CHAPTER TWO

Moral Grammar and Intuitive Jurisprudence: A Formal Model of Unconscious Moral and Legal Knowledge

John Mikhail

Contents

Abstract

Could a computer be programmed to make moral judgments about cases of intentional harm and unreasonable risk that match those judgments people already make intuitively? If the human moral sense is an unconscious computational mechanism of some sort, as many cognitive scientists have suggested, then the answer should be yes. So too if the search for reflective equilibrium is a sound enterprise, since achieving this state of affairs requires demarcating a

Psychology of Learning and Motivation, Volume 50
ISSN 0079-7421, DOI: 10.1016/S0079-7421(08)00402-7

set of considered judgments, stating them as explanandum sentences, and formulating a set of algorithms from which they can be derived. The same is true for theories that emphasize the role of emotions or heuristics in moral cognition, since they ultimately depend on intuitive appraisals of the stimulus that accomplish essentially the same tasks. Drawing on deontic logic, action theory, moral philosophy, and the common law of crime and tort, particularly Terry's five-variable calculus of risk, I outline a formal model of moral grammar and intuitive jurisprudence along the foregoing lines, which defines the abstract properties of the relevant mapping and demonstrates their descriptive adequacy with respect to a range of common moral intuitions, which experimental studies have suggested may be universal or nearly so. Framing effects, protected values, and implications for the neuroscience of moral intuition are also discussed.

A critic who wished to say something against that work [*Groundwork of the Metaphysic of Morals*] really did better than he intended when he said that there was no new principle of morality in it but only a new formula. Who would want to introduce a new principle of morality and, as it were, be its inventor, as if the world had hitherto been ignorant of what duty is or had been thoroughly wrong about it? Those who know what a formula means to a mathematician, in determining what is to be done in solving a problem without letting him go astray, will not regard a formula which will do this for all duties as something insignificant and unnecessary.

Immanuel Kant, *Critique of Practical Reason*

[I]n our science, everything depends upon the possession of the leading principles, and it is this possession which constitutes the greatness of the Roman jurists. The notions and axioms of their science do not appear to have been arbitrarily produced; these are actual beings, whose existence and genealogy have become known to them by long and intimate acquaintance. For this reason their whole mode of proceeding has a certainty which is found no where else, except in mathematics; and it may be said, without exaggeration, that they calculate with their notions.

F.C. Von Savigny, *Of the Vocation of Our Time for Legislation and Jurisprudence*

How does it happen that the prevailing public opinion about what is right and what is moral is in so many respects correct? If such a philosopher as Kant failed in the attempt to find the source of our knowledge of right and wrong, is it conceivable that ordinary people succeeded in drawing from this source?... But this difficulty... is easily resolved. We only have to reflect that much of what is present in our store of knowledge contributes toward the attainment of new knowledge without our being clearly conscious of the process.... Thus it has often been observed that for thousands of years men have drawn right conclusions without bringing the procedure and the principles which form the condition of the formal validity of the

> inference into clear consciousness by means of reflection. . .. In spite of their false conception of the true fundamental principles, these still continue to operate in their reasoning. But why do I go so far for examples? Let the experiment be made with the first ''plain man'' who has just drawn a right conclusion, and demand of him that he give you the premises of his conclusion. This he will usually be unable to do and may perhaps make entirely false statements about it.
>
> Franz Brentano, *The Origin of The Knowledge of Right and Wrong*

> The demand is not to be denied: every jump must be barred from our deductions. That this is so hard to satisfy must be set down to the tediousness of proceeding step by step.
>
> Gottlob Frege, *The Foundations of Arithmetic*

1. The Moral Grammar Hypothesis

The moral grammar hypothesis holds that ordinary individuals are intuitive lawyers, who possess tacit or unconscious knowledge of a rich variety of legal rules, concepts, and principles, along with a natural readiness to compute mental representations of human acts and omissions in legally cognizable terms (Mikhail, 2000, 2005, 2007, 2008a; see also Dwyer, 1999, 2006; Harman, 2000, 2008; Hauser, 2006; Mahlmann, 1999, 2007; Roedder and Harman, 2008; see generally Miller, 2008; Pinker, 2008; Saxe, 2005). The central aim of this chapter is to provide a preliminary formal description of some of the key mental operations implied by this hypothesis. In a comprehensive study, each of these operations would need to be described in a format suitable for explicit derivations, and many details, complications, and objections would need to be addressed. In what follows, I will be content merely to sketch some of the main ideas in quasi-formal terms, leaving further refinements, extensions, and clarifications for another occasion. My primary objective is to demonstrate that a computational theory of moral cognition, which explains an interesting and illuminating range of common moral intuitions, can indeed be formulated.

Because some readers might find the efforts at formalization in this chapter to be tedious or unnecessary, it seems useful to address this issue at the outset. Cognitive science was transformed by subjecting linguistic and visual phenomena to precise, formal analysis. The theory of moral grammar holds out the prospect of doing the same for aspects of ordinary human moral cognition, perhaps thereby lending support to the Enlightenment assumption that at least some aspects of intuitive moral judgment are ''capable of demonstration'' (Locke, 1991/1689, p. 549; cf. Hume, 1978/1740; Kant, 1993/1788;

Leibniz, 1981/1705). The alleged computational properties of moral cognition, however, must be shown and not merely asserted.

As Rawls (1971, p. 46) observes, the first step in this inquiry is to identify a class of considered judgments and a set of rules or principles from which they can be derived. As I have argued elsewhere, recent sustained efforts to explain human moral judgment in this framework suggest that untutored adults and even young children are intuitive lawyers, who are capable of drawing intelligent distinctions between superficially similar cases, although their basis for doing so is often obscure (Mikhail, 2007, 2008a; see also Alter et al., 2007; Cushman et al., 2006; Haidt, 2001; Robinson et al., 2008; Solum, 2006; Wellman and Miller, 2008; Young and Saxe, 2008; cf. Anscombe, 1958; Bradley, 1876; Cardozo, 1921; Durkheim, 1893; Freud, 1930; Gilligan, 1978; Gluckman, 1955, 1965; Holmes, 1870; Jung, 1919; Kohlberg, 1981, 1984; Piaget, 1932; Pound, 1908). If this is correct, then future research in moral psychology should begin from this premise, moving beyond pedagogically useful examples such as the trolley problem and other cases of necessity to the core concepts of universal fields like torts, contracts, criminal law, property, agency, equity, procedure, and unjust enrichment, which investigate the rules and representations implicit in common moral intuitions with unparalleled care and sophistication. Chomsky (1957) emphasized that rigorous formulation in linguistics is not merely a pointless technical exercise but rather an important diagnostic and heuristic tool, because only by pushing a precise but inadequate formulation to an unacceptable conclusion can we gain a better understanding of the relevant data and of the inadequacy of our existing attempts to explain them. Likewise, Marr (1982, p. 26) warned against making inferences about cognitive systems from neurophysiological findings without "a clear idea about what information needs to be represented and what processes need to be implemented" (cf. Mill, 1987/1843, pp. 36–38). Cognitive scientists who take these ideas seriously and who seek to understand human moral cognition must devote more effort to developing computational theories of moral competence, in addition to studying related problems, such as its underlying mechanisms, neurological signatures, cultural adaptations, or evolutionary origins. As I attempt to show in this chapter, the formalization of common legal notions can play an important part in this process.

Because the enterprise this chapter engages, the search for considered judgments in reflective equilibrium (Rawls, 1971), is controversial in some quarters, a further clarification may be helpful before we proceed. Moral judgment is a flexible, context-dependent process, which cannot be accurately described by simple consequentialist or deontological principles, and which is clearly subject to framing effects and other familiar manipulations (Doris, 2002; Kahneman and Tversky, 1984; Kelman et al., 1996; Schnall et al., 2008; Sunstein, 2005; Unger, 1996; Valdesolo and DeSteno, 2006; Wheatley and Haidt, 2005). For example, as the literature on protected

values has shown, how trade-offs among scarce resources are described can often influence how they are evaluated (Baron and Spranca, 1997; Bartels, 2008; Bartels and Medin, 2007; Fiske and Tetlock, 1997; Tetlock, 2003). Facts like these are sometimes taken to imply that moral intuitions are so malleable that the project of reflective equilibrium is quixotic. From our perspective, however, these phenomena simply reinforce the need to draw a competence–performance distinction in the moral domain and thus to take a position, fallible and revisable to be sure, on which moral judgments reflect the ideal operations of a core human competence and which are the result of various psychological limitations, performance errors, or other exogenous factors (Nozick, 1968; Rawls, 1971; cf. Chomsky, 1965; Macnamara, 1986; Marr, 1982). Hence the importance of jury instructions, rules of evidence, and other familiar methods of directing attention to precisely formulated questions and preventing irrelevant or prejudicial information from having a distorting effect on one's judgments. Unlike some researchers (e.g., Baron and Ritov, this volume), who define any deviation from utilitarianism as a cognitive "bias" — and who thus appear committed to holding that even the most basic rules of criminal and civil law reflect pervasive cognitive errors, insofar as they do not merely track outcomes, but also rely heavily on concepts like proximate causes, goals, means, side effects, and mental states generally — the approach taken here assumes that at least some of these rules are a natural benchmark with which to describe human moral cognition, at least to a good first approximation. Whether these legal norms are built into the very fabric of the human mind is one of cognitive science's deepest and most persistent questions. Our immediate concern, however, is not ontogenesis but descriptive adequacy, because without a clear understanding of the learning target in this domain, one cannot formulate, let alone endorse, one or another learning theory. Despite their obvious limitations, trolley problems are a useful heuristic for this purpose, and their artificiality is a virtue, not a vice, in this regard. These hypothetical cases must be supplemented with more realistic probes drawn from other branches of law, policy, and everyday life, however, if moral competence is to be adequately understood.

2. The Problem of Descriptive Adequacy

2.1. Twelve Considered Judgments

The provisional aim of moral theory is to solve the problem of descriptive adequacy (Rawls, 1971, 1975; cf. Chomsky, 1957, 1965). To simplify this problem, it is useful to begin by focusing our attention on the 12 problems in Table 1, which, building upon previous work (Foot, 1967; Harman, 1977; Thomson, 1985), I designed in order to investigate the mental

Table 1 Twelve Trolley Problems.[a]

1. **Bystander:** Hank is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Hank sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Hank is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the men. There is a man standing on the side track with his back turned. Hank can throw the switch, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Hank to throw the switch?
2. **Footbridge:** Ian is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Ian sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Ian is standing next to a *heavy object*, which he can throw *onto the track in the path of the train,* thereby preventing it from killing the men. *The heavy object* is a man, standing *next to Ian* with his back turned. Ian can throw the *man,* killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Ian to throw the *man*?
3. **Expensive Equipment:** Karl is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Karl sees what has happened: the driver of the train saw *five million dollars of new railroad equipment lying* across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the *equipment*. It is moving so fast that *the equipment* will be *destroyed*. Karl is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from *destroying the equipment*. There is a man standing on the side track with his back turned. Karl can throw the switch, killing him; or he can refrain from doing this, letting *the equipment* be destroyed. Is it morally permissible for Karl to throw the switch?
4. **Implied Consent:** Luke is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Luke sees what has happened: the driver of the train saw *a man* walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward

the *man*. It is moving so fast that *he* will not be able to get off the track in time. Luke is standing next to *the man*, *whom* he can throw *off the track out of the path of the train*, thereby preventing it from killing the *man*. *The man* is *frail and* standing with his back turned. Luke can throw the man, *injuring* him; or he can refrain from doing this, letting the *man* die. Is it morally permissible for Luke to throw the man?

5. **Intentional Homicide:** Mark is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Mark sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed, and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Mark is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the men. There is a man on the side track. Mark can throw the switch, killing him; or he can refrain from doing this, letting the men die. *Mark then recognizes that the man on the side track is someone who he hates with a passion. ''I don't give a damn about saving those five men,'' Mark thinks to himself, ''but this is my chance to kill that bastard.''* Is it morally permissible for Mark to throw the switch?

6. **Loop Track:** Ned is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Ned sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Ned is standing next to a switch, which he can throw, that will temporarily turn the train onto a side track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, giving the men time to escape. The heavy object is a man, standing on the side track with his back turned. Ned can throw the switch, preventing the train from killing the men, but killing the man. Or he can refrain from doing this, letting the five die. Is it morally permissible for Ned to throw the switch?

7. **Man-In-Front:** Oscar is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Oscar sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the

track in time. Oscar is standing next to a switch, which he can throw, that will temporarily turn the train onto a side track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, giving the men time to escape. *There* is a man standing on the side track *in front of the heavy object* with his back turned. Oscar can throw the switch, preventing the train from killing the men, but killing the man; or he can refrain from doing this, letting the five die. Is it morally permissible for Oscar to throw the switch?

8. **Costless Rescue:** Paul is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Paul sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Paul is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the men. Paul can throw the switch, saving the five men; or he can refrain from doing this, letting the five die. Is it morally *obligatory* for Paul to throw the switch?

9. **Better Alternative:** Richard is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Richard sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed, and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Richard is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the men. There is a man on the side track with his back turned. Richard can throw the switch, killing him; or he can refrain from doing this, letting the men die. *By pulling an emergency cord, Richard can also redirect the train to a third track, where no one is at risk. If Richard pulls the cord, no one will be killed. If Richard throws the switch, one person will be killed. If Richard does nothing, five people will be killed.* Is it morally permissible for Richard to throw the switch?

10. **Disproportional Death:** Steve is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Steve sees what has happened: the driver of the train saw *a man* walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the man. It is moving so fast that he will not be able to get off the track in time. Steve is standing next to a switch,

which he can throw, that will turn the train onto a side track, thereby preventing it from killing the man. There are *five men* standing on the side track with their backs turned. Steve can throw the switch, killing the *five men*; or he can refrain from doing this, letting the *one man* die. Is it morally permissible for Steve to throw the switch?

11. **Drop Man:** Victor is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Victor sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Victor is standing next to a switch, which he can throw, that will drop a heavy object into the path of the train, thereby preventing it from killing the men. The heavy object is a man, who is standing on a footbridge overlooking the tracks. Victor can throw the switch, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Victor to throw the switch?

12. **Collapse Bridge:** Walter is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Walter sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Walter is standing next to a switch, which he can throw, that will *collapse a footbridge overlooking the tracks* into the path of the train, thereby preventing it from killing the men. *There* is a man standing on a footbridge. Walter can throw the switch, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Walter to throw the switch?

[a] Italics in Table 1 identify salient differences between the following minimal pairs: Bystander-Footbridge, Bystander-Expensive Equipment, Footbridge-Implied Consent, Bystander-Intentional Homicide, Loop Track-Man-In-Front, Bystander-Costless Rescue, Bystander-Better Alternative, Bystander-Disproportional Death, Drop Man-Collapse Bridge. Experimental subjects were not shown these markings.

computations underlying the ordinary exercise of moral judgment (Mikhail, 2000, 2002).

In a series of experiments that began in the mid-1990s, my colleagues and I began testing these cases, and others like them based on the same basic template, on hundreds of individuals, both adults and children. The participants included several groups of American adults, several groups of

American children, one group of recent Chinese immigrants to the United States, and two groups of master's students at Harvard University's Kennedy School of Government. Collectively, the participants hailed from a diverse set of countries, including Belgium, Canada, China, Columbia, Denmark, Egypt, Finland, France, Germany, India, Iran, Israel, Italy, Japan, Korea, Lebanon, Mexico, Puerto Rico, and South Africa. Our central aim was to pursue the idea of a universal moral grammar and to begin to investigate a variety of empirical questions that arise within this framework. Our basic prediction was that the moral intuitions elicited by the first two problems (Bystander and Footbridge) would be widely shared, irrespective of demographic variables such as race, sex, age, religion, national origin, or level of formal education (see generally Mikhail, 2000, 2007; Mikhail et al., 1998). We also predicted that most individuals would be unaware of the operative principles generating their moral intuitions, and thus would be largely incapable of correctly describing their own thought processes (Mikhail et al., 1998). These predictions were confirmed, and our initial findings have now been replicated and extended with over 200,000 individuals from over 120 countries (see, e.g., Hauser et al., 2007; Miller, 2008; Pinker, 2008; Saxe, 2005). The result is perhaps the first qualitatively new data set in the history of the discipline, which has transformed the science of moral psychology and opened up many new and promising avenues of investigation (see, e.g., Bartels, 2008; Bucciarelli et al., 2008; Cushman, 2008; Cushman et al., 2006; Dupoux and Jacob, 2007; Greene et al., submitted; Koenigs et al., 2007; Lombrozo, 2008; Machery, 2007; Moore et al., 2008; Nichols and Mallon, 2006; Sinnott-Armstrong et al., 2008; Waldmann and Dieterich, 2007; Young et al., 2007).[1]

The modal responses to these 12 cases are listed in Table 2. While the variance in these intuitions is an important topic, which I discuss elsewhere (Mikhail, 2002, 2007; cf. Section 3.1), in this chapter I focus on the modal responses themselves and make the simplifying assumption that these judgments are considered judgments in Rawls' sense, that is, "judgments in which our moral capacities are most likely to be displayed without distortion" (1971, p. 47). Hence, I take them to be categorical data that a descriptively adequate moral grammar must explain.

[1] When our trolley problem studies began in Liz Spelke's MIT lab in the mid-1990s, Petrinovich and colleagues (Petrinovich and O'Neill, 1996; Petrinovich et al., 1993) had already begun using trolley problems as probes, which another lab member (Laurie Santos) brought to our attention only several years after the fact. From our perspective, the Petrinovich experiments were poorly conceived, however, because they asked participants to supply behavioral predictions ("What would you do?") rather than clearly identified moral judgments ("Is X morally permissible?"). In the context of jury trials, the former instruction has long been held to be reversible error (see, e.g., Eldredge, 1941; Epstein, 2004), while the latter more closely approximates the key theoretical issue of reasonableness or justifiability under the circumstances.

Table 2 Twelve Considered Judgments.

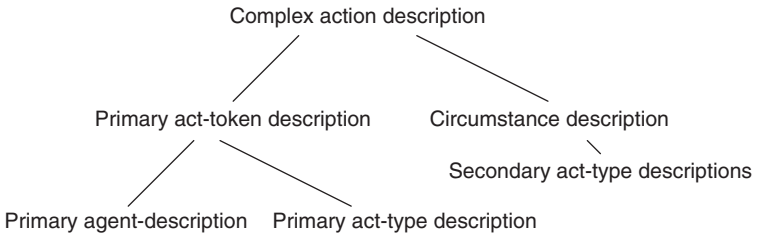
Problem	Act	Deontic status
Bystander	Hank's throwing the switch	Permissible
Footbridge	Ian's throwing the man	Forbidden
Expensive Equipment	Karl's throwing the switch	Forbidden
Implied Consent	Luke's throwing the man	Permissible
Intentional Homicide	Mark's throwing the switch	Forbidden
Loop Track	Ned's throwing the switch	Forbidden
Man-In-Front	Oscar's throwing the switch	Permissible
Costless Rescue	Paul's throwing the switch	Obligatory
Better Alternative	Richard's throwing the switch	Forbidden
Disproportional Death	Steve's throwing the switch	Forbidden
Drop Man	Victor's throwing the switch	Forbidden
Collapse Bridge	Walter's throwing the switch	Permissible

2.2. The Poverty of the Perceptual Stimulus

For convenience, let us label each of these cases a *complex action-description*. Let us say that their two main constituents are a *primary act-token description* and a *circumstance description*. The primary act-token description consists of a *primary act-type description* and a *primary agent-description*. The circumstance description also includes *secondary act-type descriptions*. Hence, our scheme for classifying the input may be rendered by Figure 1A, and the results of applying it to an example like the Bystander problem can be given by Figure 1B. Clearly, it is unproblematic to classify the remaining cases in Table 1 in these terms.

With this terminology, we may now make a simple but crucial observation about the data in Table 2. Although each of these rapid, intuitive, and highly automatic moral judgments is occasioned by an identifiable stimulus, how the brain goes about interpreting these complex action descriptions and assigning a deontic status to each of them is not something revealed in any obvious way by the surface structure of the stimulus itself. Instead, an intervening step must be postulated: an intuitive appraisal of some sort that is imposed on the stimulus prior to any deontic response to it. Hence, a simple perceptual model, such as the one implicit in Haidt's (2001) influential model of moral judgment, appears inadequate for explaining these intuitions, a point that can be illustrated by calling attention to the unanalyzed link between eliciting situation and intuitive response in Haidt's model (Figure 2A; cf. Mikhail, 2007, 2008b). Likewise, an *ad hoc* appraisal theory, such as the personal/impersonal distinction that underlies Greene's (Greene, 2005; Greene and Haidt, 2002; Greene et al., 2001) initial explanation of

A



B

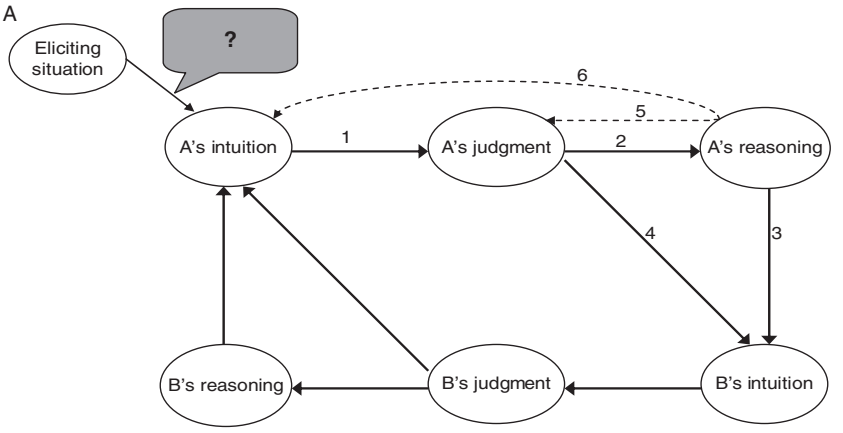
Primary act-token description	Hank to throw the switch
Circumstance description	Hank is taking his daily walk over the train tracks when he notices that the train that is approaching is out of control. Hank sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Hank is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the men. There is a man standing on the side track with his back turned. Hank can throw the switch, killing him; or he can refrain from doing this, letting the five die
Primary agent description	Hank
Primary act-type description	throw the switch
Secondary act-type descriptions	1. will turn the train onto a side track 2. preventing it from killing the men 3. killing him 4. refrain from doing this 5. letting the five die

Figure 1 **Classifying the Stimulus: (A) Scheme and (B) Application.**

the trolley problems, also fails to explain the data (Figure 2B; cf. Mikhail, 2002, 2007, 2008b; see also Greene, 2008a,b for recognition of this problem). Instead, an adequate model must be more complex and must look more like Figure 3.

Figure 3 implies that moral judgments do not depend merely on the superficial properties of an action-description, but also on how that action is *mentally represented*, a critical preliminary step in the evaluative process that jurists have frequently examined (e.g., Cardozo, 1921; Hutcheson, 1929; Oliphant, 1928; Radin, 1925; see also Grey, 1983; Kelman, 1981), but, surprisingly, many psychologists have unduly neglected. The point can be illustrated by Table 3, which supplies an exhaustive list of the primary and secondary act-type descriptions that are directly derivable from the stimuli in Table 1. As Table 3 reveals, it is not just difficult, but *impossible*, to explain the data in Table 2 by relying on these primary and secondary act-type descriptions alone. Strictly speaking, the impossibility covers only the

A



B

Problem	Personal / Impersonal	Deontic status
Bystander	Impersonal	Permissible
Footbridge	Personal	Forbidden
Expensive Equipment	Impersonal	Forbidden
Implied Consent	Personal	Permissible
Intentional Homicide	Impersonal	Forbidden
Loop Track	Impersonal	Forbidden
Man-In-Front	Impersonal	Permissible
Costless Rescue	Impersonal	Obligatory
Better Alternative	Impersonal	Forbidden
Disproportional Death	Impersonal	Forbidden
Drop Man	Impersonal	Forbidden
Collapse Bridge	Impersonal	Permissible

"A moral violation is personal if it is (i) likely to cause serious bodily harm, (ii) to a particular person, (iii) in such a way that the harm does not result from the deflection of an existing threat onto a different party. A moral violation is impersonal if it fails to meet these criteria." (Greene & Haidt 2002: 519)

Figure 2 **Two Inadequate Appraisal Theories: (A) Unanalyzed Link in Haidt's (2001) Model of Moral Judgment and (B) Inadequacy of Greene's (Greene and Haidt, 2002; Greene et al., 2001) Personal–Impersonal Distinction.**

Bystander, Intentional Homicide, Loop Track, and Man-In-Front problems, since these are the only cases whose primary and secondary act-type descriptions are completely equivalent. It is therefore logically possible to formulate *ad hoc* hypotheses that could handle the remaining eight cases. For example, each case could be explained by an elaborate conditional whose antecedent simply restates the primary and secondary act-types contained in the stimulus. Presumably, with enough effort, even such an unimaginative theory as this could be falsified, but, in any case, the point I am making should be apparent. Clearly, the brain must be generating action representations of its own that go beyond the information given. That is, much like a given patch of retinal stimulation or the acoustic stream in

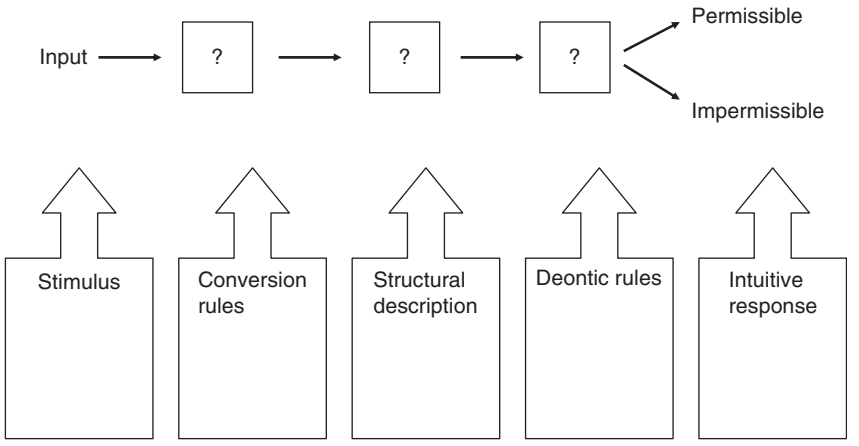


Figure 3 Expanded Perceptual Model for Moral Judgment.

speech perception, the stimulus here evidently consists merely of clues for the formation of an unconscious percept that the perceiver first constructs using her own internal resources and then projects back onto the stimulus, creating an illusion of qualities that the latter does not in fact possess (cf. Descartes, 1985/1647, p. 303; Hume, 1983/1751, p. 88; see also Chomsky, 2000; Fodor, 1985; Helmholtz, 1962/1867; Marr, 1982; Rey, 2006). Hence an adequate scientific explanation of the data in Table 2 must specify at least three elements: (i) the deontic rules operative in the exercise of moral judgment, (ii) the structural descriptions over which those computational operations are defined, and (iii) the conversion rules by which the stimulus is transformed into an appropriate structural description (Figure 3).

2.3. Simplifying the Problem

Let us break this problem into parts and attempt to treat each part systematically. Because we seek to explicate 12 distinct judgments, we must construct 12 separate derivations. To make this task more manageable, we rely on the following idealizations and simplifying assumptions. First, we assume certain basic principles of deontic logic (Figure 4). We also assume that the sole deontic primitive in our model is the concept *forbidden*, leaving the concepts *permissible* and *obligatory*, and the various logical expressions in Figure 4, to be defined by implication.[2]

[2] To generate the expressions in Figure 4, we need just two logical connectives, because out of "∼" (*not*) and any one of "." (*and*), "v" (*or*), "⊃" (*if–then*), or "≡" (*if and only if*), the others may be mechanically defined. For example, given two propositions, P and Q, and the connectives "∼" and "v," we may define "(P . Q)" as an abbreviation for "(∼((∼P) v ((∼Q)))"; "(P ⊃ Q)" as an abbreviation for "((∼P) v Q)"; and "(P ≡ Q)" as an abbreviation for "(P ⊃ Q) . (Q ⊃ P)."

Table 3 The Poverty of the Perceptual Stimulus.

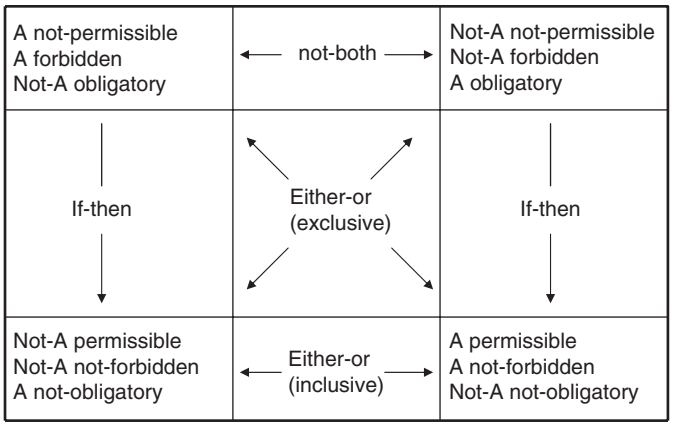
Problem	Primary act-type description	Secondary act-type descriptions	Deontic status
Bystander	throw the switch	**1.** will turn the train onto a side track **2.** preventing it from killing the men **3.** killing him **4.** refrain from doing this **5.** letting the five die	Permissible
Footbridge	throw the man	**1.** throw onto the track into the path of the train **2.** preventing it from killing the men **3.** killing him **4.** refrain from doing this **5.** letting the five die	Forbidden
Expensive Equipment	throw the switch	**1.** will turn the train onto a side track **2.** preventing it from killing the man **3.** killing them **4.** refrain from doing this **5.** letting the man die	Forbidden
Implied Consent	throw the man	**1.** throw off the track out of the path of the train **2.** preventing it from killing the man **3.** injuring him	Permissible

(*Continued*)

Table 3 (*Continued*)

Problem	Primary act-type description	Secondary act-type descriptions	Deontic status
		4. refrain from doing this **5.** letting the man die	
Intentional Homicide	throw the switch	**1.** will turn the train onto a side track **2.** preventing it from killing the man **3.** killing him **4.** refrain from doing this **5.** letting the man die	Forbidden
Loop Track	throw the switch	**1.** will turn the train onto a side track **2.** preventing it from killing the men **3.** killing him **4.** refrain from doing this **5.** letting the five die	Forbidden
Man-In-Front	throw the switch	**1.** will turn the train onto a side track **2.** preventing it from killing the men **3.** killing him **4.** refrain from doing this **5.** letting the five die	Permissible
Costless Rescue	throw the switch	**1.** will turn the train onto a side track **2.** preventing it from killing the men **3.** saving the five men **4.** refrain from doing this **5.** letting the five die	Obligatory

Better Alternative	throw the switch	**1.** will turn the train onto a side track **2.** preventing it from killing the men **3.** killing him **4.** refrain from doing this **5.** letting the man die **6.** pulling an emergency cord **7.** redirect the train to a third track	Forbidden
Disproportional Death	throw the switch	**1.** will turn the train onto a side track **2.** preventing it from killing the man **3.** killing the five men **4.** refrain from doing this **5.** letting the one man die	Forbidden
Drop Man	throw the switch	**1.** will drop a heavy object into the path of the train **2.** preventing it from killing the men **3.** killing him **4.** refrain from doing this **5.** letting the five die	Forbidden
Collapse Bridge	throw the switch	**1.** will collapse a footbridge overlooking the tracks into the path of the train **2.** preventing it from killing the men **3.** killing him **4.** refrain from doing this **5.** letting the five die	Permissible



Key: Equipollence relations (i.e. logical equivalences) are expressed in the four corners. "A" stands for *act*; "not-A" stands for *omission*.

Figure 4 Principles of Deontic Logic: Square of Opposition and Equipollence.

Second, we assume that the form of our derivations is given by the following schema:

(1) A has deontic status D ≡ A has features $F_1 \ldots F_n$
A has features $F_1 \ldots F_n$
A has deontic status D

In other words, we attempt to state necessary and sufficient conditions for assigning a deontic status to a given act or omission. As noted below (Section 4.1), this renders our model a logically closed system and, given our choice of primitive, it is equivalent to assuming that the correct closure rule is a Residual Permission Principle.

Third, we replace the letter "A" in (1) with the following formula:

(2) [S's V-ing at $t^{(\alpha)}$]^{C}

The syntax of this formula calls for comment. Drawing upon Goldman (1970) and Ginet (1990), we take the central element of what we call the normal form of a *complex act-token representation* to be a gerundive nominal, whose grammatical subject is possessive (cf. Bentham's preference for nominalization in Ogden, 1932). Following Katz (1972), we use the symbol "at t" to denote some unspecified position on an assumed time dimension, and we use superscripts on occurrences of "t" to refer to specific positions on this dimension. We assume that superscripts can be either variables or constants. We take "t" with the superscript constant "0," i.e., "$t^{(0)}$," to function as an indexical element in a complex act-token representation, serving to orient the temporal relationships holding between it and other such representations.

Superscript variables (''*n*,'' ''*m*,'' etc.) denote members of the set of natural numbers. They appear in superscripts with prefixes ''+'' and ''−,'' indicating the number of positive or negative units from the origin point (''$t^{(0)}$'') of the time dimension. For example, ''$t^{(+n)}$'' means ''*n* units to the right of the origin,'' whereas ''$t^{(-n)}$'' signifies ''*n* units to the left of the origin.'' Thus, ''$t^{(-n)}$,'' ''$t^{(0)}$,'' and ''$t^{(+m)}$'' in the following series of representations imply that Hank's seeing what happened occurs before his throwing the switch, which occurs before his killing the man:

(3) (a) [Hank's seeing what happened at $t^{(-n)}$]
(b) [Hank's throwing the switch at $t^{(0)}$]
(c) [Hank's killing the man at $t^{(+m)}$]

There is an important convention this notation incorporates, which is to date an action by its time of *completion*. Strictly speaking, an act that begins at $t^{(0)}$ and ends at $t^{(+n)}$ is performed neither at $t^{(0)}$ nor $t^{(+n)}$, but in that period of time bounded by them. We simplify this situation by following the traditional legal rule of dating an action by when it is completed (see, e.g., Salmond, 1966/1902, p. 360). Doing so enables us to avoid many problems, such as locating ''the time of a killing,'' which have been identified in the literature (Thomson, 1970; cf. Fodor, 1970; Fried, 1978; Jackendoff, 1987; Pinker, 2007). Finally, since acts always occur in particular circumstances, we need a notation for designating those circumstances. Hence, we enclose these representations in square brackets, followed by the superscript ''C'' to denote the circumstances in which act-tokens are performed.[3]

3. Intuitive Legal Appraisal

3.1. Acts and Circumstances

It is at this point that turning more directly to legal theory and the philosophy of action is useful for our topic. Together with aspects of Goldman's (1970) theory of level-generation, the substantive law of crime and tort provides us with the necessary conceptual tools for explaining the data in Table 2, as well as an indefinitely large class of structurally similar judgments.

[3] Our notation for designating act-token representations can be elaborated in simple ways, as needed. For example, we can exhibit more complex temporal relations by relying on conventions for adding and subtracting in algebra. Thus, ''$t^{(+n\ +\ (-m))}$'' signifies ''$n - m$ units to the right of the origin,'' while ''$t^{(-n\ +\ (-m)\ +\ (-o))}$'' signifies ''$n + m + o$ units to the left of the origin.'' Likewise, our generic reference to circumstances, ''C,'' can be replaced with one or more sets of circumstances, ''{C1, C2, C3,. . .,C*n*}'' (see generally Mikhail, 2000).

From a common legal point of view, an *act* is simply a voluntary bodily movement that occurs in a particular set of circumstances (see, e.g., Holmes, 1881; Terry, 1884; cf. ALI, 1965; Goldman, 1970). Those circumstances, in turn, may be regarded as a body of information that obtains at the time that the act or its omission occurs. In *De inventione*, Cicero supplies a classic list of seven probative questions that can be asked about the circumstances of any particular action:

Quis? Quid? Ubi? Quibus auxiliis? Cur? Quomodo? Quando?
Who? What? Where? By what aids? Why? How? When?

Cicero's list, which is presumably illustrative rather than exhaustive, has been the subject of philosophical analysis for centuries (see, e.g., Aquinas, 1952/1274, p. 653). For our purposes, its significance rests in the fact that the answers elicited by questions like these can transform one description of an action into another, and that the resulting set of descriptions can be arranged into hierarchical tree structures, successive nodes of which bear a generation relation to one another that is asymmetric, irreflexive, and transitive (Goldman, 1970; see also Anscombe, 1957; Davidson, 1963; Donagan, 1977; Ryle, 1968; cf. Geertz, 1973 on "thick description"). When properly constructed, these expository diagrams not only enable us to predict moral intuitions with surprising accuracy, but also to see at a glance a variety of structural relationships, including those we might have overlooked or ignored.

For example, act trees can be used not only to identify the basic differences between the Footbridge and Bystander problems, but also to explain the *variance* one finds in highly refined manipulations of these cases, such as the Loop Track, Man-In-Front, Drop Man, and Collapse Bridge problems. As Figure 5A indicates, the intuitive data in these six cases form a remarkably consistent pattern, with permissibility judgments increasing linearly across the six conditions. Moreover, as Figure 5b illustrates, these results can be tentatively explained as a function of the properties of each problem's structural description. Other things equal, acts are more likely to be judged permissible as counts of battery committed as a means decrease from three (Footbridge) to two (Drop Man) to one (Loop Track), and as these violations become side effects and additional structural features come into play. In Man-In-Front, the agent's goal presumably is to save the men by causing the train to hit the object but not the man, yet the actual result (not shown) is likely to involve hitting the man before the object; hence, from an *ex post* perspective, the agent will commit a battery prior to and as a means of achieving his good end. Likewise, in Collapse Bridge, one or more counts of battery must necessarily occur before the good end is achieved. By contrast, in Bystander, battery and homicide are side effects that occur only after the good end has been secured by turning the train onto the side track (Mikhail, 2007).

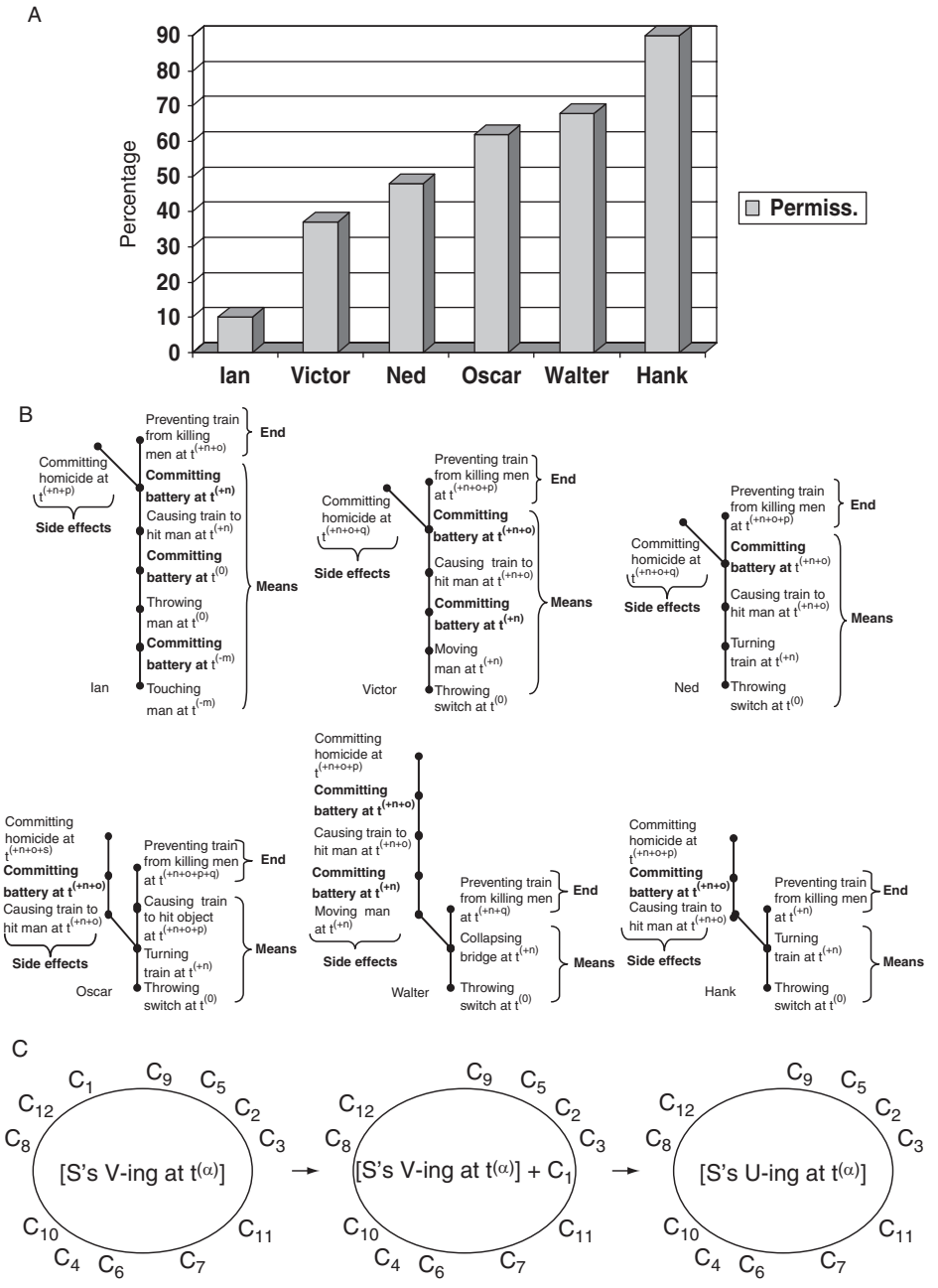


Figure 5 Circumstances Alter Cases: (A) Variance in Six Trolley Problem Experiments, (B) Six Structural Descriptions, and (C) "Mental Chemistry" as an Illustration of Moral Computation. (Data in (a) from Mikhail, 2002, 2007, in press).

Diagramming action plans is an important tool for solving the problem of descriptive adequacy, and it was a major part of the effort that began with the earliest work on moral grammar to identify the precise structural properties of the mental representations that are elicited by thought experiments such as the trolley problems (Mikhail et al., 1998; cf. Bentham, 1948/1789, p. 79; Locke, 1991/1689, pp. 550–552). In what follows, I seek to build on this foundation by explaining how iterated applications of a general computational principle, which combines an act-token representation with its circumstances to yield another act-token representation, can be utilized together with a variety of legal rules, concepts, and principles to explain how the brain computes the complex structural descriptions of a given action and its alternatives. All of these structural descriptions can be exhibited by act trees, but, as I suggest below (Section 5), other graphic devices, such as a table of recurring elements, can also be fruitfully utilized in this endeavor.

Formally, this general principle can be rendered in various ways, including the following:

(4) (a) $[\text{S's V-ing at } t^{(\alpha)}]^{C} \rightarrow [\text{S's U-ing at } t^{(\beta)}]$
(b) $[\text{S's V-ing at } t^{(\alpha)}] + C \rightarrow [\text{S's U-ing at } t^{(\beta)}]$
(c) $[\text{S's V-ing at } t^{(\alpha)} + C] \rightarrow [\text{S's U-ing at } t^{(\beta)}]$
(d) $[\text{S's V-ing at } t^{(\alpha)}]^{\{C1,\ C2,\ C3 \ldots Cn\}} \rightarrow [\text{S's V-ing at } t^{(\alpha)} + C_1]^{\{C2,\ C3 \ldots Cn\}}$
(e) $[\text{S's V-ing at } t^{(\alpha)} + C_1]^{\{C2,\ C3 \ldots Cn\}} \rightarrow [\text{S's U-ing at } t^{(\beta)}]^{\{C2,\ C3 \ldots Cn\}}$

(4a) states that a complex act-token representation can yield another act-token representation. In this formula, "→" functions as a rewrite rule that permits the object on the left side of the arrow to be replaced by the object on the right side. (4b) uses the "+" symbol to express a similar proposition, indicating that a circumstance can be added to an act-token representation to yield another act-token representation. (4c) is similar, but more precise, because it signifies that a circumstance becomes material, so to speak, by combining with an act-token representation within its corresponding brackets. Finally, (4d) and (4e) reveal how, in two distinct steps, the process might unfold in a generic case. First, a particular circumstance, C_1, is selected from the set of circumstances surrounding an act-token representation, S's V-ing at $t^{(\alpha)}$, and conjoined with the latter (4d). Next, the combination of these two elements yields a new act-token representation, S's U-ing at $t^{(\beta)}$ (4e). The set of circumstances surrounding this transformation remains intact throughout, except that C_1 is no longer an element of this set.

As Bentham (1948/1879, p. 77) observes, the basic mental processes we seek to describe can be illustrated by drawing on the etymology of the word *circumstance* — "*circum stantia*, things standing round: objects standing around another object" — and thus conceiving of "the field of circumstances,

belonging to any act'' to be ''a circle, of which the circumference is no where, but of which the act in question is the centre.'' Moreover, as Mill (1987/1843, pp. 39–40) observes, the relevant phenomena can be conceived as a kind of ''mental chemistry,'' in which simple ideas combine to generate more complex ones in a process loosely analogous to chemical combination (Figure 5C; cf. Kant, 1993/1788, pp. 169–171; D'Arcy, 1963, pp. 57–61). The particular diagrams used to exhibit these transformations are inessential, of course, and one should avoid getting carried away with metaphors that risk obscuring rather than illuminating the relevant mental operations. What matters is simply to recognize that any adequate scientific theory of moral intuition must seek to explain how the brain converts complex action-descriptions and other sensory inputs into complex act-token representations as a necessary precondition of moral judgment. The general computational principle we have identified is a plausible component of one such proposal, but, even so, it obviously cannot do the job on its own. As I argue below, however, this principle, together with other elements of moral grammar, can be used to explain the 12 cases in Table 1, along with a potentially infinite number and variety of other cases.

3.2. K-Generation and I-Generation

Modifying Goldman's (1970) analysis to suit our topic, let us begin by defining two generation relations that might hold between pairs of act-token representations:

Definition of K-Generation
Given two act-token representations, [S's V-ing at $t^{(\alpha)}$] and [S's U-ing at $t^{(\beta)}$], and a set of known circumstances, C, [S's V-ing at $t^{(\alpha)}]^{C}$ *K-generates* [S's U-ing at $t^{(\beta)}$] if and only if:

(a) $V \neq U$
(that is: [S's V-ing] and [S's U-ing] are syntactically distinct)
(b) $\beta - \alpha \geq 0$
(that is: [S's U-ing at $t^{(\beta)}$] is either time-identical or subsequent to [S's V-ing at $t^{(\alpha)}$])
(c) ([S's V-ing at $t^{(\alpha)}$ + C] → [S's U-ing at $t^{(\beta)}$]
(that is: the conjunction of [S's V-ing at $t^{(\alpha)}$] and C yields [S's U-ing at $t^{(\beta)}$])

Definition of I-Generation
Given two act-token representations, [S's V-ing at $t^{(\alpha)}$] and [S's U-ing at $t^{(\beta)}$], and a set of known circumstances, C, [S's V-ing at $t^{(\alpha)}]^{C}$ *I-generates* [S's U-ing at $t^{(\beta)}$] if and only if:

(a) [S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's U-ing at t(β)]
(b) ([S's U-ing at t(β)]=[GOAL] v [S's U-ing at t(β)] *I-generates* [GOAL])
(that is: [S's U-ing at $t^{(\beta)}$] is the goal, or *I-generates* the goal, of an action plan)

Comment: These provisional definitions of K-generation and I-generation are meant to provide a sufficient basis for our limited objectives of accounting for the fact that individuals ordinarily distinguish at least two types of effects that are caused by a moral agent: (i) effects that are *knowingly* caused (K-generation), and (ii) effects that are *intentionally* or *purposely* caused (I-generation). For simplicity, we assume here that the latter are a proper subset of the former; hence, we do not attempt to account for cases in which an agent intends to accomplish ends that she believes are unlikely to occur. Instead, we simply assume that all of the effects that are intentionally or purposefully caused by an agent are also knowingly caused by her.

In Anglo-American jurisprudence, ''intent'' is often defined or used broadly to include knowledge. For example, the American Law Institute's Restatement (Second) of Torts uses ''intent,'' ''intentional,'' and related terms ''to denote that the actor desires to cause consequences of his act, or that he believes that the consequences are substantially certain to result from it'' (ALI, 1965, p. 15; cf. Lefave and Scott, 1972, pp. 196–197). Likewise, Sidgwick holds that ''[f]or purposes of exact moral or juristic discussion, it is best to include under the term of 'intention' all the consequences of an act that are foreseen as certain or probable'' (1981/1907, p. 202). Ordinary language often exhibits a different and more precise understanding of intention, however, and distinguishes intended and foreseen effects in a variety of contexts (Bratman, 1987; Finnis, 1995; Kenny, 1995), including the trolley problems, if our hypothesis is correct. By defining K-generation and I-generation in the foregoing manner, then, we depart from certain conventional accounts of intention and attempt instead to explicate the familiar distinction between ends and means, on the one hand, and known or foreseen side effects, on the other.

Bentham (1948/1789, p. 84) succinctly captures the distinction between I-generation and K-generation when he explains that a consequence may be ''directly or lineally'' intentional, when ''the prospect of producing it constituted one of the links in the chain of causes by which the person was determined to do the act,'' or merely ''obliquely or collaterally'' intentional, when ''the consequence was in contemplation, and appeared likely to ensue in case of the act's being performed, yet the prospect of producing such a consequence did not constitute a link in the aforesaid chain.'' The definition of I-generation tracks Bentham's notion of direct intention; it can also be regarded as a rule for generating the adverb ''purposely'' (or ''intentionally'' in one of its ambiguous meanings) and conjoining it to an act-token representation that otherwise lacks this mental

state. The definition of K-generation corresponds with Bentham's notion of collateral intention; it can also be regarded as a rule for generating the adverb "knowingly" and conjoining it to an act-token representation that otherwise lacks this mental state.

The recursive aspect of the definition of I-generation (i.e., provision (b)) is meant to provide a computational interpretation of the principle Kant takes to be analytic: A rational agent who wills the end necessarily wills the known means (Kant, 1964/1785, pp. 84–85). The key insight here is that once the end, goal, or final effect of a causal chain has been identified, each of the previous links of that chain can be sequentially transformed from a representation of a mere *cause* of its subsequent effects to a representation of a *means* of its subsequent ends. In this manner, we can explain how the brain imputes intentional structure to what previously was only a projection of causes and effects. The end, goal, or final effect of an action is presupposed in this process. Later, we explain how one can compute the end, goal, or final effect of a complex act-token representation on the basis of information about its good and bad effects (see Section 6.4).

For present purposes, we do not define a separate notion of C-generation (i.e., causal generation; see Goldman, 1970) or distinguish it from K-generation. Nor do we incorporate an explicit causal requirement in our definition of K-generation. Doing so would complicate our model, and it seems unnecessary given our immediate aims. Instead, we simply assume that each agent in Table 1 both causes and knows the stipulated effects of his actions. The reference to *known* circumstances in the definition of K-generation is thus taken to mean that these circumstances, including relevant causal conditionals, are known to the agents themselves (as well as to the participants in our experiments, in a secondary sense of the term). In a fully adequate moral grammar, these assumptions would need to be scrutinized, and a separate notion of C-generation would presumably need to be analyzed, defined, and incorporated into our definition of K-generation to account for the fact that individuals ordinarily distinguish both the effects that are objectively caused by an agent and those that she knowingly caused. We leave this task for another occasion, with the expectation that by drawing on a sophisticated body of work on causation (e.g., Alicke, 1992; Hart and Honore, 1959; Mackie, 1974; Pearl, 2000; Wright, 1985), a computational theory of C-generation can be integrated into the foregoing framework.

4. Deontic Rules

The two notions we have defined, K-generation and I-generation, provide a principled basis for distinguishing what an agent knowingly does from what she purposely does, at least in a provisional way suitable for our

limited aims. We need more conceptual tools, however, to explain the data in Table 2. An adequate moral grammar must include several more concepts and principles.

4.1. The Principle of Natural Liberty

One of these principles is a so-called *closure rule* (Raz, 1970; Stone, 1968; see also Rawls, 1971), which renders our idealized model closed or complete. From a logical point of view, there are two main possibilities: (i) a *Residual Prohibition Principle*, which assumes that all permissible acts and omissions are defined and states that "whatever is not legally permitted is prohibited," and (ii) a *Residual Permission Principle*, which assumes that all forbidden acts and omissions are defined and states that "whatever is not legally prohibited is permitted."[4] The first alternative, which appears in Aristotle's discussion of law,[5] is essentially authoritarian, since it leaves little or no room for individual choice. The second alternative, which underwrites the legal maxims *nullum crimen sine lege* ("no crime without law") and *nullu peona sine lege* ("no penalty without law") and characterizes modern liberalism, is essentially libertarian, since it implies unrestricted freedom within the domain of acts that are neither obligatory nor forbidden.

The Residual Permission Principle may also be called a *Principle of Natural Liberty*, and it is this essentially libertarian principle, rather than the essentially authoritarian Residual Prohibition Principle — or, alternatively, the apparently unrestrained notion of natural liberty that one finds in legal writers like Hobbes, Blackstone, and Bentham — on which the system we describe here rests. In particular, we follow a long line of writers on natural jurisprudence (e.g., Burlamaqui, 2006/1748, p. 284; Kant, 1991/1797, pp. 63–64; Wilson, 1967/1790, pp. 587–588; cf. Mill, 1978/1859, pp. 9–10) in adopting a more restricted, yet still expansive, precept of natural liberty as our preferred closure rule, which can be rendered as follows: "If an act has features $F_1 \ldots F_n$, then it is forbidden; otherwise, it is permissible." More formally, the principle can be restated as the following conjunction of conditionals, which is simply a theorem of our model that can be derived from the schema in (1) together with the equipollence relations in Figure 4:

Principle of Natural Liberty
$[[\text{S's V-ing at } t^{(\alpha)}]^{C} \text{ has features } F_1 \ldots F_n] \supset [[\text{S's V-ing at } t^{(\alpha)}]^{C} \text{ is forbidden}]$.
$\sim[[\text{S's V-ing at } t^{(\alpha)}]^{C} \text{ has features } F_1 \ldots F_n] \supset [[\text{S's V-ing at } t^{(\alpha)}]^{C} \text{ is permissible}]$

[4] A *Residual Obligation Principle* is not a genuine third alternative, because it is logically equivalent to the Residual Prohibition Principle (cf. Figure 4).

[5] See Aristotle, *Nichomachean Ethics*, 1138a, 6–8 (observing that "the law does not expressly permit suicide, and what it does not expressly permit it forbids").

4.2. The Prohibition of Battery and Homicide

Any normative system purporting to achieve descriptive adequacy must presumably include a set of basic legal prohibitions. For our purposes, two familiar prohibitions are relevant: battery and homicide. In a moral grammar that is capable of serving as premises of a derivation, each of these trespasses would need to be clearly and comprehensively defined. Here, I will merely state provisional definitions that are suitable for our limited purposes.

First homicide: The American Law Institute's Model Penal Code defines homicide in part as an act which consists in "purposely, knowingly, recklessly, or negligently causing the death of another human being" (ALI, 1962, Section 210.1). Modifying this definition to suit our purposes by detaching its adverbial component, let us assume that the act-type *commits homicide* can be defined[6] simply as causing the death of a person. Formally, this can be stated as follows:

Definition of Homicide:
[S commits homicide at $t^{(\alpha)}$]^{C} $=_{Df}$ [S's V-ing [EFFECT (Person, Death)] at $t^{(\alpha)}$]^{C}

There is an implicit causation element in this definition that requires further analysis, but we set aside this issue here. By combining this definition with the notions of I-generation and K-generation, we can now distinguish the following two types of homicide.

Representation of Purposeful Homicide
[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's committing homicide at $t^{(\beta)}$]

Representation of Knowing Homicide
[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's committing homicide at $t^{(\beta)}$]

In our model, the first expression formalizes the complex act-type *purposely committing homicide*. The second formalizes the complex act-type *knowingly committing homicide*. As we shall see, the second formula appears operative in ten of our 12 cases. By contrast, the only case that appears to involve the first formula is the Intentional Homicide problem.

Next battery: Prosser (1941, p. 43) defines battery in part as "unpermitted, unprivileged contact with [a] person." The Restatement (Second) of Torts (ALI, 1965, p. 25) offers a more elaborate definition, which reads in part: "An actor is subject to liability for battery if (a) he acts intending to cause a harmful or offensive contact with the person of the other or a third person, or an imminent apprehension of such contact, and (b) a harmful contact with the person of the other directly or indirectly results." Modifying these accounts to suit our objectives, let us assume that the act-type

[6] On the standard form of definition used here, see generally Hempel (1955).

commits battery can be defined simply as causing harmful contact with a person without her consent.[7] Formally, this definition can be stated as follows:

Definition of Battery
[S commits battery at $t^{(\alpha)}]^{C}$ $=_{Df}$ [S's V-ing [EFFECT (Person, Contact-$_{H}$, $\sim$Consent)] at $t^{(\alpha)}]^{C}$

The concept of *contact* as it is used in this definition needs to be explained. In the common law of torts, protection against unwanted physical contact encompasses all forms of direct touching and "extends to any part of the body, or to anything which is attached to it and practically identified with it" (Prosser, 1971, p. 34). Moreover, it includes any "touching of the person, either by the defendant or any substance put in motion by him" (Hilliard, 1859, p. 191). Hence, the ordinary concept of contact is inadequate in some circumstances and must be replaced with a more expansive concept. Although we need not draw the precise contours of this broader concept here, it is important to recognize that a salient contact occurs not only when a person is (i) touched or (ii) moved by an agent, but also when she is (iii) touched by an object that is being touched by an agent, (iv) touched by an object that was previously moved by an agent, without the intervention of a more proximate cause, or (v) moved by an object that was previously moved by an agent, without the intervention of a more proximate cause. None of these effects necessarily trigger a representation of battery, but each is sufficient to generate a representation of the contact necessary for battery, at least within the confines of our model.

For example, the contact requirement can be met by shoving or grabbing another person, but also by kicking the umbrella she is holding, snatching a plate from her hand, throwing a rock at her, spitting on her, or pulling a chair out from under her as she sits down, thereby causing her to fall (see, e.g., Epstein, 2004). In our 12 cases, the requirement is satisfied by throwing a person (as in the Footbridge and Implied Consent problems), moving a person and thereby causing him to come into contact with a train (as in the Drop Man and Collapse Bridge problems), or redirecting a train so that it comes into contact with a person (as in the Bystander, Expensive Equipment, Intentional Homicide, Loop Track, Man-In-Front, Better Alternative, and Disproportional Death problems). Depending on how the implicit causation element of battery is interpreted, the requirement might also be satisfied in the Costless Rescue problem. I ignore this issue here, along with the broader question of whether battery can occur by omission, which some commentators have denied, even

[7] I focus here on harmful battery rather than offensive battery, since only the former is relevant for our purposes. On the latter, see generally the Restatement (Second) of Torts, Sections 18–20.

when the resulting harm or offense is intentional (see, e.g., the Restatement (First) of Torts, Sections 2, 13, 18, 281, and 284, and Topic 1, Scope Note).

Our definition of battery also requires that the contact be *harmful*. Hence this concept must also be analyzed, and sufficient conditions for generating it must be provided. Once again, for our purposes it is sufficient to adopt with only minor changes the concept of harm utilized by the Restatement (Second) of Torts, which provides a useful framework in this regard. First, we use the word ''harm'' and its cognates, without further qualification, to denote any kind of detriment to a person resulting from any cause (ALI, 1965, p. 12). That is, we interpret harm broadly to include any ''detriment or loss to a person which occurs by virtue of, or as a result of, some alteration or change in his person, or in physical things'' (ALI, 1965, p. 13). Second, we use the narrower notion of *bodily* harm to refer to any physical impairment of a person's body, including physical pain, illness, or alteration of the body's normal structure or function to any extent. Third, we understand the harmful contact element of battery to require bodily harm, in the sense defined. Finally, we stipulate that a harmful contact occurs whenever contact with a person *results* in bodily harm, whether or not it does so directly, immediately, or purposely. In other words, we assume that the harmful effect of an I-generated contact need not be I-generated itself for the I-generated contact to be considered harmful (cf. Bentham, 1948/1789, p. 83).

Although these analyses could be improved, they are sufficient for our limited aims. By combining our definition of battery with the notions of I-generation and K-generation, we can now formally distinguish the following two types of battery:

Representation of Purposeful Battery
[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's committing battery at $t^{(\beta)}$]

Representation of Knowing Battery
[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's committing battery at $t^{(\beta)}$]

In our model, the first expression formalizes the complex act-type *purposely committing battery*. The second formalizes the complex act-type *knowingly committing battery*. The second formula appears operative in ten of our 12 cases. By contrast, the first formula appears operative in only four cases, all of which are judged to be impermissible: Footbridge, Intentional Homicide, Loop Track, and Drop Man.

4.3. The Self-Preservation Principle

The concept of *consent* in our definition of battery, which usually operates instead as an affirmative defense (see, e.g., RST, Sections 49–62), also calls for comment. Crucial as this concept is, I do not attempt to analyze it here,

beyond stating one sufficient condition for its application. What is important for our purposes is to have a principled basis for distinguishing Luke's throwing the man in the Implied Consent problem from Ian's performing the same action in the Footbridge problem (along with numerous other cases of simple battery, in which harmful contact occurs without any possible justification). Intuitively, the relevant difference is that the man *would* consent to being thrown in the Implied Consent problem, since his own life is being saved. To generate this representation, we may assume that the moral grammar includes the following principle:

Self-Preservation Principle
[EFFECT (Person, $\text{Contact}_{\text{-H}}$)] ⊃ [EFFECT (Person, Death)] → [EFFECT (Person, $\text{Contact}_{\text{-H}}$, ∼Consent)]

Roughly, the Self-Preservation Principle affords a presumption that, if a harmful contact with a person necessitates killing her, then she would not consent to it. This presumption may, of course, be rebutted in certain contexts, such as triage, euthanasia, or physician-assisted suicide, but I set aside these potential complications here.

4.4. The Moral Calculus of Risk

If our hypothesis is correct, then the "background information" (Lashley, 1951) that must be attributed to the participants in our experiments to explain their considered judgments must include not only principles of deontic logic (Section 2.3), a general computational principle capable of transforming one act-token representation into another (Section 3.1), a set of rules for distinguishing K-generation and I-generation (Section 3.2), a closure rule (Section 4.1), and a set of presumptively prohibited acts (Section 4.2). Among other things, it also must include a moral calculus of some sort for specifying, ranking, and comparing the probabilities of an action's good and bad effects.

In our simple model, we account for the first of these three necessary operations by postulating three primary bad effects: (i) death of a person, (ii) bodily harm to a person, and (iii) destruction of a valuable thing. Formally, these three postulates can be rendered as follows:

Postulate #1:
[EFFECT [(Person, Death)] → [BAD EFFECT]

Postulate #2:
[EFFECT [(Person, $\text{Harm}_{\text{-B}}$)] → [BAD EFFECT]

Postulate #3:
[EFFECT [($\text{Thing}_{\text{-V}}$, Destroy)] → [BAD EFFECT]

The first postulate states that an effect that consists of the death of a person is a bad effect, and may be rewritten as such. In this formula, "→" is

a rewrite rule that converts the object on the left side of the arrow to the object on the right side. The second and third postulates apply the same rule to bodily harm to a person and the destruction of a valuable thing, respectively.

We also make the simplifying assumption that the only good effects in our model are those that consist of the negation of a bad effect. That is, we postulate that each bad effect has a corresponding good effect: namely, the prevention of that bad effect. In addition, we postulate a second, derivative-type of bad effect that consists of the prevention of a good effect. Formally, these two postulates can be rendered as follows:

Postulate #4:
[EFFECT [neg [BAD EFFECT]]] → [GOOD EFFECT]

Postulate #5:
[EFFECT [neg [GOOD EFFECT]]] → [BAD EFFECT]

Postulate #4 states that an effect that consists of the negation of a bad effect is a good effect, and may be rewritten as such. Postulate #5 states that an effect that consists of the negation of a good effect is a bad effect, and may be rewritten as such. In Section 6, I provide an alternative formal interpretation of these principles and explain how they can be applied directly to the underlying semantic structures of certain causative constructions in the stimulus, thereby showing how these structures can be transformed into richer representations that encode both good and bad effects.

The second operation we must explain is how to generate a moral ranking of an action's good and bad effects. In our model, we postulate a simple ordinal ranking of bad effects, according to which (i) the death of a person is morally worse than bodily harm to a person, and (ii) bodily harm to a person is morally worse than the destruction of a valuable thing. Formally, these two postulates can be rendered as follows:

Postulate #6:
[EFFECT [(Person, Death)] $<_m$ [EFFECT [(Person, $Harm_{-B}$)]

Postulate #7:
[EFFECT [(Person, $Harm_{-B}$)] $<_m$ [EFFECT [($Thing_{-V}$, Destroy)]

In these formulas, "$<_m$" symbolizes what we call the *morally worse-than* relation. Postulate #6 states that an effect that consists of the death of a person is morally worse than an effect that consists of bodily harm to a person. Postulate #7 states that an effect that consists of bodily harm to a person is morally worse than an effect that consists of destruction of a valuable thing.

In our model, the morally worse-than relation is assumed to be asymmetric, irreflexive, and transitive. If "A," "B," and "C" are three effects, then the following can be validly inferred: if A is morally worse than B, then

B is not morally worse than A (asymmetry); A is not morally worse than A (irreflexivity); if A is morally worse than B, and B is morally worse than C, then A is morally worse than C (transitivity). We also assume that each bad effect is morally worse than its corresponding good effect. Hence, we assume that (i) the death of a person is morally worse than its prevention, (ii) bodily harm to a person is morally worse than its prevention, and (iii) destruction of a valuable thing is morally worse than its prevention. Formally, these three postulates can be rendered as follows:

Postulate #8:
[EFFECT [(Person, Death)] $<_{m}$ [EFFECT [neg (Person, Death)]]

Postulate #9:
[EFFECT [(Person, Harm_{B})] $<_{m}$ [EFFECT [neg (Person, Harm_{B})]]

Postulate #10:
[EFFECT [(Thing_{V}, Destroy)] $<_{m}$ [EFFECT [neg (Thing_{V}, Destroy)]]

Finally, we postulate that the life of one person has the same moral worth as that of another. We also assume that these values can be aggregated to arrive at an ordinal (although not necessarily cardinal) ranking of multiple effects, each of which consists of the death of one or more persons. Letting "x" and "y" stand for positive integers, and letting "$>$" and "$\leq$" stand for the *is-greater-than* and *is-less-than-or-equal-to* relations (which, unlike the morally worse-than relation, are mathematical concepts, not normative ones), these assumptions imply two further postulates:

Postulate #11:
$\forall(x, y)$ [[$x > y$] $\equiv$ [(x Persons, Death)] $<_{m}$ [(y Persons, Death)]]

Postulate #12:
$\forall(x, y)$ [[$x \leq y$] $\equiv \sim$[(x Persons, Death)] $<_{m}$ [(y Persons, Death)]]

Similar formulas could presumably be constructed for the two other bad effects in our model: bodily harm to a person and the destruction of a valuable thing. However, certain complications would have to be addressed in each case. For example, even if one assumes that the physical security of one person has the same moral worth as that of another, it does not follow that bodily harm to five persons is morally worse than bodily harm to one person; to reach this conclusion, both the type and the extent of the harm must be held constant. For different types of harm, at least, a separate ranking is necessary, and problems of incommensurability can arise (Fiske and Tetlock, 1997; Hallborg, 1997; Tetlock, 2003). Likewise, it is conceivable, but not obvious, that valuable things can be monetized or otherwise ranked to permit judgments of their comparative moral worth. Nor is it clear that the destruction of a more expensive object is always morally worse than that of a less expensive

one. A comprehensive moral grammar would need to confront issues like these, but since this is not necessary for our purposes, we can set them aside.

The third operation we must explain is how to compute and compare the probabilities of an action's good and bad effects. In our model, we draw upon the common law of torts to sketch a provisional account of how this operation is performed. On this account, the reasonableness and hence justifiability of a given risk of unintentional harm can be calculated as a function of five variables: (i) the magnitude of the risk, R_M; (ii) the value of the principal object, V_P, which may be thought of as the life, safety, or property interests of the individual in question; (iii) the utility of the risk, R_U; (iv) the necessity of the risk, R_N; and (v) the value of the collateral object, V_C, which is the actor's own purpose in imposing the given risk (Terry, 1915; cf. Restatement (Second) of Torts, Sections 291–293). In particular, justifiability depends on whether R_M multiplied by V_P is greater than (i.e., morally worse than) the combined product of R_U, V_C, and R_N:

Moral Calculus of Risk

$(R_M)\ (V_P) > (R_U)\ (V_C)\ (R_N)$

The Moral Calculus of Risk is similar to the famous Hand Formula for calculating negligence liability, according to which negligence depends on whether the probability that a given accident will occur, P, multiplied by the injury or loss resulting from the accident, L, is greater than the cost or burden of preventing the accident, B; that is, on whether $PL > B$ (see, e.g., Epstein, 2004). Whereas the Hand Formula is comprised of three variables, however, the Moral Calculus of Risk relies upon five. Three of these variables are probabilities, while two of them are evaluative components that measure the comparative moral worth of the principal and collateral objects. We have already explained how the two evaluative variables can be specified and compared by means of a simple ordinal ranking of the various good and bad effects in our model. It will be useful, however, to say a further word about the three probability variables.

The *magnitude of the risk* is the probability that the principal object will be harmed in some manner; in our case, this is simply the probability that an agent will K-generate one of our three bad effects: death of a person, bodily harm to a person, or destruction of a valuable thing. The *utility of the risk* is the probability that the collateral object — the agent's purpose — will be achieved; in our model, this usually refers to the probability of K-generating a good effect (e.g., preventing the train from killing the men). The sole exception is the Intentional Homicide problem, where the agent's purpose is to achieve a bad effect. The *necessity of the risk* is the probability that the agent's purpose would *not* be achieved without risk to the principal object; in our model, this variable typically measures the likelihood that a good effect (e.g., preventing the train from killing the men) could not be achieved without K-generating a bad side effect. The sole exception is the Better

Alternative problem, where risking the bad side effect is unnecessary due to the availability of a safer alternative: turning the train onto the empty third track.

The complement to the necessity of the risk is the *gratuitousness of the risk*: the probability that the agent's purpose *would* be achieved without the risk to the principal object, or, in other words, that the risk to the principal object is useless or unnecessary. A completely gratuitous risk is one in which the necessity of the risk is 0 and the gratuitousness of the risk is 1; conversely, a completely necessary risk is one in which the gratuitousness of the risk is 0 and the necessity of the risk is 1. More generally, the gratuitousness of the risk, R_G, can be given by the formula, $1 - R_N$. Likewise, the necessity of the risk can be given by the formula, $1 - R_G$.

By substituting $(1 - R_G)$ for R_N in the Moral Calculus of Risk and by performing some simple algebra, a marginal version of the same formula can be stated as follows:

Marginal Calculus of Risk

$(R_M)(V_P) > (V_C)(R_U) - (V_C)(R_U)(R_G)$

What the Marginal Calculus of Risk makes transparent, which both the Hand Formula and, to a lesser extent, the Moral Calculus of Risk tend to obscure, is that a narrow calculation of the expected benefit of the agent's conduct, the value of the collateral object multiplied by the probability of success, is not the correct measure against which to compare the expected cost to the potential victim. Rather, what matters is the expected benefit of the necessary *risk*, that is, the difference between the expected benefit of the agent's conduct *with* the risk and the expected benefit of the agent's conduct *without* the risk. What matters, in other words, is how much the unavoidable risk of harm to the potential victim increased the likelihood that the actor's goal would be achieved. (The actor does not get credit, as it were, for the avoidable risk.) To make this calculation, one must first discount the expected benefit of the agent's conduct by its gratuitous risk, and then subtract the resulting value from the expected benefit. For ease of reference, in what follows I will refer to the value of the agent's expected benefit when it is discounted by its gratuitous risk, which can be given by either "$(R_U)(V_C)(R_N)$" or "$(V_C)(R_U) - (V_C)(R_U)(R_G)$," as the agent's *discounted expected benefit*. I will refer to the agent's expected benefit without the risk, which is given by "$(V_C)(R_U)$," as the agent's *simple expected benefit*.

With this terminology, we can now clarify an important aspect of the simple model of moral grammar outlined in this chapter and prior publications (e.g., Mikhail, 2007), which is that it generally assumes that the magnitude, utility, and necessity of the risk are to be given a value of 1, rather than some other, more realistic value. That is, our model assumes that when ordinary individuals evaluate the trolley problems, they accept the

stipulation that certain actions ''will'' have certain effects without discounting those effects by their intuitive probability.

Clearly this assumption is unrealistic. There is good reason to think that people might be discounting the stipulated outcomes by their relative likelihood. For example, they might assign a relatively low utility of the risk to throwing the man in the Footbridge problem, but a relatively high utility of the risk to throwing the switch in the Bystander problem. If this is correct, then the perceived wrongfulness of the former could be the result of two independent yet interacting factors, using battery as a means and doing something whose expected cost exceeds its discounted expected benefit, neither of which is operative in the Bystander problem. Indeed, it seems entirely possible to explain the Footbridge problem data on cost-benefit grounds alone. Holding all other factors constant, one need only postulate that people intuitively recognize that the utility of the risk of throwing the man is less than 0.2, or, put differently, that there is a less than 1 in 5 chance that this action will manage to prevent the train from killing the men. In that case, expected costs would exceed discounted expected benefits, and the conduct would be unjustifiable on that basis alone. By contrast, the intuitive mechanics of the Bystander problem are different: there is no apparent basis for doubting that the utility of the risk of turning the train to the side track is 1 (or nearly so). Nor is there any reason to doubt that the necessity of the risk is also 1 (or nearly so), as long as the stipulation that this situation is unavoidably harmful, with no latent possibility of preventing harm to all parties involved, is deemed to be credible. Hence, the discounted expected benefit in this case is equivalent to simple expected benefit, which itself equals the value of the five lives that are saved.

The same logic can be applied to other familiar thought experiments. In the Transplant problem, for instance, in which five patients are dying of organ failure, but a doctor can save all five if she removes the organs from a sixth patient and gives them to the other five (Foot, 1967), the utility of the risk might not be 1, but something much less than 1. Transplant surgery, after all, is a complicated business. Things can go wrong. It is also *expensive*. So, unlike the Trolley or Bystander problems with which it is usually compared, the discounted expected benefit of this arduous and expensive set of operations might be considerably less than it first appears. It is conceivable, although perhaps unlikely, that individuals perceive the expected costs of these operations to exceed their discounted expected benefits, and make their judgments accordingly.

Could all of the familiar trolley problem data be explained in terms of a sophisticated cost-benefit analysis, and are the complex structural descriptions we have proposed therefore unnecessary? Several factors weigh against this possibility. First, simple linguistic experiments, such as the ''by'' and ''in order to'' tests, support the hypothesis that people spontaneously compute structural descriptions of these problems that incorporate properties like ends, means, side effects, and *prima facie* wrongs, such as battery (Mikhail, 2005, 2007).

Moreover, this hypothesis is corroborated by the finding that even young children distinguish genuine moral violations, such as battery, from violations of social conventions (Smetana, 1983; Turiel, 1983; cf. Nichols, 2002) and that even infants are predisposed to interpret the acts of moral agents in terms of their goals and intentions (Gergely and Csibra, 2003; Hamlin et al., 2007; Johnson, 2000; Meltzoff, 1995; Woodward et al., 2001). It is also reinforced by a variety of recent studies at the interface of moral cognition and theory of mind (e.g., Knobe, 2005; Sinnott-Armstrong et al., 2008; Wellman and Miller, 2008; Young and Saxe, 2008). So there appears to be substantial independent evidence supporting this aspect of the moral grammar hypothesis.

Second, although we have identified a potentially important confound in the Footbridge and Bystander problems, one should not assume that it operates in all of the cases in Table 1. For example, while one might be tempted to attribute the data in Figure 5A to different utilities of the risk — it is easier to stop an onrushing train with a heavy object, such as a brick wall (Man-In-Front), than with a man (Loop Track), after all, just as one is more likely to do so with a bridge (Collapse Bridge) than with a man (Drop Man) — not all of the variance in these six cases can be explained in this manner. Whatever its value, the utility of the risk of moving a man in the path of a train, for instance, is presumably equivalent in the Footbridge and Drop Man problems. Hence the variance in these cases must be due to some other factor, which repeated applications of the battery prohibition can explain (see Figure 5B).

Third, the structural descriptions we have proposed for the Transplant, Footbridge, Loop Track, and Drop Man problems share a single, crucial property: in each case, an agent's good end cannot be achieved without committing battery as a means to this objective (Mikhail, 2007). It seems both implausible and unparsimonious to deny that this property enters into the relevant computations, particularly since it presumably operates in countless instances of ordinary battery, that is, run-of-the-mill cases which do not involve any possible justification of necessity.

Finally, while it seems reasonable to assume that individuals perceive the utility of the risk in the Transplant problem to be considerably less than 1, it also seems plausible to infer that the utility of this risk is perceived to be considerably greater than that of the structurally similar Footbridge problem. Successful transplants, after all, are much more probable than using a man to stop or slow down an onrushing train. Yet roughly the same proportion of individuals (around 90%) judges these actions to be impermissible (Mikhail, 2007). While this does not necessarily imply that these actions are held to be morally equivalent — the Footbridge problem, for example, could be held to involve reckless behavior in a way that the Transplant problem does not — it does suggest that the Moral Calculus of Risk may play a subordinate operative role in these problems, whereas a more dominant role is played by the prohibition of purposeful battery.

The precise role of the Moral Calculus of Risk in intuitive moral judgments and its relation to other moral principles is obviously an important topic, which requires careful and thorough investigation that goes beyond the scope of this chapter. We will return to it briefly in Sections 4.5, 4.6, and 5. Here I will simply make the following clarifications, as a way of summarizing the foregoing remarks and anticipating that discussion. In our model, we generally make the simplifying assumption that the magnitude, utility, and necessity of the risk in the 12 cases in Table 1 are to be given a value of 1, rather than another more realistic value. The lone exception is the Better Alternative problem, for which the perceived necessity of the risk of throwing the switch is assumed to be 0, and thus the discounted expected benefit is also held to be 0. In the other eleven cases, we assume that the necessity of the risk is 1; hence, in these cases, the discounted expected benefit is assumed to be equivalent to the simple expected benefit.

4.5. The Rescue Principle

The Rescue Principle is a familiar precept of common morality — but not the common law — which has been defended by many writers, including Bentham (1948/1789), Scanlon (1998), Singer (1972), Unger (1996), and Weinrib (1980). Briefly, it holds that failing to prevent a preventable death or other grave misfortune is prohibited, where this can be achieved without risking one's own life or safety, or without violating other more fundamental precepts. It may be presumed to contain a *ceteris paribus* clause, the precise details of which need not detain us here.

The central element of the Rescue Principle in its core application is simple and intuitive: "Failing to rescue a person in grave danger is forbidden." In this section, we briefly describe how this principle can be explicated merely by concatenating elements we have already defined.

First, we need a formula to represent an omission or "negative act" (Bentham, 1948/1789, p. 72). To do this, we place the negation symbol in front of a complex act-token representation, thus taking the normal form of a *complex omission-token representation* to be given in (5):

(5) $\sim$[S's V-ing at $t^{(\alpha)}$]^{C}

As before, we assume that any expression obtainable by substituting permissibly for the individual variables in the normal form of a complex omission-token representation is also a complex omission-token representation. For example, "$\sim$[Hank's throwing the switch at $t^{(\alpha)}$]^{C}" symbolizes an omission, which can be paraphrased as "Hank's neglecting to throw the switch at time t in circumstances C," "Hank's not throwing the switch at time t in circumstances C," "It is not the case that Hank throws the switch at time t in circumstances C," and so forth.

Second, to interpret this formula, we adopt the standard convention of using brackets to restrict the scope of the negation symbol. Thus, ''[∼[Hank's throwing the switch at $t^{(\alpha)}]^C$ has features $F_1 \ldots F_n$]'' is a statement that refers to an omission and affirms that it has certain features. By contrast, ''∼[[Hank's throwing the switch at $t^{(\alpha)}]^C$ has features $F_1 \ldots F_n$]'' does not refer to an omission; rather, it is the negation of a statement that affirms that a complex-act-token representation has certain features. It can be paraphrased as ''It is not the case that Hank's throwing the switch at time t in circumstances C has features $F_1 \ldots F_n$'' or ''Hank's throwing the switch at time t in circumstances C does not have features $F_1 \ldots F_n$.''

By relying on this formula, together with the other concepts we have already explicated, we can now individuate 12 different purposely harmful acts and omissions and 12 different knowingly harmful acts and omissions, each of which can be formally described using the resources of our model. These twenty-four expressions are listed in Table 4, where they are divided into four groups: (i) purposely harmful acts, (ii) purposely harmful omissions, (iii) knowingly harmful acts, and (iv) knowingly harmful omissions.

With the aid of these expressions, one can consider various formulations of the Rescue Principle and ascertain which, if any, are descriptively adequate. We will not pursue this inquiry here, beyond making the following general observations. First, while a simple rescue principle that forbids any knowingly harmful omission is capable of explaining the intuition that Paul has a duty to throw the switch in the Costless Rescue problem, it clearly conflicts with all those cases in Table 1 in which harmful omissions are held to be permissible. Likewise, a simple rescue principle that forbids any act of (i) letting die, (ii) failing to prevent bodily harm to a person, or (iii) failing to prevent the destruction of a valuable object can also be shown to be inadequate. The first conflicts with the Footbridge problem (among others), while the second and third can easily be falsified by designing two new problems in which killing five persons is set against preventing bodily harm to one person and destroying a valuable object, respectively. Among other things, this implies that an adequate rescue principle must be a *comparative* rather than a *noncomparative* principle, which compares an act or omission with its alternatives (Lyons, 1965; Mikhail, 2002). It further suggests, although it does not entail, that an adequate rescue principle must occupy a subordinate position in a ''lexically ordered'' scheme of principles, in which at least some negative duties to avoid harm are ranked higher than at least some positive duties to prevent harm (Rawls, 1971, pp. 40–45; cf. Foot, 1967; Russell, 1977).[8] In particular, on the basis of the Footbridge, Intentional Homicide, Loop Track, and Drop Man problems, one may infer

[8] A lexical order is not entailed because there are other ways to solve the priority problem (Rawls, 1971).

Table 4 Purposely and Knowingly Harmful Acts and Omissions.

Purposely Harmful Acts
[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's committing homicide at $t^{(\beta)}$]
[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's committing battery at $t^{(\beta)}$]
[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's U-ing at $t^{(\beta)}$ [BAD EFFECT]]
[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Person, Death)]]
[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Person, Harm-$_{B}$)]]
[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Thing-$_{V}$, Destroy)]]
Purposely Harmful Omissions
~[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's committing homicide at $t^{(\beta)}$]
~[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's committing battery at $t^{(\beta)}$]
~[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's U-ing at $t^{(\beta)}$ [BAD EFFECT]]
~[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Person, Death)]]
~[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Person, Harm-$_{B}$)]]
~[S's V-ing at $t^{(\alpha)}$]^{C} *I-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Thing-$_{V}$, Destroy)]]
Knowingly Harmful Acts
[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's committing homicide at $t^{(\beta)}$]
[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's committing battery at $t^{(\beta)}$]
[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's U-ing at $t^{(\beta)}$ [BAD EFFECT]]
[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Person, Death)]]
[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Person, Harm-$_{B}$)]]
[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Thing-$_{V}$, Destroy)]]
Knowingly Harmful Omissions
~[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's committing homicide at $t^{(\beta)}$]
~[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's committing battery at $t^{(\beta)}$]
~[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's U-ing at $t^{(\beta)}$ [BAD EFFECT]]
~[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Person, Death)]]
~[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Person, Harm-$_{B}$)]]
~[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT (Thing-$_{V}$, Destroy)]]

that purposeful homicide and, at a minimum, purposeful battery that results in knowing homicide, are each lexically prior to the Rescue Principle — at least in circumstances other than a potential catastrophe or "supreme emergency" (Rawls, 1999; Walzer, 1977; cf. Nichols and Mallon, 2006).

Determining the precise nature of a descriptively adequate rescue principle is beyond the scope of this chapter. Instead, we merely state the following relatively simple yet demanding version of the principle as it relates the death of a person, which appears to be consistent with the data in Table 2, along with some further natural extensions:

The Rescue Principle (provisional version, applied to least harmful alternative)

$\sim$[S's V-ing at $t^{(\alpha)}]^{C}$ *K-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT [neg [neg [(Person, Death)]]]]] $\supset$ ([$\sim$[S's V-ing at $t^{(\alpha)}]^{C}$ is forbidden] $\equiv$

(a) $\sim$[[S's V-ing at $t^{(\alpha)}]^{C}$ *I-generates* [S's committing homicide at $t(^{\beta})$]];

(b) $\sim$[[S's V-ing at $t^{(\alpha)}]^{C}$ *I-generates* [S's committing battery at $t(^{\beta})$]];

(c) $\sim$[[S's V-ing at $t^{(\alpha)}]^{C}$ *K-generates* [S's U-ing at $t(^{\beta})$ [BAD EFFECT]] $<_{m}$ [S's V-ing at $t^{(\alpha)}]^{C}$ *K-generates* [S's U-ing at $t(^{\beta})$ [EFFECT [neg [BAD EFFECT]]]]])

Several aspects of this provisional formula merit attention. First, the principle is formulated as a comparative rather than a noncomparative principle; specifically, it compares one type of knowingly harmful omission with its least harmful alternative, the precisely relevant act-token being omitted under the circumstances, and it forbids the former just in case the latter does not possess certain features. Second, the principle holds that an omission that K-generates the double negation of the death of a person is forbidden just in case its least harmful alternative neither: (a) I-generates homicide, (b) I-generates battery, nor (c) K-generates bad effects that are morally worse than the negation (i.e., prevention) of the bad effects that it K-generates. This sounds exceedingly complex, but in plain English it simply means that the only justifications for knowingly letting a person die in our simple model are that doing so constitutes purposeful homicide, purposeful battery, or knowing homicide whose (discounted) expected benefits do not exceed its expected costs. More simply, preventing death is obligatory in our model unless doing so requires purposeful homicide, purposeful battery, or unjustified costs. The principle thus explains the Costless Rescue problem, yet it is also consistent with the other eleven problems in Table 1. Third, the principle is limited to the *knowingly* harmful omission of letting die. While one could expand the principle to include *purposely* harmful omissions, such as the deliberate letting die that Rachels (1975) depicts in the second of his famous Bathtub examples, in which a man purposely refrains from saving his drowning cousin in order to receive a large inheritance, this is unnecessary for our purposes: in light of our theory of how intentional structure is computed (Section 6.4), we may safely assume that none of the omissions in Table 1 are represented as purposely harmful (with the possible exception of the Intentional Homicide problem, where the actor's bad intent conflicts with a default rule of good intentions that we assume operates in this context; see Section 6.4). Fourth, the principle implies that pursuing the greater good in the Bystander, Implied Consent, Man-In-Front, and Drop Man problems is not only permissible, but obligatory, a strong assumption that is consistent with, but goes beyond, the data in Table 2. A weaker explanation might seek to

accommodate a principle of pacifism, according to which knowing homicide is never obligatory, at least in the type of circumstances at issue here (cf. Thomson, 1985, p. 280).

Fifth, the principle incorporates the assumptions we made in Section 4.4 about the magnitude, utility, and necessity of the risk. Condition (c) merely specifies the Moral Calculus of the Risk under those assumptions, as does restricting the scope of the principle to the least harmful alternative. We return to the significance of this restriction in Section 4.6. Sixth, the fact that condition (b) is given as purposeful battery, rather than purposeful battery that results in knowing homicide, reflects the stronger of two possible explanations from a deontological perspective of what constitutes an independent and adequate ground for precluding a duty to rescue that is consistent with the data in Table 2. A weaker assumption would appeal to purposeful battery that results in knowing homicide as the operative analysis of the Footbridge, Loop Track, and Drop Man problems. Finally, the presence of conditions (a) and (b) in the principle reflects the presumed lexical priority in common morality of at least some negative duties to avoid harm over some positive duties to prevent harm, and the possibility of justifying breaches of the latter, but not the former, by the Moral Calculus of Risk. Put differently, the Rescue Principle as it is formulated here is both consistent with and closely related to the Principle of Double Effect (PDE), a topic to which we now turn.

4.6. The Principle of Double Effect

The PDE is a complex principle of justification, which is narrower in scope than the traditional necessity defense because it places limits on what might otherwise be justified on grounds of necessity. Historically, the principle traces to Aquinas' attempt to reconcile the prohibition of intentional killing with the right to kill in self-defense. Denying any contradiction, Aquinas (1988/1274, p. 70) observed: ''One act may have two effects only one of which is intended and the other outside of our intention.'' On Aquinas' view, the right to kill in self-defense is thus apparently limited to cases in which death is a side effect of defending oneself against attack. It does not apply when the attacker's death is directly intended.

Our question here is not whether the PDE is a sound principle of normative ethics, but whether it is descriptively adequate, or at least captures the implicit logic of common moral intuitions to a useful first approximation (Harman, 1977; Nagel, 1986). In other words, our practical concern is whether the PDE can be strategically utilized to identify elements of moral grammar and other building blocks of intuitive jurisprudence. Likewise, because our main objective is to construct a computational theory of moral cognition along the lines of Marr's (1982) first level, we are not

concerned here with how the PDE or whatever mental operations it implies are actually implemented in our psychology, nor with whether those operations are modular in Fodor's (1983) sense, or otherwise informationally encapuslated (for some interesting discussion of these topics, see, e.g., Dupoux and Jacob, 2007; Greene, 2008b; Hauser et al., 2008a,b; Mallon, 2008; Nichols, 2005; Patterson, 2008; Prinz, 2008a,b; Sripada, 2008a,b; Stich, 2006). We merely assume that they are implemented in some manner or other, and that our analysis will help guide the search for underlying mechanisms, much as the theory of linguistic and visual perception has improved our grasp of the cognitive architecture and underlying mechanisms in these domains.

Many different versions of the PDE exist in the literature (see, e.g., Woodward, 2001; cf. Mikhail, 2000, pp. 160–161). According to the version we will develop here, the principle holds that an otherwise prohibited action, such as battery or homicide, which has both good and bad effects may be permissible if the prohibited act itself is not directly intended, the good but not the bad effects are directly intended, the good effects outweigh the bad effects, and no morally preferable alternative is available. In this section, we briefly describe how this principle can be rendered in a format suitable for premises of a derivation. Moreover, as we did with the Rescue Principle, we explain how this can be accomplished merely by concatenating elements we have already defined. In this manner, we show how what appears on the surface to be a rather complex moral principle can be broken down into its relatively simple psychological constituents.

At least six key terms in the PDE must be explained: (i) otherwise prohibited action, (ii) directly intended, (iii) good effects, (iv) bad effects, (v) outweigh, and (vi) morally preferable alternative. In our model, we interpret these concepts as follows.

First, we use the notions of I-generation, homicide, and battery to explicate the meanings of ''otherwise prohibited action'' and ''directly intended.'' While there are four *prima facie* wrongs in our simple model — purposeful homicide, purposeful battery, knowing homicide, and knowing battery — only the first two are directly intended under the meaning we assign them here, which equates ''directly intended'' with ''I-generated.'' As a result, these two actions cannot be justified by PDE, as we interpret it here. By contrast, knowing homicide and knowing battery can in principle be justified by the PDE. Unless they are justified in this manner, however, knowing homicide and knowing battery are forbidden.[9] There is no circularity, therefore, in referring to them as ''otherwise prohibited actions'' that can be justified under certain limited circumstances.

[9] Because our model is concerned only with explicating the data in Table 2, we need not consider other possible justifications or excuses, such as self-defense, duress, mental illness, etc.

Formally, these four *prima facie* prohibitions can be stated as follows:

Prohibition of Purposeful Homicide
[S's V-ing at $t^{(\alpha)}]^{C}$ *I-generates* [S's committing homicide at $t^{(\beta)}$] $\supset$ [[S's V-ing at $t^{(\alpha)}]^{C}$ is prohibited]

Prohibition of Purposeful Battery
[S's V-ing at $t^{(\alpha)}]^{C}$ *I-generates* [S's committing battery at $t^{(\beta)}$] $\supset$ [[S's V-ing at $t^{(\alpha)}]^{C}$ is prohibited]

Prohibition of Knowing Homicide
[S's V-ing at $t^{(\alpha)}]^{C}$ *K-generates* [S's committing homicide at $t^{(\beta)}$] $\supset$ [[S's V-ing at $t^{(\alpha)}]^{C}$ is prohibited]

Prohibition of Knowing Battery
[S's V-ing at $t^{(\alpha)}]^{C}$ *K-generates* [S's committing battery at $t^{(\beta)}$] $\supset$ [[S's V-ing at $t^{(\alpha)}]^{C}$ is prohibited]

In these formulas, we use the predicate *prohibited*, rather than the predicate *forbidden*, to differentiate the function of these *prima facie* prohibitions in our model from those all-things-considered deontic rules that assign a status of forbidden to complex act-tokens, if they have certain features.

Second, the PDE requires that the good effects but not the bad effects must be directly intended. Because we equate the meaning of ''directly intended'' with ''I-generated,'' and because we have already specified the only good and bad effects in our model, it is simple enough to combine these representations in a manner that explicates the meaning of this condition. Formally, these two requirements can be rendered as follows:

Good Effects Directly Intended
[S's V-ing at $t^{(\alpha)}]^{C}$ *I-generates* [S's U-ing at $t^{(\beta)}$ [GOOD EFFECT]]

Bad Effects Not Directly Intended
$\sim$[[S's V-ing at $t^{(\alpha)}]^{C}$ *I-generates* [S's U-ing at $t^{(\beta)}$ [BAD EFFECT]]]

Third, the PDE requires that the good effects outweigh the bad effects. Because we have already stipulated that the only good effects in our model consist of the negation of a bad effect, and because we have already relied on the morally worse-than relation to provide an ordinal ranking of bad effects, this condition can also be straightforwardly explained, at least insofar as we limit our attention to the 12 cases in Table 1. The key observation is that the good effects of an action can be said to outweigh its bad effects just in case the bad effects that the action *prevents* are morally worse than the bad effects that the action *causes*. Here, it should be recalled that the only good effects in our simple model consist in the negation (or prevention) of certain specified bad effects (Section 4.4). Consequently, this condition can be formalized as follows:

Good Effects Outweigh Bad Effects (Full Version)
[[S's V-ing at $t^{(\alpha)}]^{C}$ *K-generates* [S's U-ing at $t^{(\beta)}$ [BAD EFFECT]] $<_{m}$ [S's V-ing at $t^{(\alpha)}]^{C}$ *K-generates* [S's U-ing at $t^{(\beta)}$ [EFFECT [neg [BAD EFFECT]]]]]

Good Effects Outweigh Bad Effects (Abbreviated Version)
[BAD EFFECT_{P}] $<_{m}$ [BAD EFFECT_{C}]

The first formula, which we already encountered in the Rescue Principle, holds that the bad effects K-generated by a complex act-token representation are morally worse than the negation of those bad effects that are also K-generated by that act-token representation. The second formula abbreviates and incorporates a new notation for stating the same proposition, using "BAD EFFECT_{P}" to refer to the bad effects an actor knowingly prevents and "BAD EFFECT_{C}" to refer to the bad effects that she knowingly causes. Because the second formula shifts our focus from the (good) effect that consists of the negation of a bad effect to the bad effect that an actor knowingly prevents, the two sides of the relation are exchanged (cf. Sections 4.4 and 4.5).

Finally, the PDE demands that no morally preferable alternative be available. This is an important condition of the PDE that is often overlooked or ignored, causing the principle to seem unduly lax because it appears to justify knowingly harmful acts as long as their good effects outweigh their bad effects, without further qualification. Among other things, to understand this condition we need to know the meaning of "morally preferable" and "alternative." In our model, we explicate this condition as follows. First, we take the alternative to a given action to refer in the first place to *omission* rather than *inaction*; that is, to the failure to perform a specific act-token, rather than the failure to do anything at all (cf. Section 4.5). Second, we interpret the no-morally preferable-alternative condition to require comparing a given action to its *least harmful* omission. In all but one of our examples, there is only one possible alternative to the given action, hence the least harmful omission is identical with failing to perform that action. In the Better Alternative problem, by contrast, there are two possible alternatives, only one of which is the least harmful. Third, to decide which of several possible omissions is least harmful, we fall back on two comparative measures we have already explicated: (i) the morally worse-than relation, and (ii) the Moral Calculus of Risk. Finally, to decide whether the least harmful omission is morally preferable to the given action, we rely not only on (i) and (ii), but also (iii) the presumed lexical priority of the prohibition of purposeful homicide to the prohibition of knowing homicide, and the presumed lexical priority of the prohibition of purposeful battery (or, alternatively, the prohibition of purposeful battery that results in knowing homicide) to the Rescue Principle (cf. Section 4.5). By drawing on (i)–(iii), the computations for deciding whether a morally preferable alternative exists can be made without introducing any new evaluative concepts into our model, which thus can be kept as parsimonious as possible.

Formally, the comparison-to-the-least-harmful-omission component of the no-morally-preferable-alternative condition can be rendered for our purposes as follows:

No Less Harmful Alternative (Full Version)
$\sim$[S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's U-ing at $t^{(\beta)}$ [BAD EFFECT $_{LHA}$]] $<_{m}$ [S's V-ing at $t^{(\alpha)}$]^{C} *K-generates* [S's U-ing at $t^{(\beta)}$ [BAD EFFECT]]

No Less Harmful Alternative (Abbreviated Version)
[BAD EFFECT $_{LHA}$] $<_{m}$ [BAD EFFECT$_{C}$]

The first formula holds that the bad effect K-generated by the least harmful alternative to a complex act-token representation is morally worse than the bad effect K-generated by that act-token representation. In this formula, "BAD EFFECT$_{LHA}$" refers to the bad effect of the least harmful alternative (which must be calculated separately, of course, a task whose complexity grows with the increase of available alternatives and may become computationally intractable or inefficient beyond a certain point, one plausible source of so-called "omission bias"; cf. Baron and Ritov, this volume). The second formula abbreviates the same proposition, again using "BAD EFFECT$_{C}$" to refer to the bad effects that are caused.

The PDE has been the subject of intense scrutiny in the literature in recent years. Nonetheless, this discussion has often obscured both its virtues and limitations, and the foregoing analysis indicates one reason why. Many writers have assumed that the "natural application" (Quinn, 1993, p. 179) of the PDE is to state conditions under which actions are prohibited. This way of putting the matter seems potentially misleading. The PDE is not a direct test of whether an action is right or wrong; rather, its status is that of a second-order priority rule (Rawls, 1971) or ordering principle (Donagan, 1977) whose proper application is to state the only conditions under which otherwise prohibited actions are (or may be) permissible. Put differently, the principle's natural application is to serve as a principle of justification, which states necessary and sufficient conditions for a presumptively wrong action to be justified. As such, it constitutes a precise explication of yet another commonsense principle: "A knowingly harmful action which would otherwise be wrong may be justifiable, if but only if no better option exists."

5. A Periodic Table of Moral Elements

All of the foregoing definitions could presumably be improved, but they are satisfactory to our purposes. By utilizing these concepts, we can now construct a "periodic" table of moral elements, which identifies the key recurring properties of the structural descriptions elicited by the 12 cases in Table 1, and which can be used to explain their deontic status (Table 5).

Table 5 A Periodic Table of Moral Elements (Version 1).

	Structural features									
	I-generates (end or means)			K-generates (side effect)			Remaining conditions of PDE			
Problem	Homicide	Battery	Bad Effect	Homicide	Battery	Bad Effect	E/M = Good Effect	BAD EFFECTS$_P$ $<_m$ Bad Effects$_C$	Bad Effects$_{LHA}$ $<_m$ Bad Effects$_C$	Deontic status
Bystander										
Act (throw switch)				X	X	X	X	X	X	Permissible
Omission (~throw switch)				X	X	X	X			*?*
Footbridge										
Act (throw man)		X		X	X	X	X	X	X	Forbidden
Omission (~throw man)				X	X	X	X			*Obligatory*
Expensive Equipment										
Act (throw switch)				X	X	X	X			Forbidden
Omission (~throw switch)						X	X	X	X	*Obligatory*
Implied Consent										
Act (throw man)						X	X	X	X	Permissible
Omission (~throw man)				X	X	X	X			*?*

Intentional Homicide										
Act (throw switch)	X	X	X	X	X	X		X	X	Forbidden
Omission (~throw switch)				X	X	X				*Obligatory*
Loop Track										
Act (throw switch)		X		X	X	X	X	X	X	Forbidden
Omission (~throw switch)				X	X	X	X			*Obligatory*
Man-In-Front										
Act (throw switch)				X	X	X	X	X	X	Permissible
Omission (~throw switch)				X	X	X	X			*?*
Costless Rescue										
Act (throw switch)							X	X	X	Obligatory
Omission (~throw switch)				X	X	X				*Forbidden*
Better Alternative										
Act (throw switch)				X	X	X	X	X		Forbidden
Omission: Alternative #1 (~throw switch, pull cord)							X	X	X	*Obligatory*
Omission: Alternative #2 (~throw switch, do nothing)				X	X	X	X			*Forbidden*
Disproportional Death										
Act (throw switch)				X	X	X	X			Forbidden
Omission (~throw switch)				X	X	X	X	X	X	*Obligatory*

(*Continued*)

Table 5 (*Continued*)

	Structural features									
	I-generates (end or means)			K-generates (side effect)			Remaining conditions of PDE			
Problem	Homicide	Battery	Bad Effect	Homicide	Battery	Bad Effect	E/M = Good Effect	BAD EFFECTS$_P$ $<_m$ Bad Effects$_C$	Bad Effects$_{LHA}$ $<_m$ Bad Effects$_C$	Deontic status
Drop Man										
Act (throw switch)		X		X	X	X	X	X	X	Forbidden
Omission (~throw switch)				X	X	X	X			*Obligatory*
Collapse Bridge										
Act (throw switch)				X	X	X	X	X	X	Permissible
Omission (~throw switch)				X	X	X	X			*?*

Like any graphic device for displaying certain properties or relations, the layout of Table 5 is meant to provide a systematic arrangement of its essential information. Beginning at the top and working down, the table is divided into three main columns: Problem, Structural Features, and Deontic Status, respectively. Broadly speaking, this layout matches that of Tables 2 and 3, as well as the other schemas and models we have previously used to develop the moral grammar hypothesis. The table's main value from this vantage point is the ability to predict the deontic status of a given act or omission based entirely upon its structural features and those of its available alternatives. All of these features are included in the Structural Features column. From another perspective, the table can simply be viewed as an alternative method for exhibiting part of the structural description of a given action, for which act trees are also a useful method.[10]

The Structural Features column is itself divided into three groups. The first group includes three of the six purposely harmful features that can be I-generated in our model: homicide, battery, or a bad effect. The last of these, it will be recalled, is a broad category that can include either death of a person, bodily harm to a person, or destruction of a valuable thing (Section 4.4). For convenience, I have listed only the Bad Effect category itself in Table 5, even though this results in some redundancy. A different table might list all six features, or perhaps only the three bad effects themselves (cf. Table 6). Because the notion of I-generation is meant to incorporate and replace what are commonly referred to as ends or means, these notions are included parenthetically in the heading of this first group of properties.

The second group of properties includes three of the six knowingly harmful features that can be K-generated in our model: homicide, battery, or a bad effect. Because K-generation is meant to incorporate and serve as a replacement for side effects, this notion is included parenthetically in the heading of this group of properties. Finally, the third group includes the three remaining conditions of the PDE not already encompassed by the first group: (i) good effects are directly intended, (ii) good effects outweigh bad effects, and (iii) no less harmful alternative. Table 5 uses ''E/M = Good Effect'' to label the first condition (where ''E/M'' is itself an abbreviation of ''End or Means''), ''BAD EFFECTS$_P$ $<_m$ BAD EFFECTS$_C$'' to label the second condition, and ''BAD EFFECTS$_{LHA}$ $<_m$ BAD EFFECTS$_C$'' to label the third condition.

Each problem in Table 5 has two or more rows, one each for an act and its alternatives. Since eleven of our cases afford only one alternative, the set of alternatives is generally listed as *omission*. The sole exception is the Better Alternative problem, whose alternatives are given as ''Omission: Alternative

[10] Note, however, that Table 5 conveys both more and less information than the act trees in Figure 5A. The computations required by the PDE are exhibited, for example, but temporal information is not.

Table 6 A Periodic Table of Moral Elements (Version 2).

	Structural features									
	Homicide		Battery		Other bad effect		Justification			Deontic
Problem	Purpose	Knowledge	Purpose	Knowledge	Purpose	Knowledge	Good	Useful	Necessary	status
Bystander										
Act (throw switch)		X		X			X	X	X	Permissible
Omission (~throw switch)		X		X			X			*?*
Footbridge										
Act (throw man)		X	X	X			X	X	X	Forbidden
Omission (~throw man)		X		X			X			*Obligatory*
Expensive Equipment										
Act (throw switch)		X		X			X			Forbidden
Omission (~throw switch)						X	X	X	X	*Obligatory*
Implied Consent										
Act (throw man)							X	X	X	Permissible
Omission (~throw man)		X		X			X			*?*
Intentional Homicide										
Act (throw switch)	X	X	X	X				X	X	Forbidden
Omission (~throw switch)		X		X						*Obligatory*

Loop Track							
Act (throw switch)	X	X	X	X	X	X	Forbidden
Omission (~throw switch)	X		X	X			*Obligatory*
Man-In-Front							
Act (throw switch)	X		X	X	X	X	Permissible
Omission (~throw switch)	X		X	X			*?*
Costless Rescue							
Act (throw switch)				X	X	X	Obligatory
Omission (~throw switch)	X		X				*Forbidden*
Better Alternative							
Act (throw switch)	X		X	X	X		Forbidden
Omission: Alternative #1 (~throw switch, pull cord)				X	X	X	*Obligatory*
Omission: Alternative #2 (~throw switch, do nothing)	X		X	X			*Forbidden*

(*Continued*)

Table 6 (*Continued*)

	Structural features									
	Homicide		Battery		Other bad effect		Justification			Deontic
Problem	Purpose	Knowledge	Purpose	Knowledge	Purpose	Knowledge	Good	Useful	Necessary	status
Disproportional Death										
Act (throw switch)		X		X			X			Forbidden
Omission (∼throw switch)		X		X			X	X	X	*Obligatory*
Drop Man										
Act (throw switch)		X	X	X			X	X	X	Forbidden
Omission (∼throw switch)		X		X			X			*Obligatory*
Collapse Bridge										
Act (throw switch)		X		X		X	X	X	X	Permissible
Omission (∼throw switch)		X		X			X			*?*

#1'' and ''Omission: Alternative #2,'' and listed in descending order from least to most harmful, in order to facilitate the required comparison with an act's least harmful alternative. While the prevalence of single-alternative acts in Table 5 might suggest otherwise, it is important to emphasize that trolley problems are exceptional in this regard. In most real-life situations, there are many alternatives to a given action (i.e., many possible omissions), and in these situations identifying the least harmful alternative will take on much greater importance than it does here, a point of considerable significance for civil litigation (see, e.g., Grady, 1989).

The last column lists the deontic status of each act and omission. The judgments gleaned directly from experiments are given in normal typeface, while those that were not, but which can be logically derived from them, are italicized. Because the principles of deontic logic imply that both the doing and the forbearing of a given action can be permissible without contradiction, but the same is not true of the other two deontic operators (see, e.g., Mikhail, 2008b), one cannot simply infer the deontic status of omissions in the Bystander, Implied Consent, Man-In-Front, and Collapse Bridge problems. They are thus marked as open questions, which could of course be investigated empirically.

Turning to the table's individual cells, the presence or absence of an ''X'' in each cell indicates the presence or absence of a given feature. As indicated, the only case in which the first (I-generates homicide), third (I-generates bad effect), or seventh (E/M = Good Effect) features are atypical is the Intentional Homicide problem. No other problem involves death or another bad effect as a means or an end. By contrast, the second feature (I-generates battery) is implicated in four cases, all of which are forbidden: namely, the Footbridge, Intentional Homicide, Loop Track, and Drop Man problems. Next, eleven structural descriptions include one or more knowingly harmful acts (K-generates homicide, battery, or bad effect). The only exception is the Costless Rescue problem. Likewise, eleven structural descriptions include one or more knowingly harmful omissions. The only exception is the Better Alternative problem.[11] Finally, the three residual conditions of the PDE are satisfied in eight cases. In four of these cases — the Bystander, Implied Consent, Man-In-Front, and Collapse Bridge problems — these conditions can be invoked to explain why otherwise prohibited actions are held to be justified.

Table 5 is not the only way to exhibit structural features in a tabular format. Another instructive example is Table 6. On this layout, which closely resembles but in some ways improves upon the basic conceptual scheme of both the first and second Restatements of Torts and the Model

[11] Here one should recall that each I-generated homicide, battery, or bad effect is also K-generated (Section 3.2). Hence these cells are checked in both the first and second groups in the Intentional Homicide problem.

Penal Code, structural features are divided into four groups: Homicide, Battery, Bad Effect, and Justification. The first three groups are each divided into two subgroups: Purpose and Knowledge. These labels replace their technical counterparts in Table 5, as do the three subgroups of the Justification category: Good, Useful, and Necessary.

Table 6 has several advantages over Table 5. As indicated, one advantage is that how structural features are now labeled largely comports with a common type of legal analysis. Moreover, the exceptions tend to be virtues rather than vices. For example, Table 6 implies that one can commit battery by omission. This stipulation is potentially at odds with the first and second Restatements of Torts, which include a voluntary act requirement for battery (cf. Section 4.2). Still, this layout enables us to exhibit certain priority rules that might otherwise go unnoticed, as I explain below. Likewise, Table 6 avoids relying on the intuitive but often misleading terms, ''killing'' and ''letting die,'' while nonetheless identifying two ways each of these acts can occur in our model, resulting in four different possibilities in all: purposely killing, knowingly killing, purposely letting die, and knowingly letting die. The table thus reinforces Thomson's (1985, pp. 283–284) apt observation ''that 'kill' and 'let die' are too blunt to be useful tools'' for solving the trolley problems, and that one therefore ought to look within these acts ''for the ways in which the agents would be carrying them out.''

A further advantage of Table 6 is that its justifications closely track the Moral Calculus of Risk (Section 4.4). As such, they largely reflect the common sense analysis of unintentional harm that underlies the common law of negligence. Ordinarily, when a reasonable person seeks to justify a knowingly or foreseeably harmful or risky act, she asks the following questions: Is it good? (That is, is the act directed toward a good or worthwhile end?) Is it useful? (That is, does the act promote utility, insofar as the harm avoided outweighs the harm done?) Is it necessary? (That is, is there a less harmful alternative?) These questions not only capture the core residual features of the PDE; they also are basically utilitarian, much like the traditional necessity defense. This is to be expected, since the residual features of the PDE and the necessity defense are largely identical within the confines of our model. It is important to recognize, however, that neither the PDE nor the traditional necessity defense is utilitarian in the conventional sense; rather, each is a species of ''negative utilitarianism'' (Popper, 1945; Smart, 1958), which justifies the lesser of two evils, but not knowingly harming another individual simply because doing so maximizes aggregate welfare.

Perhaps the biggest advantage of Table 6 is that it aligns structural features in a regular order that reflects the apparent lexical priority of some prohibitions over others in common morality. In particular, prohibited acts prioritized over prohibited omissions, and purposeful harms are prioritized over knowing harms. In addition, homicides as a group are prioritized over batteries as a group, which in turn are prioritized over bad effects as a group. Further, unlike Table 5, the Bad Effect category in

Table 6 is limited to the destruction of a valuable thing in order to avoid unnecessary overlap with those bad effects that are already implicit in the homicide (death of a person) and battery (bodily harm to a person) categories, respectively. The result is that each individual cell in Table 6 represents a prohibition that is presumably lexically prior (subsequent) to the cells to the right (left) of it. Likewise, with respect to act and omission, each cell represents a prohibition that is lexically prior (subsequent) to the one immediately below (above) it. Finally, this layout naturally suggests a series of novel experiments that can be used to test, refine, and, if necessary, revise these assumptions, while rounding out our analysis of the behavior of structural features by considering them in all logically possible permutations. In particular, a new set of probes can be designed that systematically manipulate as far as possible each of the eighteen variables (9 columns × 2 rows) into which the Structural Features column is divided (see, e.g., Table 7). Together with sophisticated techniques for measuring neurological activity, reaction-time, implicit bias, and other familiar psychological phenomena, these probes can be used to improve our understanding of moral competence beyond that which has been previously contemplated.[12] I will not pursue these lines of inquiry further here; instead, I simply identify them as objects of future research that grow directly out of the foregoing analysis.

6. Conversion Rules

As we have seen, for the PDE or another extensionally equivalent principle to be operative in moral cognition, the brain must have the resources to compute representations of an agent's ends, means, side effects, and available alternatives. It also must incorporate a calculus of some sort capable of identifying, ranking, and computing the probabilities of an act's good and bad effects. In *Rawls' Linguistic Analogy*, I sought to provide a basis for generating these representations by formalizing five commonsense moral postulates: (i) the death of a person is bad, (ii) preventing a bad effect is good, (iii) failing to prevent a bad effect is bad, (iv) good is to be done and pursued, and evil avoided, and (v) from a moral point of view, the life of one person is worth the same as that of another (Mikhail, 2000, pp. 163–169). However, I did not provide complete account of the relevant conversion rules. Such an

[12] For example, some readers might wonder why all of our stimulus materials have thus far used only Anglophone male names. One objective of doing so is precisely to study the potential prejudicial effect of other names—and their associated sexual, racial, religious, or other identities—on intuitive moral judgment. Regrettably, it is not unreasonable to assume that "Helen," Isabel," "Karen," or "Lydia" will yield different data than "Hank," "Ian," "Karl," or "Luke" (or that both might yield different data than, e.g., "Fawzy," "Hussein," "Mourid," or "Naguib"). To understand this phenomena and thus integrate the theory of moral grammar with research on implicit bias (see, e.g., Banaji et al., 2007; Kang, 2005), one must first establish a descriptive baseline from which comparisons can be made.

Table 7 Manipulating Structural Features: Fourteen New Trolley Problems.[a]

1a. **Bystander: Purposely Homicidal Omission (not throwing the switch, for purpose of letting die).** Adam is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Adam sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Adam is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the men. There is a man standing on the side track with his back turned. Adam can throw the switch, killing him; or he can refrain from doing this, letting the five die. *Adam then recognizes that the men walking across the tracks are individuals whom he hates with a passion and wants to see dead. ''I don't give a damn [am indifferent/care deeply] about that guy on the side track,'' Adam thinks to himself, ''but this is my chance to sit by and watch those bastards get killed.''* Is it morally permissible [*obligatory*] for Adam to throw the switch?

1b. **Bystander: Purposely Homicidal Act and Omission (throwing the switch for the purpose of killing; not throwing the switch, for the purpose of letting die).** Bob is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Bob sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Bob is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the men. There is a man standing on the side track with his back turned. Bob can throw the switch, killing him; or he can refrain from doing this, letting the five die. *Bob then recognizes that the men walking across the tracks and the man on the side track are individuals whom he hates with a passion and wants to see dead. ''Wow, this is my lucky day,'' Bob thinks to himself, ''this is my chance either to kill that jerk on the side track or to sit by and watch those bastards get killed.''* Is it morally permissible [*obligatory*] for Bob to throw the switch?

2a. **Footbridge: Purposely Homicidal Act (throwing the man for the purpose of killing).** Charles is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Charles sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the

brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Charles is standing next to a heavy object, which he can throw onto the track in the path of the train, thereby preventing it from killing the men. The heavy object is a man, standing next to Charles with his back turned. Charles can throw the man, killing him; or he can refrain from doing this, letting the five die. *Charles then recognizes that the man standing next to him is someone whom he hates with a passion and wants to see dead. ''I don't give a damn [am indifferent/care deeply] about saving those five men,'' Charles thinks to himself, ''but this is my chance to kill that bastard.''* Is it morally permissible for Charles to throw the man?

2b. **Footbridge: Purposely Homicidal Omission (not throwing the man, for the purpose of letting die).** Donald is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Donald sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Donald is standing next to a heavy object, which he can throw onto the track in the path of the train, thereby preventing it from killing the men. The heavy object is a man, standing next to Donald with his back turned. Donald can throw the man, killing him; or he can refrain from doing this, letting the five die. *Donald then recognizes that the men walking across the tracks are individuals whom he hates with a passion and wants to see dead. ''I don't give a damn [am indifferent/care deeply] about this guy next to me,'' Donald thinks to himself, ''but this is my chance to sit by and watch those bastards get killed.''* Is it morally permissible [*obligatory*] for Donald to throw the man?

2c. **Footbridge: Purposely Homicidal Act and Omission (throwing the man for the purpose of killing; not throwing the man, for the purpose of letting die).** Edward is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Edward sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Edward is standing next to a heavy object, which he can throw onto the track in the path of the train, thereby preventing it from killing the men.

The heavy object is a man, standing next to Edward with his back turned. Edward can throw the man, killing him; or he can refrain from doing this, letting the five die. *Edward then realizes that the men walking across the tracks and the man standing next to him are individuals whom he hates with a passion and wants to see dead. ''Wow, this is my lucky day,'' Edward thinks to himself, ''this is my chance either to kill this jerk standing next to me or to sit by and watch those bastards get killed.*'' Is it morally permissible *[obligatory]* for Edward to throw the man?

3a. **Expensive Rescue (destroying an expensive thing as a side effect of saving life).** Fred is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Fred sees what has happened: the driver of the train saw *a man walking* across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the *man*. It is moving so fast that *he* will *not be able to get off the track in time*. Fred is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from *killing the man*. There is *five million dollars of new railroad equipment lying across* the side track. Fred can throw the switch, *destroying the equipment;* or he can refrain from doing this, letting the *man die*. Is it morally permissible [*obligatory*] for Fred to throw the switch?

3b. **Inexpensive Rescue (destroying an inexpensive thing as a side effect of saving life).** George is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. George sees what has happened: the driver of the train saw *a man walking* across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the *man*. It is moving so fast that *he* will *not be able to get off the track in time*. George is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from *killing the man*. There is *an old wagon worth about five hundred dollars lying across* the side track. George can throw the switch, *destroying the wagon*; or he can refrain from doing this, letting the *man die*. Is it morally permissible [*obligatory*] for George to throw the switch?

4a. **Substituted Consent (harmful contact as a means of saving life; prevent suicide).** Jack is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Jack sees what has happened: the driver of the train saw a man walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the man. It is moving so fast that he will not be able to get off the track in time. Jack is standing next to the man, whom he can throw

off the track out of the path of the train, thereby preventing it from killing the man. The man is frail and standing with his back turned. *Jack believes he is trying to commit suicide by walking in front of the train.* Jack can throw the man, injuring him; or he can refrain from doing this, letting the man die. Is it morally permissible [*obligatory*] for Jack to throw the man?

4b. **Hypothetical Consent (harmful contact as a means of saving life; remove obstruction).** Quinn is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Quinn sees what has happened: the driver of the train saw *five men* walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the *men*. It is moving so fast that *they* will not be able to get off the track in time. Quinn is standing next to *a switch*, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the *men*. There is a man standing *in front of the switch* with his back turned. *To reach the switch in time, Quinn will need to grab and throw the man out of the way, thereby injuring him.* Quinn can throw the man, injuring him, *and then throw the switch, saving the men*; or he can refrain from doing this, letting the five die. Is it morally permissible [*obligatory*] for Quinn to throw the man?

8a. **Suicidal Rescue (knowingly killing oneself as a side effect of saving life).** Thomas is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Thomas sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Thomas is standing next to *the will tracks and can throw himself in front of the train*, thereby preventing it from killing the men. *Doing so will put his own life at risk, however, and will almost surely kill him.* Thomas can throw himself in front of the train, *killing himself*, but saving the five men; or he can refrain from doing this, letting the five die. Is it permissible [*obligatory*] for Thomas to throw *himself in front of the train*?

10a. **Efficient Risk (destroying a valuable thing as a side effect of saving a more valuable thing).** Upton is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Upton sees what has happened: the driver of the train saw *five million dollars of new railroad equipment lying* across the tracks and slammed on the brakes, but the brakes

failed and the driver fainted. The train is now rushing toward the *equipment*. It is moving so fast that *the equipment* will *be destroyed*. Upton is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from *destroying the equipment*. There *is an old wagon worth about one hundred dollars lying across* the side track. Upton can throw the switch, *destroying the wagon*; or he can refrain from doing this, letting *the equipment be destroyed*. Is it morally permissible *[obligatory]* for Upton to throw the switch?

10b. **Inefficient Risk (destroying a valuable thing as a side effect of saving a less valuable thing).** Xavier is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Xavier sees what has happened: the driver of the train saw *an old wagon worth about five hundred dollars lying across* the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the *wagon*. It is moving so fast that *the wagon* will *be destroyed*. Xavier is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from *destroying the wagon*. There *is five million dollars of new railroad equipment lying across* the side track. Xavier can throw the switch, *destroying the equipment*; or he can refrain from doing this, letting *the wagon be destroyed*. Is it morally permissible for Xavier to throw the switch?

11a. **Drop Equipment (destroying a valuable thing as a means of saving life).** Yale is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Yale sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Yale is standing next to a switch, which he can throw, that will drop a heavy object into the path of the train, thereby preventing it from killing the men. The heavy object is *five million dollars of new railroad equipment, which* is standing on a footbridge overlooking the tracks. Yale can throw the switch, *destroying the equipment*; or he can refrain from doing this, letting the five die. Is it morally permissible [*obligatory*] for Yale to throw the switch?

12a. **Collapse Bridge: Destroy Equipment (destroying a valuable thing as a side effect of saving life).** Zach is taking his daily walk near the train tracks when he notices that the train

that is approaching is out of control. Zach sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Zach is standing next to a switch, which he can throw, that will collapse a footbridge overlooking the tracks into the path of the train, thereby preventing it from killing the men. There is *five million dollars of new railroad equipment* standing on the footbridge. Zach can throw the switch, *destroying the bridge and equipment*; or he can refrain from doing this, letting the five die. Is it morally permissible [*obligatory*] for Zach to throw the switch?

[a] Italics in Table 7 identify salient differences between the given problem and its correspondingly numbered problem in Table 1.

account is needed, however, because a key theoretical question implied by the moral grammar hypothesis is how the brain manages to compute a full structural description that incorporates properties like ends, means, side effects, and *prima facie* wrongs, such as battery, even when the stimulus contains no direct evidence for these properties.

As Figure 6A implies, this problem may be divided into at least five parts. To compute an accurate structural description of a given act and its alternatives, the systems that support moral cognition must generate complex representations that encode pertinent information about their temporal, causal, moral, intentional, and deontic properties. An interesting question is whether these computations must be performed in any particular order. Offhand, it might seem that the order is irrelevant; however, this impression appears to be mistaken. In fact, it seems that these computations must be performed in the order depicted in Figure 6A, at least in our 12 primary cases, because to recognize the deontic structure of these actions, one must already grasp their intentional structure; to recognize their intentional structure, one must already grasp their moral structure; to recognize their moral structure, one must already grasp (at least part of) their causal structure; and finally, to recognize their (full) causal structure, one must already grasp their temporal structure. These assumptions reflect some classical philosophical ideas about the relevant mental operations. But how exactly does each individual manage to extract the relevant cues from an impoverished stimulus (Figure 6B) and convert what is given into a full structural description? The process appears to include the following main steps.

First, the brain must identify the relevant action-descriptions in the stimulus and order them serially according to their relative *temporal* properties (Figure 6C). Second, it must identify their *causal* structure by

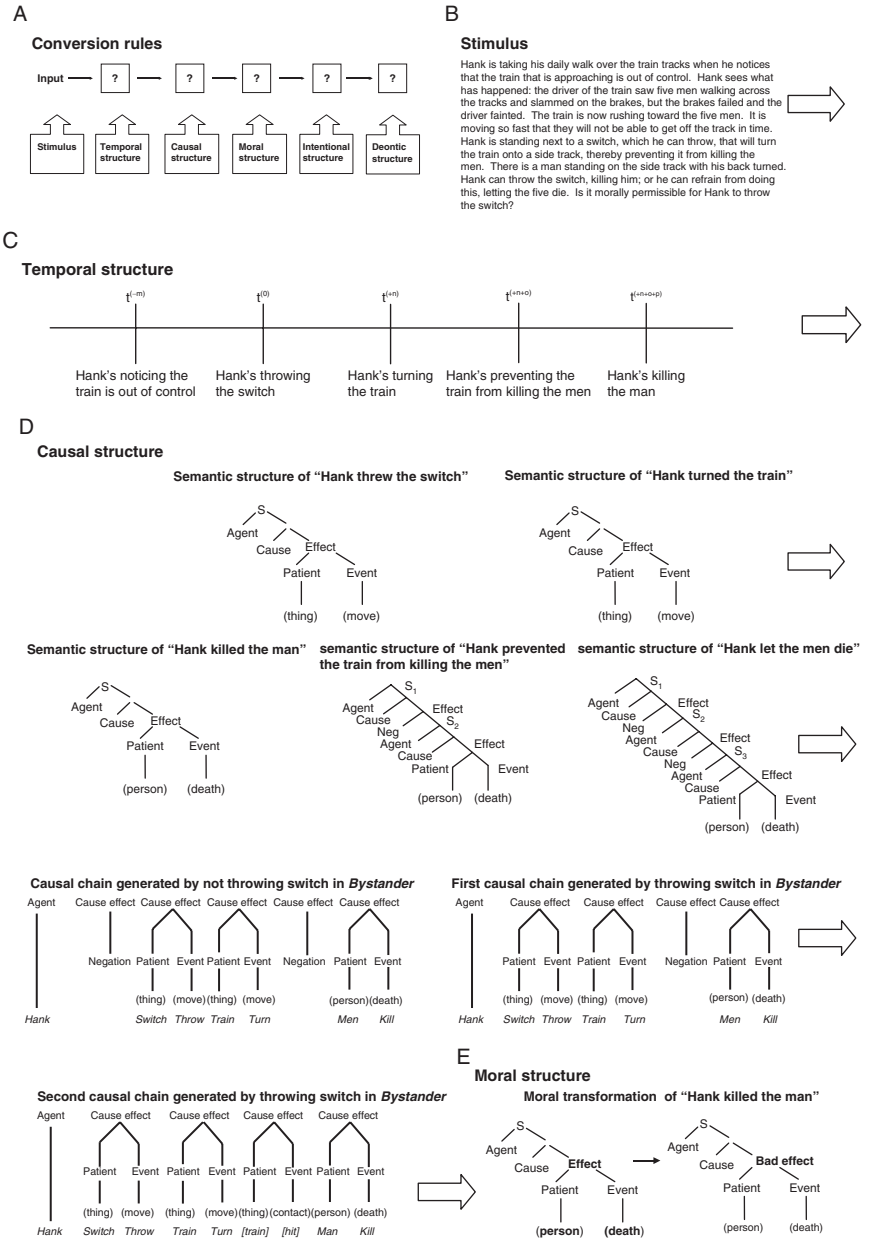


Figure 6 Computing Structural Descriptions.

decomposing the relevant causative constructions into their underlying semantic properties (Figure 6D). In addition, presumably by relying on temporal information, it must compute the full causal structure of the

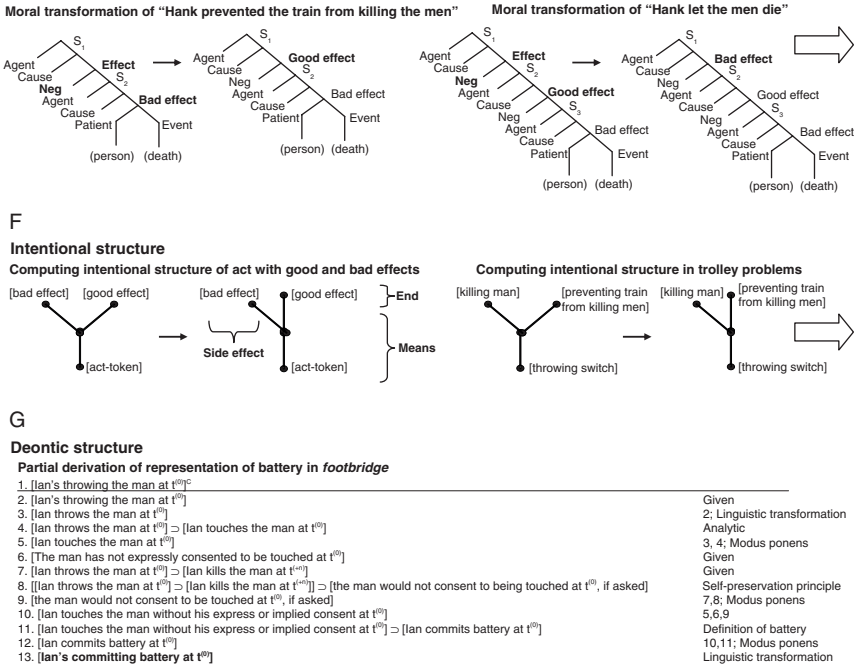


Figure 6 (Continued)

relevant acts and omissions by combining these percepts into ordered sequences of causes and effects (''causal chains''), supplying missing information where necessary (cf. Kant, 1965/1787). Figure 6D illustrates the three chains at issue in the Bystander Problem, linking (i) Hank's not throwing throw the switch to the effect of letting the men die, (ii) Hank's throwing the switch to the effect of preventing the train from killing the men, and (iii) Hank's throwing the switch to the effect of killing the man. In (iii), causing the train to hit the man is placed in brackets because this percept is not derived directly from the stimulus, but must be inferred from how objects interact with one another, presumably in accord with certain core knowledge of contact mechanics (Carey and Spelke, 1994; Spelke et al., 1992). In other words, the brackets identify one location in the causal chain where the brain supplies the missing information that killing the man requires causing the train to come into contact with him.

Third, the brain must compute the *moral* structure of the relevant acts and omissions by applying the following rewrite rules to the causal structures in Figure 6D: (i) an effect that consists of the death of a person is bad, (ii) an effect that consists of the negation of a bad effect is good, and (iii) an effect that consists of the negation of a good effect is bad. As a result, these

causal structures are transformed into richer representations that encode good and bad effects (Figure 6E). Moreover, since the second and third operations can be attributed to simple logical reasoning, and the first can be attributed, at least indirectly, to an instinct for self-preservation — the same likely source as that of the Prohibition of Homicide (Section 4.2) and the Self-Preservation Principle (Section 4.3) — we can explain this entire process merely by appealing to a common sociobiological instinct (cf. Darwin, 1981/1871, pp. 85–87; Hobbes, 1968/1651, p. 189; Leibniz, 1981/1705, p. 92; Proudhon, 1994/1840, pp. 170–174; Pufendorf, 2003/1673, p. 53).

Fourth, one must apply a presumption of good intentions, or what might be called a presumption of innocence, to the structures generated up to this point, thereby converting them into new structures that represent the *intentional* properties of the given action. That is, taking an act-token representation with both good and bad effects as a proximal input, the brain must (in the absence of countervailing evidence) generate its intentional structure by identifying the good effect as the *end* or *goal* and the bad effect as the *side effect* (cf. Section 3.2). This operation also can be represented graphically (Figure 6F). Note that some procedure of this general type must be postulated to explain how the brain computes ends, means, and side effects, since — crucially — there is no goal or mental state information in the stimulus itself. In Figure 6F, the presumption of good intentions acts as a default rule which says, in effect, that unless contrary evidence is given or implied, one should assume that S is a person of good will, who pursues good and avoids evil — another principle commonly held to be an innate instinct (see, e.g., Hume, 1978/1740, p. 438; cf. Aquinas, 1988/1274, p. 49; St. Germain, 1874/1518, p. 39). By relying on this principle, one can perhaps explain how the brain regularly computes representations of *mens rea*, even though goals and mental states are never directly observable.

Fifth, because the foregoing steps are necessary but not sufficient to explain the data in Table 2, the brain must supply some additional structure to the foregoing representations. What additional structure is necessary? One key insight of the moral grammar hypothesis is that adequate structural descriptions must also incorporate *prima facie* legal wrongs, such as battery or homicide. For example, in the Footbridge Problem, the brain must derive a representation of battery by inferring that (i) the agent must *touch* and *move* the man in order to throw him onto the track in the path of the train, and (ii) the man would not *consent* to being touched or moved in this manner, because of his desire for self-preservation (and because no contrary evidence is given). Utilizing standard notation in deductive logic (e.g., Leblanc and Wisdom, 1993), this line of reasoning argument can be also formalized (Figure 6G).

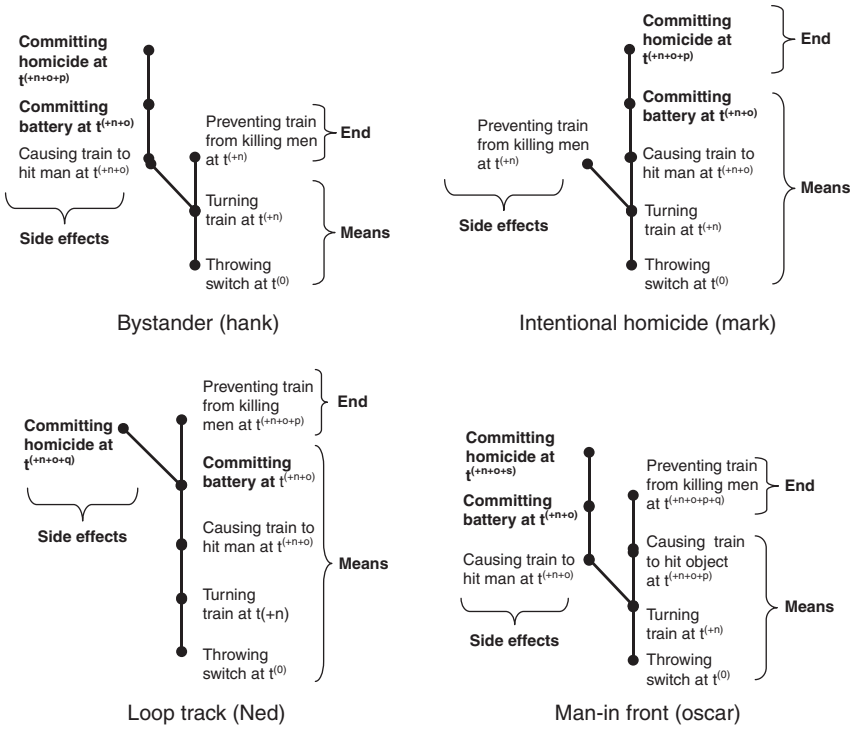


Figure 7 **Moral Geometry: Structural Descriptions of Four Equivalent Act-Type Descriptions.**

Sixth, because merely recognizing that the 12 cases in Table 1 implicate the legal categories of battery and homicide does not yet enable us to explain the data in Table 2, these violations must also be situated in the correct location of their associated structural descriptions, thereby identifying whether they are a means, end, or side effect. Figure 7 illustrates the outcome of this process for the four cases that we encountered in Section 2.2 whose act-type descriptions are completely equivalent.

Finally, once accurate structural descriptions of a given act-token representation and its alternatives (or at least the least harmful alternative of this potentially infinite set) are generated, the correct deontic rules must be applied to these descriptions to yield a considered judgment. Moreover, as we have observed (Section 4.4), the magnitude, utility, and necessity of the risk, and the comparative moral worth of the principal and collateral objects, must also be calculated and incorporated into these evaluations. This chapter has avoided many of the complexities that arise in this context, but they must be squarely confronted by any theory which purports to be descriptively adequate.

7. Conclusion

The model outlined in this chapter remains incomplete in many aspects, some of which have been highlighted along the way. For example, compensating for an apparent overemphasis on the role of emotions and heuristics in recent literature, I have avoided discussing these and other important topics in order to analyze a set of basic computations in ordinary moral cognition, whose subtlety, complexity, and explanatory power are often underestimated. The foregoing model is merely one component of an adequate moral psychology, however, and it must be integrated with general theories of affect, emotion, memory, motivation, prejudice, probabilistic reasoning, situationism, and a range of other cognitive systems and processes, particularly causal cognition and theory of mind, all of which have been fruitfully investigated in recent years. Moreover, the chapter does not supply any formal proofs, of course, and many gaps remain in the derivations I have sketched. Still, it seems clear from what has been achieved here that a complete theory of the steps converting proximal stimulus to intuitive response by means of an unconscious structural description could be given along the foregoing lines. In principle, a computer program could be devised that could execute these rapid, intuitive, and highly automatic operations from start to finish. The model outlined here thus goes some way toward achieving the first of Marr's (1982) three levels at which any information-processing task may be understood, the level of computational theory, because the abstract properties of the relevant mapping have been defined and its adequacy for the task at hand has been demonstrated. The model thus appears to be a significant advance in our understanding of intuitive moral judgment.

At the same time, we have discovered how certain fundamental legal conceptions can be utilized in this endeavor to explain an interesting range of moral intuitions, which prior experimental studies have indicated may be universal, or nearly so. By postulating latent knowledge of these and other basic legal norms, we can accurately predict human moral intuitions in a huge number and variety of actual cases. How this knowledge is acquired and put to use in different cultural, social, and institutional contexts thus emerge as pressing questions for law, philosophy, the social sciences, and the cognitive and brain sciences, broadly construed. As difficult to accept as it may seem, there are grounds for thinking that much of this knowledge may be innate or rooted in universal human instincts, as many cognitive scientists, philosophers, and jurists have often assumed. The argument is not conclusive, however, and more cross-disciplinary research is needed to clarify the relevant conceptual and evidentiary issues.

ACKNOWLEDGMENTS

Thanks to Dan Bartels, Noam Chomsky, Michael Dockery, Steve Goldberg, Tom Grey, Lisa Heinzerling, Ray Jackendoff, Emma Jordan, Mark Kelman, Greg Klass, Don Langevoort, Amanda Leiter, David Luban, Matthias Mahlmann, Doug Medin, Mitt Regan, Whitman Richards, Henry Richardson, Rebecca Saxe, Josh Tenenbaum, Robin West, and Allen Wood for helpful feedback, suggestions, and encouragement. This research was supported in part under AFOSR MURI award FA9550-05-1-0321.

REFERENCES

Alicke, M. (1992). Culpable Causation. *Journal of Personality and Social Psychology*, 63, 368–378.

Alter, A. L., Kernochan, J. and Darley, J. M. (2007). Morality Influences How People Apply the Ignorance of the Law Defense. *Law and Society Review*, 41, 819–864.

American Law Institute (1938). Restatement of the Law of Torts, as Adopted and Promulgated by the American Law Institute at Washington, DC. May 12, 1938, American Law Institute, St. Paul, MN.

American Law Institute (1965). *Restatement (Second) of Torts*. American Law Institute, St. Paul, MN.

American Law Institute (1985/1962). *Model Penal Code*. American Law Institute, Philadelphia, PA.

Anscombe, G. E. M. (1957). *Intention*. Basil Blackwell, Oxford.

Anscombe, G. E. M. (1958). Modern Moral Philosophy. *Philosophy*, 33, 1–19.

Aquinas, T. (1988/1274). In edited by Sigmund, P., (ed.), *St. Thomas Aquinas on Politics and Ethics*. Norton, New York.

Aquinas, T. (1952/1274). *The Summa Theologica of St. Thomas Aquinas*. Encyclopedia Britannica, Inc. (W. Benton, Publisher), Chicago.

Aristotle (1954). In edited by Ross, W. D. (ed.) *Nichomachean Ethics*. Oxford University Press, Oxford.

Banaji, M. R., Baron, A., Dunham, Y. and Olson, K. (2007). The Development of Intergroup Social Cognition: Early Emergence, Implicit Nature, and Sensitivity to Group Status, in edited by Killen, M. and Levy, S. (eds.), *Intergroup Relationships: An Integrative Developmental and Social Psychology Perspective* (pp. 87–102). Oxford University Press, Oxford.

Baron, J. and Ritov, I. (in press). Protected Values and Omission Bias as Deontological Judgments, in edited by D. Medin, L. Skitka, D. Bartels and C. Bauman (Eds.), *The Psychology of Learning and Motivation*, Vol. 50.

Baron, J. and Spranca, M. (1997). Protected Values. *Organizational Behavior and Human Decision Processes*, 70, 1–16.

Bartels, D. M. (2008). Principled Moral Sentiment and the Flexibility of Moral Judgment and Decision Making. *Cognition*, 108, 381–417.

Bartels, D. M. and Medin, D. L. (2007). Are Morally Motivated Decision Makers Insensitive to the Consequences of Their Choices? *Psychological Science*, 18, 24–28.

Bentham, J. (1948/1789). *An Introduction to the Principles of Morals and Legislation*. Halfner Press, New York.

Bradley, F. H. (1962/1876). *Ethical Studies*. Oxford University Press, Oxford.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA.

Brentano, F. (1969/1889). In edited by Chisholm, R., (ed.), *The Origin of The Knowledge of Right and Wrong*. Humanities Press, New York.

Bucciarelli, M., Khemlani, S. and Johnson-Laird, P. N. (2008). The Psychology of Moral Reasoning. *Judgment and Decision Making*, 3(2), 121–139.
Burlamaqui, J. (2006/1748). *The Principles of Natural Law and Politic*. Liberty Classics, Indianapolis.
Carey, S. and Spelke, E. (1994). Domain-Specific Knowledge and Conceptual Change, in edited by Hirschfield, L.A. and Gelmen, S.A., (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 169–200). Cambridge University Press, New York.
Cardozo, B. (1921). *The Nature of the Judicial Process*. Yale University Press, New Haven.
Chomsky, N. (1957). *Syntactic Structures*. Mouton, the Hague.
Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge.
Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge University Press, Cambridge.
Cushman, F. A. (2008). Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment. *Cognition*, 108(2), 353–380.
Cushman, F. A., Young, L. and Hauser, M. D. (2006). The Role of Reasoning and Intuition in Moral Judgments: Testing Three Principles of Harm. *Psychological Science*, 17(12), 1082–1089.
D'Arcy, E. (1963). *Human Acts: An Essay in Their Moral Evaluation*. Clarendon Press, Oxford.
Darwin, C. (1981/1871). *The Descent of Man, and Selection in Relation to Sex*. Princeton University Press, Princeton.
Davidson, D. (1963). Actions, Reasons, and Causes. *Journal of Philosophy*, 60, 685–700.
Descartes, R. (1985/1647). Comments on a Certain Broadsheet, in edited by Cottingham, J., Stoothoff, R., and Murdoch, D. (eds.), *The Philosophical Writings of Descartes*, Vol. 1. Cambridge University Press, Cambridge.
Donagan, A. (1977). *The Theory of Morality*. University of Chicago Press, Chicago.
Doris, J. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge University Press, Cambridge.
Dupoux, E. and Jacob, P. (2007). Universal Moral Grammar: A Critical Appraisal. *Trends in Cognitive Sciences*, 11, 373–378.
Durkheim, E. (1997/1893). *The Division of Labor in Society*. New York: The Free Press.
Dwyer, S. (1999). Moral Competence, in edited by Murasugi, K. and Stainton, R., (eds.), *Philosophy and Linguistics* (pp. 169–190). Westview Press, Boulder, CO.
Dwyer, S. (2006). How Good is the Linguistic Analogy? In edited by Carruthers, P., Laurence, S., and Stich, S. (eds.), *The Innate Mind*, Vol. 2, *Culture and Cognition*. Oxford University Press, Oxford.
Eldredge, L. (1941). *Modern Tort Problems*. George T. Bisel Company, Philadelphia.
Epstein, R. (2004). *Cases and Materials on Torts*. Aspen Publishers, New York.
Finnis, J. (1995). Intention in Tort Law, in edited by Owen, D. (ed.), *Philosophical Foundations of Tort Law*. Clarendon Press, Oxford.
Fiske, A. P. and Tetlock, P. E. (1997). Taboo Trade-Offs: Reactions to Transactions that Transgress the Spheres of Justice. *Political Psychology*, 18, 255–297.
Fodor, J. (1970). Three Reasons for Not Deriving "Kill" from "Cause to Die." *Linguistic Inquiry*, 1, 429–138.
Fodor, J. (1983). *The Modularity of Mind*. MIT Press, Cambridge.
Fodor, J. (1985). Precis of Modularity of Mind. *Behavioral and Brain Sciences*, 8(1), 1–42.
Foot, P. (1992/1967). The Problem of Abortion and the Doctrine of Double Effect, in Fischer, J. M. and Ravizza, M. (1992). *Ethics: Problems and Principles,* (pp. 60–67). Harcourt Brace Jovanovich, Fort Worth (Reprinted from *Oxford Review*, 5, 5–15).
Frege, G. (1980/1884). *The Foundations of Arithmetic,* in Austin, J. L. (trans.). Northwestern University Press, Evanston.
Freud, S. (1994/1930). *Civilization and Its Discontents*. Dover, New York.
Fried, C. (1978). *Right and Wrong*. Harvard University Press, Cambridge.

Geertz, C. (1973). Thick Description: Toward an Interpretative Theory of Culture, *in The Interpretation of Cultures: Selected Essays*. Basic Books, New York.
Gergely, G. and Csibra, G. (2003). Teleological Reasoning in Infancy: The Naive Theory of Rational Action. *Trends in Cognitive Sciences*, 7, 287–292.
Gilligan, C. (1978). *In a Different Voice*. Harvard University Press, Cambridge, MA.
Ginet, C. (1990). *On Action*. Cambridge University Press, Cambridge.
Gluckman, M. (1955). *The Judicial Process among the Barotse of Northern Rhodesia (Zambia)*. Manchester University Press, Manchester.
Gluckman, M. (1965). *The Ideas in Barotse Jurisprudence*. Manchester University Press, Manchester.
Goldman, A. (1970). *A Theory of Human Action*. Princeton University Press, Princeton.
Grady, M. (1989). Untaken Precautions. *Journal of Legal Studies*, 18, 139.
Greene, J. D. (2005). Cognitive Neuroscience and the Structure of the Moral Mind, in edited by Laurence, S., Carruthers, P., and Stich, S. (eds.), *The Innate Mind, Vol. 1, Structure and Contents*. Oxford University Press, New York.
Greene, J. D. (2008a). The Secret Joke of Kant's Soul, in edited by Sinnott-Armstrong, W. (ed.), *Moral Psychology*, Vol. 3, *The Neuroscience of Morality: Emotion, Disease, and Development*. MIT Press, Cambridge, MA.
Greene, J. D. (2008b). Reply to Mikhail and Timmons, in edited by Sinnott-Armstrong, W. (ed.), *Moral Psychology*, Vol. 3, *The Neuroscience of Morality: Emotion, Disease, and Development*. MIT Press, Cambridge, MA.
Greene, J. and Haidt, J. (2002). How (and Where) Does Moral Judgment Work? *Trends in Cognitive Sciences*, 6(12), 517–523.
Greene, J. D., Lindsell, D., Clarke, A. C., Nystrom, L. E. and Cohen, J. D. (submitted). Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. and Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, 293, 2105–2108.
Grey, T. (1983). Langdell's Orthodoxy. *University of Pittsburgh Law Review*, 45, 1–53.
Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, 108, 814–834.
Hallborg, R., Jr. (1997). Comparing Harms: The Lesser Evil Defense and the Trolley Problem. *Legal Theory*, 3, 291–316.
Hamlin, J. K., Wynn, K. and Bloom, P. (2007). Social Evaluation by Preverbal Infants. *Nature*, 450, 557–559.
Harman, G. (1977). *The Nature of Morality: An Introduction to Ethics*. Oxford University Press, New York.
Harman, G. (2000). *Explaining Value: And Other Essays in Moral Philosophy*. Oxford University Press, Oxford.
Harman, G. (2008). Using a Linguistic Analogy to Study Morality, in edited by Sinnott-Armstrong, W. (ed.), *Moral Psychology*, Vol. 1, *The Evolution of Morality: Adaptation and Innateness* (pp. 345–351). MIT Press, Cambridge.
Hart, H. L. A. and Honore, A. M. (1959). *Causation in the Law*. Oxford University Press, Oxford.
Hauser, M. (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. Harper Collins, New York.
Hauser, M. D., Cushman, F. and Young, L. (2008a). Reviving Rawls' Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions, in edited by Sinnott-Armstrong, W. (ed.), *Moral Psychology*, Vol. 2, *The Cognitive Science of Morality: Intuition and Diversity* (pp. 107–143). MIT Press, Cambridge.

Hauser, M. D., Cushman, F. and Young, L. (2008b). On Misreading the Linguistic Analogy: A Response to Jesse Prinz and Ron Mallon, in edited by Sinnott-Armstrong, W. (ed.), *Moral Psychology*, Vol. 2, *The Cognitive Science of Morality: Intuition and Diversity* (pp. 171–179). MIT Press, Cambridge.

Hauser, M. D., Cushman, F., Young, L., Jin, R. X. and Mikhail, J. (2007). A Dissociation between Moral Judgments and Justifications. *Mind & Language*, 22, 1–22.

Helmholtz, H. V. (1962/1867). *Helmholtz's Treatise on Physiological Optics*, *in* Southhall, J.P. C. (ed. and trans.). Dover, New York.

Hempel, C. (1955). Fundamentals of Concept Formation in Empirical Science, in edited by Neurath, O., Carnap, R., and Morris, C. (eds.), *International Encyclopedia of Unified Science*, Vol. 2. No. 7.

Hilliard, F. (1859). *The Law of Torts*. 2 volumes. Little, Brown and Company, Boston.

Hobbes, T. (1968/1651). *Leviathan*, in edited by Macpherson, C.B. (ed.), Penguin, New York.

Holmes, O. W. (1870). Codes, and the Arrangement of Law. *American Law Review*, 5, 1.

Holmes, O. W. (1991/1881). *The Common Law*. Dover, New York.

Hume, D. (1978/1740). *A Treatise of Human Nature*, in edited by Nidditch, P.H. (ed.), L.A., Selby-Bigge, *Analytical Index*. Clarendon Press, Oxford.

Hume, D. (1983/1751). In edited by Schneewind, J. B., (ed.), *An Enquiry Concerning the Principles of Morals*. Hackett, Indianapolis.

Hutcheson, J. (1929). The Judgment Intuitive. *Cornell Law Quarterly*, 14, 274.

Jackendoff, R. (1987). The Status of Thematic Relations in Linguistic Theory. *Linguistic Inquiry*, 18(3), 369–411.

Johnson, S. (2000). The Recognition of Mentalistic Agents in Infancy. *Trends in Cognitive Sciences*, 4(1), 22–28.

Jung, C. (1919). Instinct and the Unconscious. *British Journal of Psychology*, 10, 15–26; reprinted in *The Portable Jung* in edited by Campbell, J. (ed.). Viking Press, New York.

Kahneman, D. and Tversky, A. (1984). Choices, Values and Frames. *American Psychologist*, 39, 341–350.

Kang, J. (2005). Trojan Horses of Race. *Harvard Law Review*, 118, 1491–1593.

Kant, I. (1964/1785). *Groundwork of the Metaphysics of Morals,* in Paton, H. J. (trans.). Harper Perennial, New York.

Kant, I. (1965/1787). *Critique of Pure Reason,* in Smith, N. K. (trans.). New York: St. Martin's Press.

Kant, I. (1991/1797). *The Metaphysics of Morals,* in Gregor, M. (trans.). Cambridge University Press, Cambridge.

Kant, I. (1993/1788). *Critique of Practical Reason,* in Beck, L. W. (trans.). MacMillan, New York.

Katz, J. (1972). *Semantic Theory*. Harper & Row, New York.

Kelman, M. (1981). Interpretive Construction in the Substantive Criminal Law. *Stanford Law Review*, 33, 591.

Kelman, M., Rottenstreich, Y. and Tversky, A. (1996). Context-Dependence in Legal Decision Making. *Journal of Legal Studies*, 25, 287.

Kenny, A. (1995). Philippa Foot on Double Effect, in edited by Hursthouse, R., Lawrence, G. and Quinn, W. (eds.), *Virtues and Reasons: Philippa Foot and Moral Theory: Essays in Honor of Philippa Foot*. Clarendon Press, Oxford.

Knobe, J. (2005). Theory of Mind and Moral Cognition: Exploring the Connections. *Trends in Cognitive Sciences*, 9, 357–359.

Kohlberg, L. (1981). *Essays on Moral Development,* Vol. 1, *The Philosophy of Moral Development*. Harper & Row, New York.

Kohlberg, L. (1984). *Essays on Moral Development,* Vol. 2, *The Psychology of Moral Development*. Harper & Row, New York.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F. A., Hauser, M. D. and Damasio, T. (2007). Damage to Ventromedial Prefrontal Cortex Increases Utilitarian Moral Judgments. *Nature*, 446, 908–911.

Lashley, K. (1951). The Problem of Serial Order in Behavior, in edited by Jeffress, L. (ed.), *Hixon Symposium on Cerebral Mechanisms in Behavior*. John Wiley & Sons, New York.

Leblanc, H. and Wisdom, W. A. (1993). *Deductive Logic*. Prentice Hall, Englewood Cliffs, NJ.

LeFave, W. R. and Scott, A. W. (1972). *Handbook on Criminal Law*. West Publishing Co, St. Paul, MN.

Leibniz, G. W. (1981/1705). In edited by Remnant, P. and Bennet, J. (eds.), *New Essays on Human Understanding*. Cambridge University Press, Cambridge.

Locke, J. (1991/1689). In edited by Nidditch, P. (ed.), *An Essay Concerning Human Understanding*. Oxford University Press, Oxford.

Lombrozo, T. (2008). The Role of Moral Theories in Moral Judgment. *Poster Presented to the 34th Annual Meeting of the Society of Philosophy and Psychology*. Philadelphia, PA.

Lyons, D. (1965). *Forms and Limits of Utilitarianism*. Clarendon Press, Oxford.

Machery, E. (2007). The Folk Concept of Intentional Action: Philosophical and Experimental Issues. *Mind & Language*, 23, 165–189.

Mackie, J. (1974). *The Cement of the Universe: A Study of Causation*. Clarendon Press, Oxford.

Mahlmann, M. (1999). *Rationalismus in der praktishen Theorie: Normentheorie und praktische kampetenz*. Nomos Verlagsgesellschaft, Baden-Baden.

Mahlmann, M. (2007). Ethics, Law, and the Challenge of Cognitive Science. *German Law Journal*, 8, 577–615.

Mallon, R. (2008). Reviving Rawls' Linguistic Analogy Inside and Out, in edited by Sinnott-Armstrong, W. (ed.), *Moral Psychology*, Vol. 2, *The Cognitive Science of Morality: Intuition and Diversity* (pp. 145–155). MIT Press, Cambridge.

Macnamara, J. (1986). *A Border Dispute: The Place of Logic in Psychology*. MIT Press, Cambridge.

Marr, D. (1982). *Vision*. Freeman, San Francisco.

Meltzoff, A. (1995). Understanding the Intentions of Others: Re-enactment of Intended Acts by 18-Month-Old-Children. *Developmental Psychology*, 31, 838–850.

Mikhail, J. (2000). Rawls' Linguistic Analogy: A Study of the 'Generative Grammar' Model of Moral Theory Described by John Rawls in "A Theory of Justice." Unpublished PhD Dissertation, Cornell University.

Mikhail, J. (2002). Aspects of the Theory of Moral Cognition: Investigating Intuitive Knowledge of the Prohibition of Intentional Battery and the Principle of Double Effect. Georgetown University Law Center Public Law & Legal Theory Working Paper No. 762385. Available at http://ssrn.com/abstract=762385.

Mikhail, J. (2005). Moral Heuristics or Moral Competence? Reflections on Sunstein. *Behavioral and Brain Sciences*, 28(4), 557–558.

Mikhail, J. (2007). Universal Moral Grammar: Theory, Evidence, and the Future. *Trends in Cognitive Sciences*, 11, 143–152.

Mikhail, J. (2008a). The Poverty of the Moral Stimulus, in edited by Sinnott-Armstrong. W. (ed.), *Moral Psychology*, Vol. 1, *The Evolution of Morality: Adaptation and Innateness* (pp. 345–351). MIT Press, Cambridge.

Mikhail, J. (2008b). Moral Cognition and Computational Theory, in edited by Sinnott-Armstrong, W. (ed.), *Moral Psychology,* Vol. 3*, The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (pp. 81–91). MIT Press, Cambridge.

Mikhail, J. (in press). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge University Press, New York.

Mikhail, J., Sorrentino, C. and Spelke, E. (1998). Toward a Universal Moral Grammar, in edited by Gernsbacher, M. A. and Derry, S. J. (eds.), *Proceedings, Twentieth Annual*

Conference of the Cognitive Science Society (p. 1250). Lawrence Erlbaum Associates, Mahwah, NJ.
Mill, J. S. (1978/1859). In edited by Rappaport, E. (ed.), *On Liberty*. Hackett Publishing Co., Indianapolis.
Mill, J. S. (1987/1843). In edited by Ayer, A. J. (ed.), *The Logic of the Moral Sciences*. Duckworth, London.
Miller, G. (2008). The Roots of Morality. *Science*, 320, 734–737.
Moore, A., Clark, B. and Kane, M. (2008). Who Shall Not Kill? Individual Differences in Working Memory Capacity, Executive Control, and Moral Judgment. *Psychological Science*, 19, 549–557.
Nagel, T. (1986). The View from Nowhere, in edited by Fischer, J.M. and Ravizza, M. (eds.), *Ethics: Problems and Principles* (pp. 165–179). Harcourt Brace Jovanovich, Fort Worth, TX.
Nichols, S. (2002). *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford University Press, Oxford.
Nichols, S. (2005). Innateness and Moral Psychology, in edited by Laurence, S., Carruthers, P., and Stich, S. (eds.) *The Innate Mind*, Vol. 1. *Structure and Contents*. Oxford University Press, New York.
Nichols, S. and Mallon, R. (2006). Moral Dilemmas and Moral Rules. *Cognition*, 100(3), 530–542.
Nozick, R. (1968). Moral Complications and Moral Structures. *Natural Law Forum*, 13, 1–50.
Ogden, C. K. (1932). *Bentham's Theory of Fictions*. Kegan Paul, London.
Oliphant, H. (1928). A Return to Stare Decisis. *American Bar Association Journal*, 14, 71.
Patterson, D. (2008). On the Conceptual and the Empirical: A Critique of John Mikhail's Cognitivism. *Brooklyn Law Review*, 73, 611–623.
Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.
Petrinovich, L. and O'Neill, P. (1996). Influence of Wording and Framing Effects on Moral Intuitions. *Ethology and Sociobiology*, 17, 145–171.
Petrinovich, L., O'Neill, P. and Jorgensen, M. (1993). An Empirical Study of Moral Intuitions: Toward an Evolutionary Ethics. *Journal of Personality and Social Psychology*, 64, 467–478.
Piaget, J. (1965/1932). *The Moral Judgment of the Child*. The Free Press, New York.
Pinker, S. (2007). *The Stuff of Thought: Language as a Window into Human Nature*. Viking, New York.
Pinker, S. (2008). The Moral Instinct. *The New York Times Magazine*, January 13, 2008.
Popper, K. (1945). *The Open Society and Its Enemies*. Routledge, London.
Pound, R. (1908). Mechanical Jurisprudence. *Columbia Law Review*, 8, 620.
Prinz, J. (2008a). Is Morality Innate? In edited by Sinnott-Armstrong, W. (ed.), *Moral Psychology*, Vol. 1, *The Evolution of Morality: Adaptation and Innateness* (pp. 367–406). MIT Press, Cambridge.
Prinz, J. (2008b). Resisting the Linguistic Analogy: A Commentary on Hauser, Young, and Cushman. in edited by Sinnott-Armstrong, W., (ed.), *Moral Psychology*, Vol. 2, *The Cognitive Science of Morality: Intuition and Diversity* (pp. 157–170). MIT Press, Cambridge.
Prosser, W. (1941). *Casebook on Torts*. University of Minnesota Press, Minneapolis.
Prosser, W. (1971). *Casebook on Torts*. (4th edn.) University of Minnesota Press, Minneapolis.
Proudhon, P. J. (1994/1840). In edited by Kelly, D.R. and Smith, B.G. (eds.),. *What is Property?* Cambridge University Press, Cambridge.
Pufendorf, S. (2003/1673). In edited by Hunter, I. and Saunders, D., (eds.), *The Whole Duty of Man, According to the Law of Nature*, Tooke, A. (trans. 1691). Liberty Fund, Indianapolis.

Quinn, W. S. (1993). *Morality and Action*. Cambridge University Press, Cambridge.
Rachels, J. (1975). Active and Passive Euthanasia. *The New England Journal of Medicine*. In Fischer, J. M. and Ravizza, M. (1992), *Ethics: Problems and Principles* (pp. 111–116). Harcourt Brace Jovanovich, Fort Worth.
Radin, M. (1925). The Theory of Judicial Decision. *American Bar Association Journal*, 11, 357.
Rawls, J. (1971). *A Theory of Justice*. Harvard University Press, Cambridge, MA.
Rawls, J. (1975). The Independence of Moral Theory. *Proceedings and Addresses of the American Philosophical Association*, 48, 5–22.
Rawls, J. (1999). *The Law of Peoples*. Harvard University Press, Cambridge, MA.
Raz, J. (1970). *The Concept of a Legal System*. Clarendon Press, Oxford.
Rey, G. (2006). Conventions, Intuitions, and Linguistic Inexistents: A Reply to Devitt. *Croatian Journal of Philosophy*, 18, 549–569.
Robinson, P. H., Kurzban, R. and Jones, O. D. (2008). The Origins of Shared Intuitions of Justice. *Vanderbilt Law Review*, 60, 1633–1688.
Roedder, E. and Harman, G. (2008). Moral Theory: The Linguistic Analogy. Forthcoming, in edited by Doris, J., Nichols, S., and Stich, S. (eds.). *Empirical Moral Psychology*. Oxford University Press, Oxford.
Ryle, G. (1968). The Thinking of Thoughts: What is 'Le Pensuer' Doing? In *Collected Papers*, Vol. 2. *Collected Essays*. Hutchinson, London.
Russell, B. (1977). On the Relative Strictness of Negative and Positive Duties. in edited by Fischer, J. M. and Ravizza, M., (eds.), *Ethics: Problems and Principles*. Harcourt Brace Jovanovich, Fort Worth, TX.
Salmond, J. (1966/1902). In edited by Fitzgerald, P. J., (ed.), *Salmond on Jurisprudence*. (12th edn.). Sweet & Maxwell, London.
Savigny, F. C. V. (1881/1814). *On the Vocation of Our Age for Legislation and Jurisprudence,* in Hayward, A. (trans.). Littleward and Co, London.
Saxe, R. (2005). Do the Right Thing: Cognitive Science's Search for a Common Morality. *Boston Review*, September–October, 2005.
Scanlon, T. M. (1998). *What We Owe to Each Other*. Harvard University Press, Cambridge.
Schnall, S., Haidt, J., Clore, G. L. and Jordan, H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, 34(8), 1096–1109.
Sidgwick, H. (1907). *The Methods of Ethics*. (7th edn.). Hacket, Indianapolis.
Singer, P. (1972). Famine, Affluence, and Morality. *Philosophy and Public Affairs*, 1(3), 229–243.
Sinnott-Armstrong, W., Mallon, R., McCoy, T. and Hull, J. (2008). Intention, Temporal Order, and Moral Judgments. *Mind & Language*, 23(1), 90–106.
Smart, R. N. (1958). Negative Utilitarianism. *Mind*, 67, 542–543.
Smetana, J. (1983). Social Cognitive Development: Domain Distinctions and Coordinations. *Developmental Review*, 3, 131–147.
Solum, L. (2006). Natural Justice. *American Journal of Jurisprudence*, 51, 65–105.
Spelke, E. S., Breinlinger, K. and Jacobson, K. (1992). Origins of Knowledge. *Psychological Review*, 99, 605–632.
Sripada, C. S. (2008a). Nativism and Moral Psychology: Three Models of the Innate Structure that Shapes the Contents of Moral Norms, in edited by Sinnott-Armstrong, W. (ed.), *Moral Psychology*, Vol. 1, *The Evolution of Morality: Adaptation and Innateness* (pp. 319–343). MIT Press, Cambridge.
Sripada, C. S. (2008b). Reply to Harman and Mikhail, in edited by Sinnott-Armstrong, W. (ed.), *Moral Psychology*, Vol. 1, *The Evolution of Morality: Adaptation and Innateness* (pp. 361–365). MIT Press, Cambridge.
St. Germain, C. (1874/1518). *Doctor and Student, or Dialogues between a Doctor of Divinity and a Student in the Laws of England*. Robert Clarke & Co, Cincinnati.

Stich, S. (2006). Is Morality an Elegant Machine or a Kludge? *Journal of Cognition and Culture*, 6(1–2), 181–189.
Stone, J. (1968). *Legal System and Lawyers' Reasonings*. Stanford University Press, Stanford.
Sunstein, C. (2005). Moral Heuristics. *Behavioral and Brain Sciences*, 28, 531–573.
Terry, H. (1884). Some Leading Principles of Anglo-American Law, Expounded with a View to Its Arrangement and Codification. T. & J. W. Johnson & Co., Philadelphia.
Terry, H. (1915). Negligence. *Harvard Law Review*, 29, 40.
Tetlock, P. E. (2003). Thinking About the Unthinkable: Coping with Secular Encroachments on Sacred Values. *Trends in Cognitive Sciences*, 7, 320–324.
Thomson, J. J. (1970). The Time of a Killing. *Journal of Philosophy*, 68, 115–132.
Thomson, J. J. (1985). The Trolley Problem, in edited by Fischer, J. M. and Ravizza, M. (eds.), *Ethics: Problems and Principles* (pp. 67–76). Harcourt Brace Jovanovich, Fort Worth, TX.
Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press, Cambridge.
Unger, P. (1996). *Living High and Letting Die: Our Illusion of Innocence*. Oxford University Press, Oxford.
Valdesolo, P. and DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological Science*, 17(6), 476–477.
Waldmann, M. R. and Dieterich, J. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, 18(3), 247–253.
Walzer, M. (1977). *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. Basic Books, New York.
Wellman, H. and Miller, J. (2008). Including Deontic Reasoning as Fundamental to Theory of Mind. *Human Development* 51, 105–135.
Wheatley, E. (1980). The Case for a Duty to Rescue. *Yale Law Journal*, 90, 247.
Weinrib, T. and Haidt, J. (2005). Hypnotic Disgust Makes Moral Judgments More Severe. *Psychological Science*, 16, 780–784.
Wilson, J. (1967/1790–91). In edited by McCloskey, R. (ed.), *The Works of James Wilson*. 2 volumes. Havard University Press, Cambridge.
Woodward, P. A. (2001). *The Doctrine of Double Effect: Philosophers Debate a Controversial Moral Principle*. University of Notre Dame Press, Notre Dame.
Woodward, A. L., Sommerville, J. A. and Guajardo, J. J. (2001). How Infants Make Sense of Intentional Action, in edited by Malle, B., Moses, L., and Baldwin, D. (eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 149–169). MIT Press, Cambridge, MA.
Wright, R. (1985). Causation in Tort Law. *California Law Review*, 73, 1735.
Young, L., Cushman, F. A., Hauser, M. D. and Saxe, R. (2007). Brain Regions for Belief Attribution Drive Moral Condemnation for Crimes of Attempt. *Proceedings of the National Academy of Science*, 104(20), 8235–8240.
Young, L. and Saxe, R. (2008). The Neural Basis of Belief Encoding and Integration in Moral Judgment. *NeuroImage*, 40, 1912–1920.