[67] Contemporary formulations of *M'Naghten* generally identify insanity with a cognitive incapacity. In California, for example, a defendant is insane if "he or she was incapable of knowing or understanding the nature and quality of his or her act and of distinguishing right from wrong at the time of . . . the offense."[68] The *M'Naghten* test thus formulated presents the jury with an algorithm: if it is true that the defendant was unable to know that his act was wrong, find that he is insane. An obvious inadequacy in the *M'Naghten* test is that it fails to allow the insanity defense to a defendant who, by hypothesis, knew very well that an act was wrong, but could not help but commit the wrongful act.[69] When confronted with

---

67 A canonical academic formulation of and apology for the *M'Naghten* test is Livermore & Meehl, *supra* note 2.

68 CAL. PENAL CODE § 25(b) (West 2021).

69 *See* MODEL PENAL CODE § 4.01 cmt. 2 (AM. L. INST. 1985) ("

> A . . . more pervasive difficulty with the M'Naghten standard appears in cases in which the defendant's disorder prevents his awareness of the wrongfulness of his conduct from restraining his action. Stated otherwise, these are cases in which mental disease or defect destroys or overrides the defendant's power of self-control.

"); Durham v. United States, 214 F.2d 862, 872-73 (D.C. Cir. 1954) ("[I]n 1929, we reconsidered [the right-wrong test] in response to 'the cry of scientific experts' and added the irresistible impulse test as a supplementary test for determining criminal responsibility."); Corrado, *supra* note 3, at 506–08.

such a defendant, a jury strictly following the *M'Naghten* algorithm will necessarily refuse to grant the insanity defense. But if it is true that "most people agree [that] incapacity [to conform to the law] excuses," the *M'Naghten* algorithm will have given the wrong result.[70] An attempt to rectify this defect produces the second moment in the dialectic, which is the *M'Naghten* test *supplemented* by an excuse for volitional incapacity.[71] *M'Naghten* as supplemented by an excuse for volitional incapacity provides the jury with a new, perhaps more adequate algorithm: if it is true that the defendant was unable to know the wrongfulness of his act or to conform his conduct to the requirements of the law, find that he is insane.

But Professor Garvey identifies a problem for the *M'Naghten* test, even as supplemented with a volitional incapacity excuse.[72] Statements about capacity are plausibly analyzed as counterfactual statements: that a person has the capacity to know right from wrong, or to conform his conduct to the law, if there exists a possible world, different from but sufficiently similar to the actual world, in which the person would have known that the act was wrong or would have conformed his conduct to the law.[73] Now the traditional *M'Naghten* test, on its terms, allows the insanity defense only to those with "total" incapacity.[74] In counterfactual terms, the traditional *M'Naghten* test does not allow the insanity defense to someone even if the only possible worlds in which that person would have known that his act was wrong or would have conformed his conduct to the law are significantly different from the actual world.[75] In particular, the cognitive incapacity element of the *M'Naghten* test will be met only if there exists a possible world in which the defendant would not have known that his act was wrong even had a police officer or other figure of authority appeared to tell him just that.[76] Likewise, the

---

70 Stephen P. Garvey, *Insanity*, *in* THE PALGRAVE HANDBOOK OF APPLIED ETHICS AND THE CRIMINAL LAW 385, 396 (L. Alexander, K. K. Ferzan eds., 2019) .

71 *See, e.g.*, State v. Hartley, 565 P.2d 658, 660 (N.M. 1977) (recognizing in addition to the two elements of *M'Naghten* an excuse if the defendant "as a result of disease of the mind 'was incapable of preventing himself from committing' the crime").

72 *See* Garvey, *supra* note 3, at 132–35.

73 *Id.* at 133; *see also* Moore, *supra* note 3, at 208 ("[I]t is plausible to analy[z]e 'X could have A-ed' in terms of the counterfactual, 'X would have A-ed if C', where 'C' represents a change from the actual world (in which X did not A).") (quoted in Garvey, *supra* note 70, at 390 n.18).

74 Garvey, *supra* note 3, at 133.

75 *Id.* (calling relevant counterfactuals "unforgiving").

76 *Id.*

volitional incapacity element of the supplemented *M'Naghten* test will be met only if there exists a possible world in which a defendant would not have conformed his conduct to the law even had he been presented with the prospect of certain and immediate death if he failed to.[77] As Professor Garvey observes, Daniel M'Naghten himself is unlikely to be judged insane by the lights of the supplemented *M'Naghten* test, because, in fact, he did know the wrongfulness of what he was doing, and was not apparently subject to any compulsion.[78] The failure of the traditional *M'Naghten* test to excuse M'Naghten himself exemplifies its more general defects: it "does a pretty poor job overall sorting the intuitively sane from the intuitively insane" and, as in the case of M'Naghten, tends "to produce false-negatives."[79]

One major American court, the Court of Appeals for the District of Columbia, responded at first to the evident inadequacy of the traditional *M'Naghten* rule by adopting a test that asked merely whether the "unlawful act was the product of mental disease or mental defect."[80] "The fundamental objection to" the traditional *M'Naghten* rule, as the court in *Durham v. United States* described it, was "not that criminal irresponsibility is made [under *M'Naghten*] to rest upon an inadequate, invalid or indeterminable symptom or manifestation, but that it is made to rest upon *any* particular symptom."[81] The *Durham* court found that "[i]n attempting to define insanity in terms of a symptom, the courts have assumed an *impossible* role."[82] The solution was to leave "the fact finder . . . free to consider all information advanced by relevant scientific disciplines."[83] These observations bespeak a pessimism about insanity theory. It is *impossible*, the *Durham* court says, that insanity should be defined in terms of a single symptom. The jury should be left with as much information as possible about the defendant but should not be given any specific instruction about what facts, if true, would be excusing. The general rule adopted in *Durham*, however, had an effect opposite to what the court had contemplated. Ultimate conclusions about whether a defendant had a mental disease, and whether an act was the product of that disease, became the purview of psychiatric ex-

77 *Id.*
78 *Id.* at 134.
79 *Id.* at 135.
80 Durham v. United States, 214 F.2d 862, 874–75 (D.C. Cir. 1954).
81 *Id.* at 872.
82 *Id.* (emphasis added).
83 *Id.*

perts, and the jury's ultimate decision-making prerogative was effectively canceled.[84]

Less than two decades after the D.C. Circuit decided *Durham*, it abandoned the product rule in *United States v. Brawner* and adopted a variation on *M'Naghten*.[85] Judge Bazelon, the author of the opinion in *Durham*, wrote in dissent. He counseled against a return to the *M'Naghten* rule, or to any other "simple, scientific formula that will [purport to] provide a clear-cut answer to every case."[86] No such formula exists, according to Judge Bazelon. Instead, courts

> have no choice . . . but to tell the truth: that the jury, not the experts, must judge the defendant's blameworthiness; that a calibrated, easily-applied standard is not yet available to guide that decision; and that the jury must resolve the question with reference to its own understanding of community concepts of blameworthiness.[87]

To that end Judge Bazelon proposed that the jury be instructed simply "that a defendant is not responsible if at the time of his unlawful conduct his mental or emotional processes or behavior controls were impaired to such an extent that he cannot justly be held responsible for his act."[88] Judge Bazelon expressly renounced the enterprise of insanity theory. "This Court's search for a new set of words to define the elusive concept of responsibility has a distinctively archaic quality."[89] But "[w]hat should by now be clear is that the problems of the responsibility defense cannot be resolved by adopting for the standard or the jury instruction any new formulation of words."[90] Judge Bazelon did not, to be sure, propose that the jury be liberated to decide insanity according to its moral sentiments, at least in those terms. But he did recognize the futility of attempting to devise an algorithm of the form, "if a response-independent fact is true of the defendant, find him insane," that would adequately define the class of insane people. He therefore appealed to the jury's "understanding of community

---

84 *See* United States v. Brawner, 471 F.2d 969, 1019 (D.C. Cir. 1972) (en banc) (Bazelon, C.J., dissenting) ("The purpose [of *Durham* and its progeny] was to give the jury an adequate basis for deciding whether the disability was such that it would be unjust to condemn the defendant for his conduct. In practice, however, under *Durham* and its progeny psychiatrists have continued to make moral and legal judgments beyond the proper scope of their professional expertise.").

85 *Id.* at 973 (majority opinion).

86 *Id.* at 1012 (Bazelon, C.J., dissenting).

87 *Id.*

88 *Id.* at 1032 (emphasis omitted).

89 *Id.* at 1039.

90 *Id.*

concepts of blameworthiness" as ultimately determinative of insanity.[91] Judge Bazelon's idea may be stated, without too much injustice, as that a person is irresponsible because the jury, acting under "community concepts of blameworthiness," finds that person irresponsible.

Judge Bazelon's renunciation of insanity theory and his appeal to the jury's sense of blameworthiness could have arrested the dialectic.[92] But his attitude to insanity found no favor and insanity commentators have remained committed to theory.[93] However, their project has not been to rehabilitate the traditional *M'Naghten* rule. Instead, these commentators are "revisionists" who argue "that what makes such people insane and what excuses them is that they or their acts are *irrational* in some way."[94] The leading irrationality theorist is Michael Moore. He accepts as given that *M'Naghten* and its variants are defective when assessed by the usual method of insanity theory (comparison of the results given by the theory in question with uncontroversial intuitions about blameworthiness).[95] Moore argues that an irrationality theory, by contrast, can answer the defects of traditional theory by "adequately captur[ing] our moral sentiments."[96] According to Moore, a

---

91 *Id.* at 1012.

92 Judge Bazelon's proposed instruction, with its emphasis on *justice*, might be understood to subordinate jurors' moral sentimental responses to normative theory. *See id.* at 1032. Understood this way, it fails to affect the conceptual reversal urged in this Note. *Cf. supra* text accompanying notes 32–38 (arguing that Wallace, in defining the class of the responsible as comprehending those it is *fair* to hold responsible, fails to articulate a response-dependent thesis). But the germ of the response-dependent thesis is clearly evident in Judge Bazelon's admonition that the "search for a new set of words to define the elusive concept of responsibility has a distinctively archaic quality," *Brawner*, 471 F.2d at 1039, and his appeal to the prerogative of the jury, *id.* at 1012.

93 For example, Professor Garvey in one instance has acknowledged the virtues of Judge Bazelon's quietist or "no-test" proposal but affirmed his commitment "to keep searching for the right test." Garvey, *supra* note 70, at 404. Professor Garvey's reasons appear to be two-fold. First is a suspicion that a "jury's good judgment," untutored by an algorithmic test, might not be very good at all. *Id.* Second, and related, is a thought that "[i]n some cases the law might make all the difference" by giving the jury a definition of insanity. *Id.* These observations, insofar as they are impediments to accepting the "no-test" proposal, are considered *infra* pp. 123–125.

94 Sinnott-Armstrong, *supra* note 2 (emphasis added).

95 MOORE, *supra* note 66, at 602 ("[O]nce one makes their meaning [viz. the meaning of the excusing terms in *M'Naghten* and its variants] more precise, it [will] turn out that some criminals we want to excuse would not be excused if the only grounds for excuse were" those allowed by *M'Naghten* and its variants.).

96 *Id.* It is remarkable that Moore should describe the project of insanity theory in this way. Indeed, as argued *infra* pp. 120–122, Moore's irrationality theory may be understood as a confused formulation of the response-dependent thesis.

person is irrational, and therefore insane, if the reasons for that person's actions defy explanation by a practical syllogism. In general, a syllogism is a deductive argument with two premises. The premises of an Aristotelian practical syllogism, on Moore's telling, specify, first, "what the agent desire[s]" and second, "the beliefs that [the agent] ha[s] as the means available to that desire's fulfillment."[97] But Moore believes that this simple syllogism contains several concealed premises. For example, a rational agent forms *intentions* to take some action A when the agent desires X and believes that A is a means of obtaining X.[98] And a rational agent *wills* a bodily movement M if the agent intends to do action A, and believes that M will cause A.[99] Moore supposes that a fully elaborated practical syllogism contains, in all, *nine* premises.[100] But how is an agent's failure to follow a valid practical syllogism relevant to that person's being insane? Moore answers by reference to the ways in which we (at least according to Moore) respond to or treat people who act under defective practical syllogisms.[101] Moore argues that "[o]nly if we can see another being" as practically rational, "will we understand her in the same fundamental way that we understand ourselves and others in everyday life."[102] "We regard as moral agents only those beings we can understand in this way."[103]

Here Moore very nearly states a thesis of response-dependent insanity. But he does not follow the thought to its natural conclusion. Instead, he insists that our ways of "see[ing]," "understand[ing]," or "regard[ing]" another person—in short, our susceptibility to the moral sentiments with respect to that person—are not *causa sui* (explanatorily basic) but in fact just the results of determinations that one or more of the nine premises in that person's practical syllogism is defective. As a matter of moral phenomenology, this is facially implausible: whether "we understand [someone] in the same fundamental way that we understand ourselves and others *in everyday life*" cannot depend on evaluation of the validity of a nine-premise syllogism.[104] The graver defect in Moore's account, however, is that it cannot be made coherent on its own terms. Instead, to be

97 MOORE, *supra* note 66, at 603.
98 *Id.* at 604.
99 *Id.* at 604–05.
100 *Id.*
101 *See id.* at 608.
102 *Id.*
103 *Id.*
104 *Id.* (emphasis added).

coherent, it must be restated as a response-dependent account. Observe to begin with that Moore cannot be understood to argue that every person who acts under a defective syllogism is insane. Some measure of irrationality is inevitable and common but has no excusing or exempting effect.[105] But, Moore may be imagined to reply, that is just *de minimis* irrationality; the truly exempting irrationality causes a person to become *unintelligible*, that is, deranged, delirious, "totally deteriorated," "hopelessly psychotic," or even catatonic.[106] It is obvious that unintelligibility so understood is unacceptable as a criterion of insanity.[107] But Moore may be imagined once more to reply: unintelligibility is not total deterioration or hopeless psychosis. It is simply an irrationality that precludes us from "understand[ing] [an agent] in the same fundamental way that we understand ourselves and others in everyday life."[108] To answer this way, however, is to define the condition of irresponsibility—unintelligibility—not by fundamental reference to any fact about the agent's rationality, but by reference to the way we respond to "others in everyday life." To make fundamental reference in defining the condition of irresponsibility to the way we respond to "others in everyday life" is to concede the correctness of the response-dependent thesis.

The foregoing rehearsal of a stylized dialectic in insanity theory suggests that insanity theorists have been led by righteous dissatisfaction with *M'Naghten* and its variants to seek alternative theories of insanity. Two of the most significant statements to that end, those by Judge Bazelon and Michael Moore, contained within them the germ of a fully satisfactory account, because both concentrated attention not on facts about the putatively insane person, but on the attitudes to that person taken by those evaluating his sanity. Judge Bazelon recognized that the "search for a new set of words to define the elusive concept of responsibility" was basically Sisyphean and proposed that the question of insanity be put to the jury's sense of blameworthiness.[109] But because Judge Bazelon

---

105 *See* Corrado, *supra* note 3, at 503 ("[A]ll of us are guilty of reasoning incorrectly from time to time; behavioral economics has shown us that this is so. But that goes absolutely no distance toward excusing ordinary behavior . . . . ").

106 Garvey, *supra* note 3, at 146 (interpreting unintelligibility standard).

107 *Id.* at 147 ("The unintelligibility theory is under-inclusive: it will too often put the seal of sanity on the certifiably insane.").

108 *See* MOORE, *supra* note 66, at 608.

109 United States v. Brawner, 471 F.2d 969, 1039 (D.C. Cir. 1972) (en banc) (Bazelon, C.J., dissenting); *see also id.* at 1012 ("[T]he jury must resolve the question with reference to its own understanding of community concepts of blameworthiness.").

could not articulate *why* the proper definition of insanity was elusive, nor why the jury was any better at deciding who deserved blame than a normative theorist, his proposal had the character of a surrender. For his part in the dialectic, Moore recognized that a person's responsibility is in some respect a function of the way we "see," "understand," or "regard" that person.[110] But Moore reintroduced perplexity by supposing that a person's rationality is the fundamental determinant of whether we "see," "understand," or "regard" a person as responsible or not.

III
THEORETICAL AND PRACTICAL CONSEQUENCES OF THE RESPONSE-DEPENDENT THESIS

With the benefit of the response-dependent innovation, it is now possible to give fuller expression to the thoughts Judge Bazelon and Moore expressed only inchoately. The proper definition of the class of responsible people is "elusive," as Judge Bazelon recognized, because that class is ultimately delimited by our moral sentimental responses.[111] The aptness of these responses can be judged by no other standard than the *causa sui* responses themselves. The jury has a special decision-making prerogative in matters of insanity that should not be relieved by a formulaic instruction, as Judge Bazelon also recognized, because the jury's moral sentimental responses are the ultimate determinants of who is insane; when they are suspended by a formulaic instruction, a "wrong" result is likely to obtain.[112] Moore's thought can be completed simply by omitting his preoccupation with rationality as the determinant of responsibility. His account began with the observation that the purpose of an insanity theory is "adequately [to] capture[ ] our moral sentiments."[113] It should have ended with the thought that nothing "captures" our moral sentiments—nothing determines under what circumstances a particular sentimental response is apt—except the moral sentimental responses themselves.

The response-dependent thesis about insanity can be articulated more exactly as the claim that a person is sane if and only if, and in virtue of the fact that, she has been a proper object of the moral sentimental responses, where propriety is

110 MOORE, *supra* note 66, at 608.
111 *Brawner*, 471 F.2d at 1039 (Bazelon, C.J., dissenting).
112 *Id.*
113 MOORE, *supra* note 66, at 602.

determined by reference to qualities of the *subject*, rather than the *object*, of the moral sentimental responses. The response-dependent thesis entails a theoretical quietism—a retreat from the project of trying to say "what makes a mind insane" or "in what insanity consists"—because it affirms that (a) a person is insane in virtue of being responded to in a certain way and (b) these responses are *causa sui*, that is, explanatorily basic. Because the responses are *causa sui*, or explanatorily basic, it cannot be said *why* they are, only observed *that* they are. To foreclose the possibility of saying why is to foreclose the possibility of theory. If this conclusion is accepted, quietism in insanity theory is the only rational posture. Now, to accept quietism in this sense is not to deny that it is possible to make more or less accurate observations about the circumstances under which the moral sentimental responses are likely to be suspended. Indeed, traditional insanity theory might be understood to be just this sort of descriptive enterprise. But to accept quietism is to hold that traditional insanity theory is *mere* description, which errs when it pretends not just to describe but to *explain* and *regulate* the moral sentimental responses, by specifying antecedent conditions of their correctness. This explanatory pretension is error because, again, the moral sentimental responses must be supposed to be *causa sui*, that is, explanatorily basic.

The response-dependent thesis about responsibility should, therefore, have a therapeutic effect in insanity theory. But insanity theory is not undertaken for its own sake. It is undertaken by judges and jurisprudents whose final concern is to say how a jury should distinguish the sane from the insane. Because insanity theory is, in the final analysis, a practical project, it is necessary to make at least some comment on how the response-dependent thesis and the quietism it has been supposed to entail might affect the administration of the insanity defense in fact.

Why not simply liberate a jury that has been appropriately solemnized to decide according to its moral sentimental responses?[114] The objection with most force is that, even if in

---

114 *Cf. Brawner*, 471 F.2d at 1032 (Bazelon, C.J., dissenting) ("Our instruction to the jury should provide that a defendant is not responsible *if at the time of his unlawful conduct his mental or emotional processes or behavior controls were impaired to such an extent that he cannot justly be held responsible for his act.*"); ROYAL COMMISSION ON CAPITAL PUNISHMENT, *supra* note 2, at 116 ("[A] preferable amendment of the law would be . . . to leave the jury to determine whether at the time of the act the accused was suffering from disease of the mind (or mental deficiency) to such a degree that he ought not to be held responsible.").

principle nothing can be said about in what insanity consists, it may be that the law, which abhors arbitrariness and subjectivity, must do better than to hope that jurors will know insanity when they see it.[115]

There is reason to think, however, that whatever the soundness of the general principle that it is the business of the law to provide rules, no such rule need govern administration of the insanity defense. Empirical study of jury treatment of the insanity defense suggests that, instruction in the standard insanity tests notwithstanding, juries *already* decide who is insane according to their responses.[116] One commentator, summarizing the empirical literature, has written that the "studies show that none of the de jure legal tests produce significantly different verdict patterns from any other de jure test *or from a de facto test* (i.e., giving jurors no test at all)."[117] Another commentator on the empirical literature has observed that "although it is possible to specify the factors that juries take into account, it is not possible to systematize those factors into a stateable rule or test."[118] He explains this impossibility by likening jury decisions about insanity to "phenomena involving pattern recognition," such as "the identification of faces . . . the construction of grammatical sentences . . . [or] the discernment of anger or fear in others," which "individuals can [competently] perform but not explain in a verbally cogent manner."[119] Insanity, he says, is something of which juries have a "perception."[120]

It may be rejoined that the fact that juries *do* decide insanity according to their perceptions (or, one might say, their moral sentimental responses) is no reason that they *should be encouraged* to decide insanity according to their perceptions (or moral sentimental responses). This rejoinder only has force, however, if it is assumed that jury perceptions or responses are in some way arbitrary, unacceptably subjective, or beneath the cognizance of the law.[121] It is beyond the scope of this Note to

---

115 *Cf.* Garvey, *supra* note 70, at 404 ("The no-test test is one solution to the problem of insanity . . . . Like any proposal to leave hard questions to a jury's good judgment, it's only as good as the jurors on whose judgment it depends.").

116 *See* Norman J. Finkel, *De Facto Departures from Insanity Instructions*, 14 L. & HUM. BEHAV. 105, 113 (1990).

117 *Id.* (emphasis added).

118 Dan M. Kahan, *Lay Perceptions of Justice vs. Criminal Law Doctrine: A False Dichotomy?*, 28 HOFSTRA L. REV. 793, 795 (2000).

119 *Id.*

120 *Id.*

121 In a critical discussion of the test proposed by the Royal Commission on Capital Punishment, Livermore and Meehl declare that "since intuitions will vary,

evaluate that assumption. But it can be observed, at least, that it is open to dispute that perceptions or moral sentimental responses are indeed arbitrary or unacceptably subjective, even if the reasons for them are inarticulable. It is evident that in ordinary life, "judgments [are not] deemed outside of reason and rationality just because they are automatic or hard to explain."[122] An adult can *feel* insulted or angry; *see* that a friend is sad, stressed, or happy; or be *aware* that someone approaching on the street is threatening, without reflection on the aptness of these feelings or perceptions. The law does recognize that jurors are competent to evaluate credibility by demeanor—in Judge Learned Hand's terms, to "take into consideration the whole nexus of sense impressions which they get from a witness"[123]—even though that "process is subjective and difficult to describe."[124] A topic for further inquiry is whether jurors, appropriately solemnized, might be liberated to decide who is insane by the same emotional and perceptive competencies they rely upon to conduct ordinary adult life and to assess witness demeanor. The response-dependent thesis is at least a suggestion that, upon scrutiny, the disfavor the law has traditionally shown to the discretion of the jury deciding insanity may be without justification.[125]

---

similar cases will be treated differently," a variation which "it has been the purpose of all legal systems to minimize rather than accentuate." Livermore & Meehl, *supra* note 2, at 825. Notice, however, that if insanity is response-dependent, no sense can be made of the notion that cases similar in the relevant respect (the insanity of the defendant) can be differently treated: different treatment, on the response-dependent account, just means that two cases are not similar in the relevant respect (whether the defendant is insane).

122 Paul Gewirtz, *On "I Know It When I See It,"* 105 YALE L.J. 1023, 1030 (1996).

123 Dyer v. MacDougall, 201 F.2d 265, 269 (2d Cir. 1952).

124 James P. Timony, *Demeanor Credibility*, 49 CATH. U. L. REV. 903, 904–05 (2000).

125 *See, e.g.*, MODEL PENAL CODE § 4.01 cmt. 3 (AM. L. INST. 1985) (rejecting the no-test proposal of the Royal Commission on Capital Punishment on the grounds that it "fails to focus the attention of the trier of fact on the specific manifestations and effects of mental disease or defect that are relevant to the justice of conviction and punishment"); Livermore & Meehl, *supra* note 2, at 824–25 (criticizing as unduly "intuitive[ ]" the no-test test proposed by the Royal Commission on Capital Punishment); Garvey, *supra* note 70, at 404 (expressing concern that a no-test test is "only as good as the jurors on whose judgment it depends"); Donald H.J. Hermann & Yvonne S. Sor, *Convicting or Confining? Alternative Directions in Insanity Law Reform: Guilty But Mentally Ill Versus New Rules for Release of Insanity Acquittees*, 1983 B.Y.U. L. REV. 499, 525 (1983) (observing that the no-test proposal has been criticized for purportedly allowing "jurors . . . to use unreviewable personal criteria for assessing blame").

CONCLUSION

The criminal law has benefitted and will continue to benefit from concepts and arguments developed in academic philosophy. This Note has proposed that the theory of insanity, which remains in an unsatisfactory confusion even after more than six decades of concentrated thought by judges and legal scholars, would benefit from taking heed of a relatively novel thesis in moral philosophy: that a person is responsible in virtue of being *held* responsible. So too might a person be insane in virtue of her being treated as insane. A stylized dialectic suggested that the response-dependent thesis has been recognized, in an inchoate way, by two of the leading revisionists in insanity theory, Judge Bazelon and Michael Moore. This Note proposes that the dialectic of insanity theory can now be arrested by making explicit that a person is sane in virtue of the fact that she has been an object of properly operating moral sentimental responses. Because these responses are explanatorily basic, insanity theory should adopt a quietist posture. A topic for further study is whether, assuming that the response-dependent thesis about insanity is correct, the jury could be liberated to decide who is insane according to its moral sentimental responses without offense to basic principles of the criminal law.