Opinion TRENDS in Cognitive Sciences Vol.11 No.4



# Universal moral grammar: theory, evidence and the future

John Mikhail

Georgetown University Law Center, 600 New Jersey Avenue NW, Washington, DC 20001, USA

**Scientists from various disciplines have begun to focus attention on the psychology and biology of human morality. One research program that has recently gained attention is universal moral grammar (UMG). UMG seeks to describe the nature and origin of moral knowledge by using concepts and models similar to those used in Chomsky's program in linguistics. This approach is thought to provide a fruitful perspective from which to investigate moral competence from computational, ontogenetic, behavioral, physiological and phylogenetic perspectives. In this article, I outline a framework for UMG and describe some of the evidence that supports it. I also propose a novel computational analysis of moral intuitions and argue that future research on this topic should draw more directly on legal theory.**

## Introduction

This article outlines a framework for the study of human moral cognition, currently one of the liveliest topics in the cognitive sciences. The framework has come to be known as universal moral grammar (UMG) [1–6] because it seeks to describe the nature and origin of moral knowledge by using concepts and models similar to those used in the study of language [7]. UMG shares many features with other important research programs in moral psychology [8–15], particularly an emphasis on describing the operative principles of intuitive moral judgment and a departure from Kohlberg's [16] once-dominant paradigm, which shifted attention away from moral intuitions and towards the interpretive or 'hermeneutic' evaluation of articulate justifications. However, UMG also has certain distinctive characteristics that set it apart from other influential approaches, such as those of Greene [10], Haidt [11], Moll [12] and Sunstein [12]. First, UMG is organized around five main questions (Box 1), each of which has a direct parallel in linguistics and is interpreted in light of concepts that Chomsky used to clarify their linguistic counterparts, such as the distinctions between (i) competence and performance, (ii) descriptive and explanatory adequacy, and (iii) the perception and production problems [17–19]. Second, UMG proceeds from assumptions that are mentalist, modular and nativist [1–6,17–21]. Third, in keeping with Marr's [22] analysis of the three levels at which any information-processing task can be understood, UMG focuses special attention on the top level, the level of computational theory. The view it adopts is that, like other domains, an adequate scientific theory of moral cognition will often depend more on the computational problems that have to be solved than on the neurophysiological mechanisms in which those solutions are implemented [22].

## Initial evidence

Initial evidence for UMG comes from multiple sources, including psychology, linguistics, anthropology and cognitive neuroscience. Although none of this evidence is univocal or conclusive, collectively it provides at least modest support for the hypothesis that humans possess an innate moral faculty that is analogous, in some respects, to the language faculty that has been postulated by Chomsky and other linguists.

First, developmental psychologists have discovered that the intuitive jurisprudence of young children is complex and exhibits many characteristics of a well-developed legal code. For example, 3–4-year-old children use intent or purpose to distinguish two acts that have the same result [23]. They also distinguish 'genuine' moral violations (e.g. battery or theft) from violations of social conventions (e.g. wearing pajamas to school) [24]. 4–5-year-olds use a proportionality principle to determine the correct level of punishment for principals and accessories [25]. 5–6-year-olds use false factual beliefs but not false moral beliefs to exculpate [26].

Second, every natural language seems to have words or phrases to express basic deontic concepts, such as *obligatory*, *permissible*, and *forbidden*, or their equivalents [27]. Moreover, deontic logic is formalizable [28]. The three primary deontic operators can be placed in a square of opposition and equipollence, similar to those for quantified and modal forms (Box 2).

Third, prohibitions of murder, rape and other types of aggression appear to be universal or nearly so [29,30], as do legal distinctions that are based on causation, intention and voluntary behavior [30–32]. Furthermore, comparative legal scholars have suggested that a few basic distinctions capture the 'universal grammar' of all systems of criminal law [31,32].

Finally, functional imaging and patient studies have led some researchers to conclude that a fairly consistent network of brain regions is involved in moral cognition, including the anterior prefrontal cortex, medial and lateral orbitofrontal cortex, dorsolateral and ventromedial prefrontal cortex, anterior temporal lobes, superior temporal sulcus and posterior cingulate/precuneus region [8,12]. However, these findings are preliminary and controversial [8]. Moreover, some of the moral-judgment tasks they rely on seem to be poorly motivated, because they involve

*Corresponding author:* Mikhail, J. (mikhail@law.georgetown.edu).
Available online 27 February 2007.

 1364-6613/$ – see front matter  doi:10.1016/j.tics.2006.12.007

**Box 1. Five main questions of universal moral grammar**

- What constitutes moral knowledge?
- How is moral knowledge acquired?
- How is moral knowledge put to use?
- How is moral knowledge physically realized in the brain?
- How did moral knowledge evolve in the species?

stand-alone sentences that allegedly have moral or non-moral content (e.g. 'The elderly are useless' versus 'Stones are made of water'), rather than acts of conspecifics that can be carefully manipulated to test specific theories of mental representation. Nonetheless, these probes are likely to become more refined as our understanding of moral competence improves.

## Two fundamental arguments

In addition to providing an explanatory framework for these and related observations, UMG relies on two fundamental arguments: the argument for moral grammar and the argument from the poverty of the moral stimulus [6,20]. The argument for moral grammar holds that the properties of moral judgment imply that the mind contains a moral grammar: a complex and possibly domain-specific set of rules, concepts and principles that generates and relates mental representations of various types. Among other things, this system enables individuals to determine the deontic status of an infinite variety of acts and omissions [6,7]. The argument from the poverty of the moral stimulus holds that the manner in which this grammar is acquired implies that at least some of its core attributes are innate, where 'innate' is used in a dispositional sense to refer to cognitive systems whose essential properties are largely pre-determined by the inherent structure of the mind, but whose ontogenetic development must be triggered and shaped by appropriate experience and can be impeded by unusually hostile learning environments [1–6,18–21]. Both arguments are nondemonstrative and presuppose a familiar set of idealizations and simplifying assumptions [7,17–20]. Moreover, both arguments have direct parallels in the case of language and, like their linguistic counterparts, can be depicted graphically by simple perceptual and acquisition models (Figure 1).

## Socratic method

The models in Figure 1 are abstract, but one can begin to put some flesh on the bones by squarely confronting the problem of descriptive adequacy. UMG attempts to solve this problem by the 'Socratic' method [6,7] – that is, a method in which individuals are asked to provide their moral intuitions about a carefully selected class of real or hypothetical fact patterns. The cases are drawn primarily from various branches of law, such as criminal law, torts, contracts or agency. The aim is to discover whether individuals have stable and systematic intuitions about cases of first impression in these areas and, if so, whether these intuitions are best explained by assuming they possess tacit knowledge of specific rules, concepts or principles [33].

### *Trolley problems*

One set of examples that are particularly interesting and useful in this context are trolley problems [1,2,5,6,8,9,33–36], a well-known family of cases that inquire whether it is permissible to harm one or more individuals in the course of saving others (Box 3). Unlike Kohlberg's dilemmas [16], the moral judgments that these problems elicit are rapid, intuitive and made with a high degree of certitude – all properties that one associates with probes that are used elsewhere in cognitive science, such as language, vision, musical cognition and face recognition [20]. Moreover, the judgments appear to be widely shared among demographically diverse populations, including young children; even in large cross-cultural samples, participants' responses to these problems cannot be predicted by variables such as age, sex, race, religion or education [36]. Furthermore, individuals typically have difficulty producing compelling justifications for these judgments: thus, trolley-problem intuitions exhibit a dissociation between judgments and justifications [36,37], thereby illustrating the distinction between operative and express principles [5]. Finally, it is clear upon analysis that it is

**Box 2. Deontic concepts and deontic logic**

Every natural language appears to have words or phrases to express the three main deontic concepts (Figure Ia in this box). They comprise the basic categorization scheme of most human moral, legal and religious systems, and their natural domain is the voluntary acts and omissions of moral agents. These concepts also bear systematic logical relationships to one another, which can be represented in a square of opposition and equipollence, similar to those for modal and quantified forms (Figure Ib). This greatly reduces the complexity of the description of the deontic component of moral competence. Given these equations, only one of these concepts needs to be selected and taken to be primitive. Then, with the aid of two logical connectives and the concepts of act ('A') and omission ('not-A'), the expressions in Figure Ib can be mechanically defined [6].

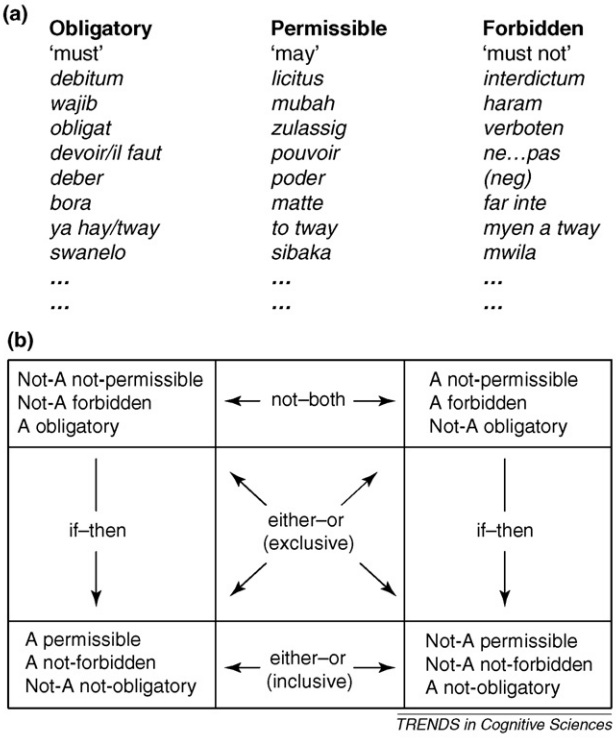


**Figure I.** Three deontic concepts **(a)** and square of opposition and equipollence **(b)**.

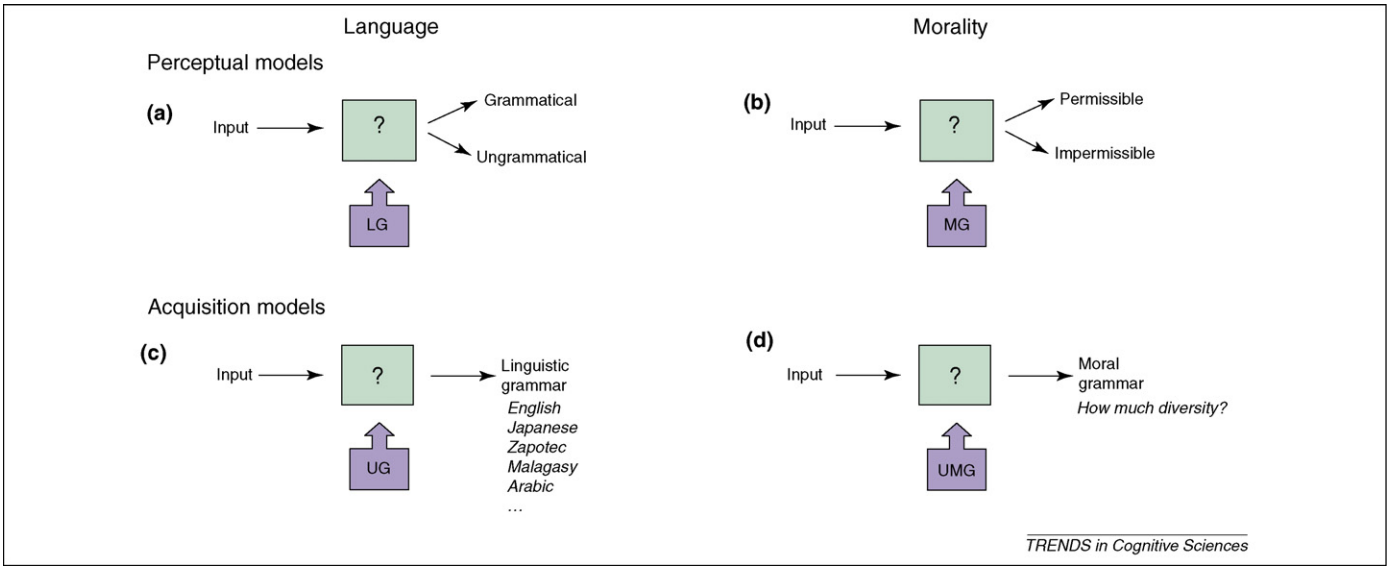


**Figure 1**. Simple perceptual and acquisition models for language and morality. Chomsky [17] clarified the objectives of linguistic theory by proposing that certain aspects of verbal behavior can be studied as an input–output relationship and by distinguishing two models that a linguistic theory must specify: a perceptual model and an acquisition model. **(a–b)** The initial goal in the theory of language is to determine how people can intuitively recognize the properties of novel expressions in their language, such as whether or not they are grammatical. This can be compared to the ability to determine whether a given action is permissible or impermissible. The question that motivated Chomsky's [17,56] early work is 'On what basis do people actually go about distinguishing grammatical from ungrammatical sentences?'. In (b), the related question is 'On what basis do people actually go about distinguishing permissible from impermissible acts?'. Provisional answers are provided by a linguistic grammar (LG) and moral grammar (MG) respectively. **(c)** Although the argument from the poverty of the stimulus implies that some linguistic knowledge is innate, the variety of human languages provides an upper bound on this hypothesis; what is innate must be consistent with the observed diversity of human languages. Hence, universal grammar (UG) must be rich and specific enough to enable each child to get over the learning hump, but flexible enough to enable him or her to acquire different grammars in different linguistic contexts [17–19,57]. **(d)** In the case of moral development, it remains unclear whether models that incorporate parametric variation will likewise enter into the best explanation of universal moral grammar (UMG), the innate function or acquisition device that maps the child's early experience onto the mature state of his or her moral competence. What one needs are answers to two questions: (i) what are the properties of the moral grammars that people do in fact acquire, and (ii) how diverse are they? Although it is plausible to suppose that some aspects of moral judgment are innate, it seems clear that cultural elements also have a dramatic influence [14,15,55,58,59].

difficult, if not impossible, to construct a descriptively adequate theory of these intuitions – and others like them in a potentially infinite series – based exclusively on the information given. Although each of these intuitions is triggered by an identifiable stimulus, how the mind goes about interpreting these novel fact patterns, and assigning a deontic status to the acts they depict, is not revealed in any obvious way by the scenarios themselves. Instead, an intervening step must be postulated: a pattern of organization that is imposed on the stimulus by the mind itself. Hence, a simple perceptual model, such as the one that is implicit in Haidt's social intuitionist model of moral judgment, seems inadequate for explaining these intuitions (see, e.g., the unanalyzed link between eliciting situation and intuitive response in Figure 2 of Ref. [11]). Instead, as is the case with language perception, an adequate model must be more complex and must specify at least three elements: (i) the deontic rules that are 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 converted into an appropriate structural description [38].

## Descriptive adequacy

### *Deontic rules*

Trolley problems are what jurists call 'cases of necessity' [39], and they can be solved by assuming individuals are intuitive lawyers who possess a natural readiness to compute mental representations of human acts in legally cognizable terms. In particular, an indefinitely large class of such cases can be explained by postulating tacit knowledge of two specific legal rules: the prohibition of intentional battery and the principle of double effect. The prohibition of intentional battery forbids purposefully or knowingly causing harmful or offensive contact with another individual or otherwise invading another individual's physical integrity without his or her consent [39,40]. The principle of double effect is a complex principle of justification, narrower in scope than the traditional necessity defense, which holds that an otherwise prohibited action, such as battery, that 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 [35]. Both rules require clarification but, taken together and suitably formalized [6], they can be invoked to explain the relevant pattern of intuitions in a relatively straightforward manner. The key distinction that explains the standard cases in the literature is that the agent commits one or more distinct batteries prior to and as a means of achieving his good end in the impermissible conditions (*Transplant* and *Footbridge*), whereas these violations are subsequent side effects in the permissible conditions (*Trolley* and *Bystander*) (Box 3).

### *Structural descriptions*

The moral grammar hypothesis holds that when people encounter trolley problems, they unconsciously compute structural descriptions such as those in Figure Id of Box 3. Note that in addition to explaining the relevant intuitions, this hypothesis has further testable implications. For

**Box 3. Trolley problems and their implications**

Trolley problems imply that moral judgments do not depend solely on the consequences or superficial description of an action but also on how that action is mentally represented (Figure Ia in this box). Hence, the problem of descriptive adequacy in the theory of moral cognition can be divided into at least three parts, involving the description of (i) deontic rules, (ii) structural descriptions, and (iii) conversion rules [36] (Figure Ib). Although the difficulty that individuals have explaining their judgments [5,36,37] suggests that they are unaware of the principles that guide their moral intuitions (Figure Ic), the judgments can be explained by assuming that these individuals are intuitive lawyers who implicitly recognize the relevance of ends, means, side effects and *prima facie* wrongs, such as battery, to the analysis of legal and moral problems. For example, the *Transplant* and *Trolley* findings can be partly explained in terms of the distinction between battery as a means and battery as a side effect [41]. The structural descriptions that are implied by this explanation can be exhibited in a two-dimensional tree diagram, or act tree, successive nodes of which bear a generation relation to one another that is asymmetric, irreflexive and transitive [60] (Figure Id).

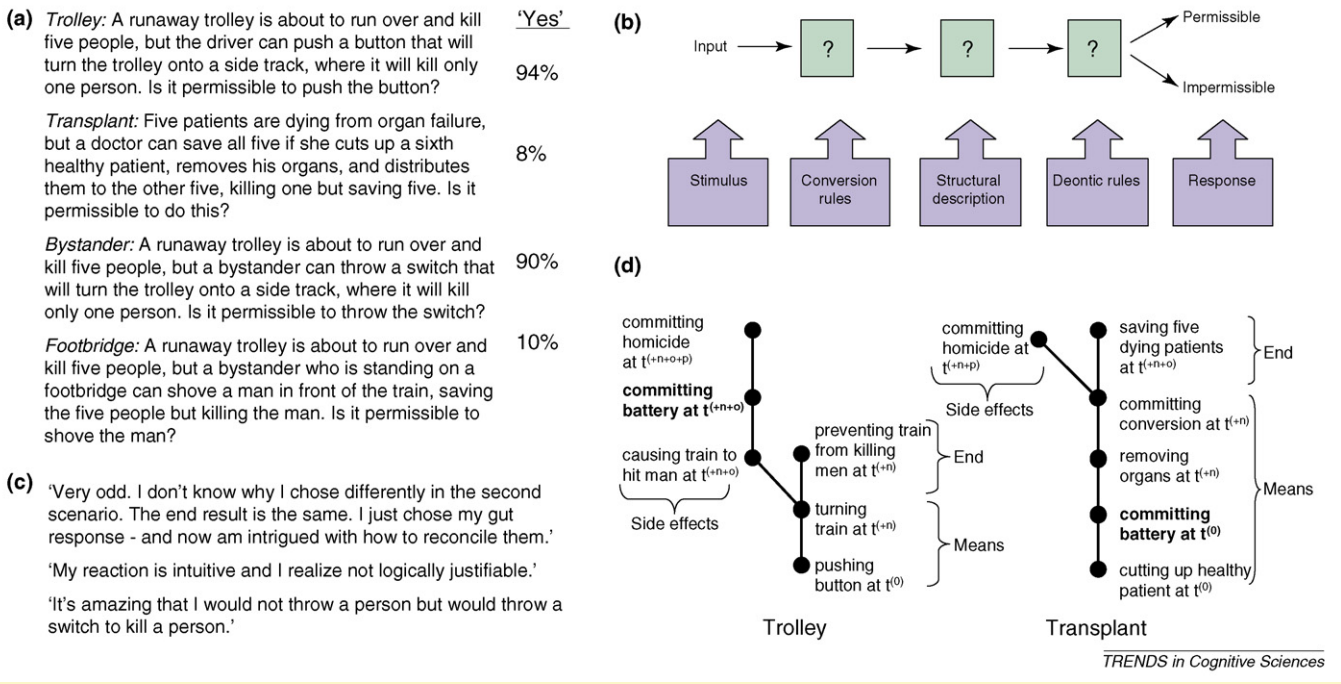


**Figure I**. Trolley problems and moral judgments **(a)**, expanded perceptual model **(b)**, judgment explanations **(c)** and structural descriptions **(d)**. Data in (a) and explanations in (c) from Ref. [6].

example, the structural properties of these representations can be investigated by asking subjects to evaluate probative descriptions of the relevant actions. Descriptions that use the word 'by' to connect individual nodes of act trees in the downward direction (e.g. 'D turned the train by throwing the switch' or 'D killed the man by turning the train') will be deemed acceptable; by contrast, causal reversals that use 'by' to connect nodes in the upward direction (e.g. 'D threw the switch by turning the train' or 'D turned the train by killing the man') will be deemed unacceptable. Likewise, descriptions that use connectors like 'in order to' or 'for the purpose of' to link nodes in the upward direction along the vertical chain of means and ends (e.g. 'D threw the switch in order to turn the train') will be deemed acceptable. By contrast, descriptions of this type that link means with side effects (e.g. 'D threw the switch in order to kill the man') will be deemed unacceptable. In short, there is an implicit geometry to these representations, which many neo-emotivist theories of moral cognition [8–15] neglect, but which an adequate theory must account for [41].

### *Conversion rules*

The main theoretical problem that is implied by the foregoing account is how people manage to compute a full structural description of the relevant action 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. This is a distinct poverty of the stimulus problem [42], similar in principle to determining how people manage to recover a three-dimensional representation from a two-dimensional stimulus in the theory of vision [22], or to determining how people recognize word boundaries in unmarked auditory patterns in the theory of language [20]. Figure 2 depicts how these properties can be recovered from the stimulus by a sequence of operations that are largely mechanical. These operations include (i) identifying the various action descriptions in the stimulus, (ii) placing them in an appropriate temporal order, (iii) decomposing them into their underlying causative and semantic structures, (iv) applying certain moral and logical principles to these underlying structures to generate representations of good and bad effects, (v) computing the intentional structure of the relevant acts and omissions by inferring (in the absence of conflicting evidence) that agents intend good effects and avoid bad ones, and (vi) deriving representations of morally salient acts like battery and situating them in the correct location of one's act tree [38]. Although each of these operations is relatively

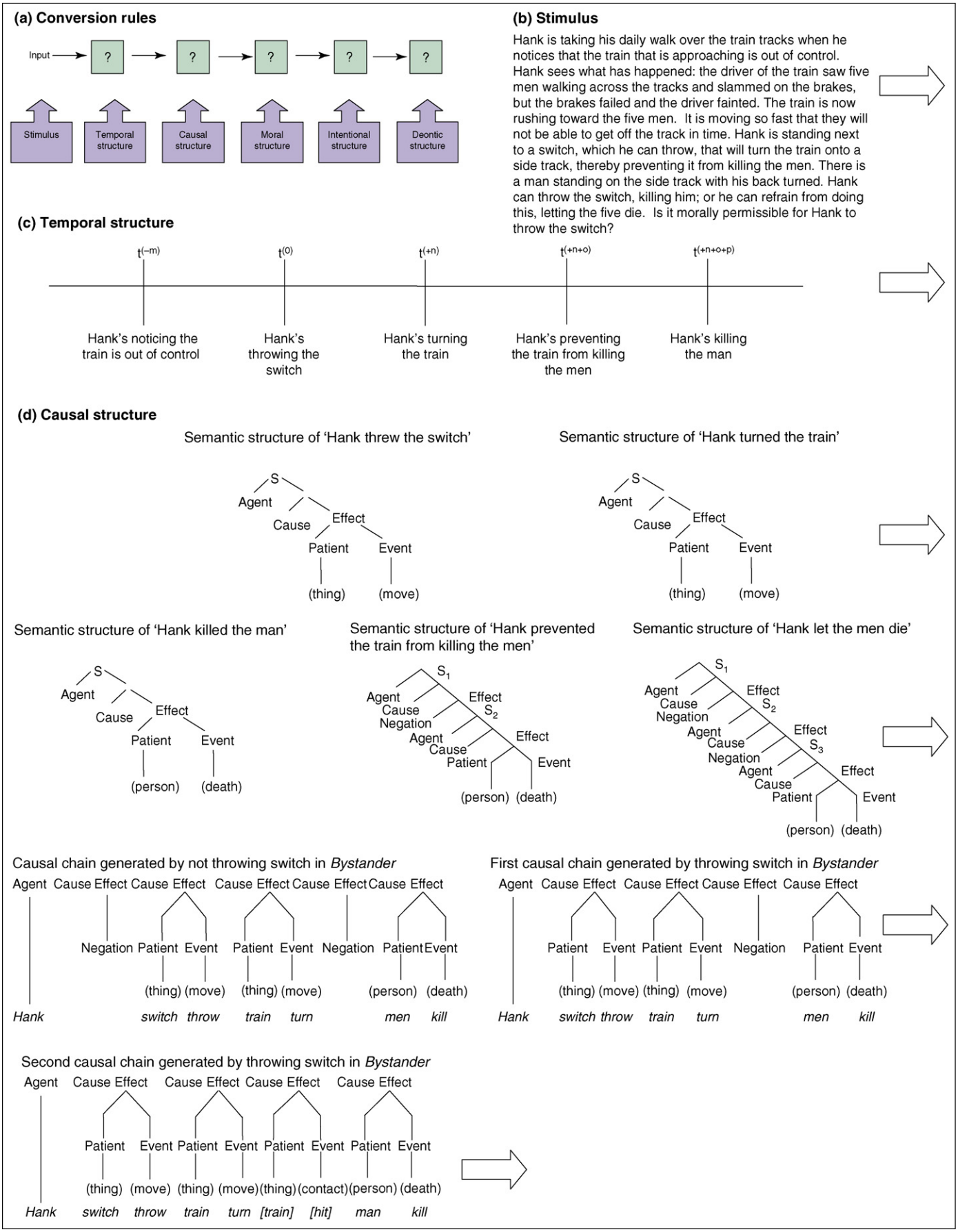
(a) Conversion rules
Input
?
?
?
?
?
Stimulus
Temporal structure
Causal structure
Moral structure
Intentional structure
Deontic structure
(b) Stimulus
Hank is taking his daily walk over the train tracks when he notices that the train that is approaching is out of control. Hank sees what has happened: 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. Is it morally permissible for Hank to throw the switch?
(c) Temporal structure
t(−m)
t(0)
t(+n)
t(+n+o)
t(+n+o+p)
Hank's noticing the train is out of control
Hank's throwing the switch
Hank's turning the train
Hank's preventing the train from killing the men
Hank's killing the man
(d) Causal structure
Semantic structure of 'Hank threw the switch'
S
Agent
Cause
Effect
Patient
Event
(thing)
(move)
Semantic structure of 'Hank turned the train'
S
Agent
Cause
Effect
Patient
Event
(thing)
(move)
Semantic structure of 'Hank killed the man'
S
Agent
Cause
Effect
Patient
Event
(person)
(death)
Semantic structure of 'Hank prevented the train from killing the men'
S1
Agent
Cause
Negation
Effect
S2
Agent
Cause
Patient
Effect
Event
(person)
(death)
Semantic structure of 'Hank let the men die'
S1
Agent
Cause
Negation
Effect
S2
Agent
Cause
Negation
Effect
S3
Agent
Cause
Patient
Effect
Event
(person)
(death)
Causal chain generated by not throwing switch in Bystander
Agent Cause Effect Cause Effect Cause Effect Cause Effect Cause Effect
Negation Patient Event Patient Event Negation Patient Event
(thing) (move) (thing) (move) (person) (death)
Hank switch throw train turn men kill
First causal chain generated by throwing switch in Bystander
Agent Cause Effect Cause Effect Cause Effect Cause Effect
Patient Event Patient Event Negation Patient Event
(thing) (move) (thing) (move) (person) (death)
Hank switch throw train turn men kill
Second causal chain generated by throwing switch in Bystander
Agent Cause Effect Cause Effect Cause Effect Cause Effect
Patient Event Patient Event Patient Event Patient Event
(thing) (move) (thing) (move) (thing) (contact) (person) (death)
Hank switch throw train turn [train] [hit] man kill

**Figure 2**

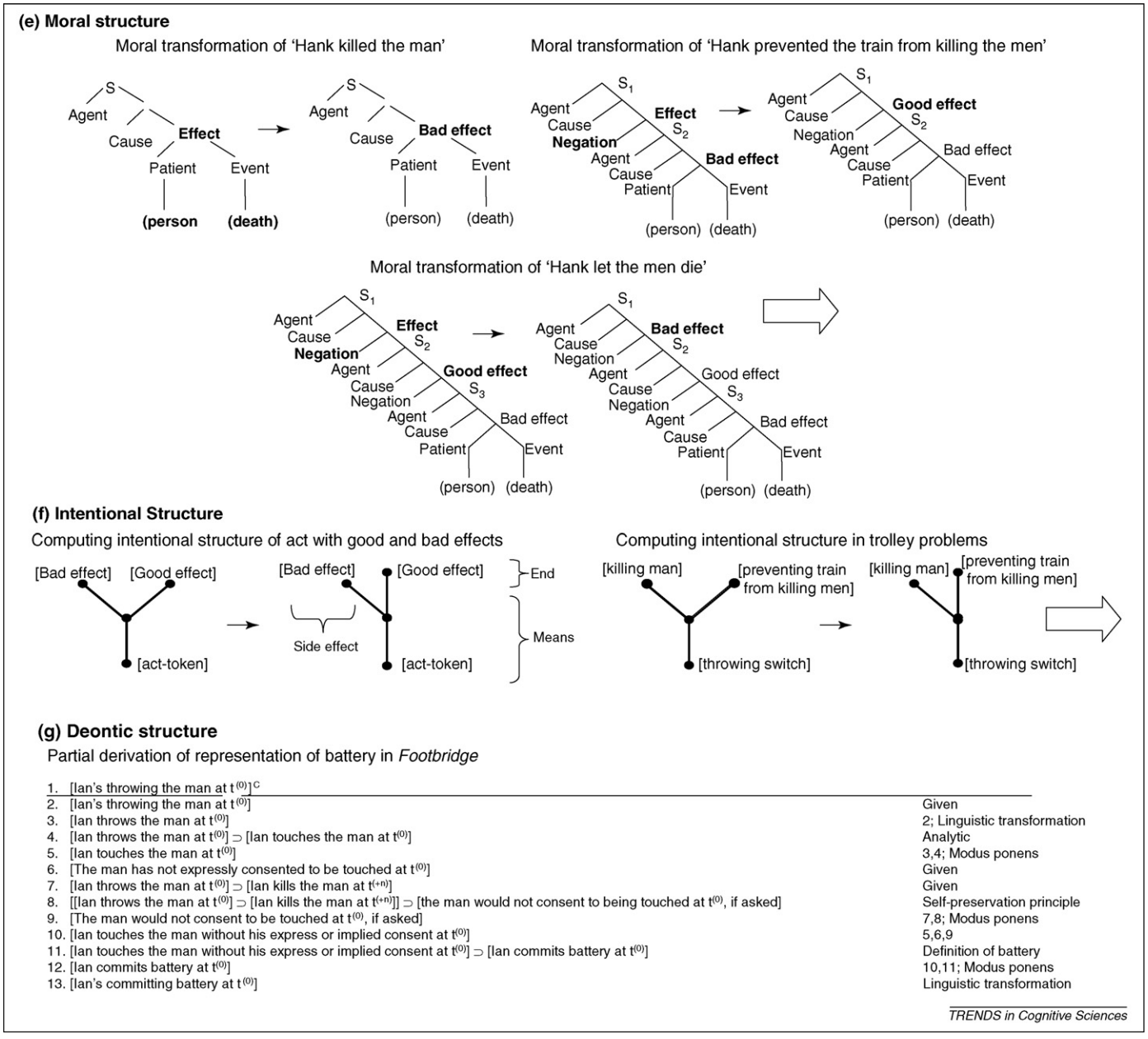


**Figure 2**. (This and opposite page.) Computing structural descriptions. **(a)** To compute an accurate structural description of a given action, the systems that support moral cognition must generate a complex representation of the action that encodes pertinent information about its temporal, causal, moral, intentional and deontic properties. Although the order of these computations might seem optional, I assume here that they occur in the following order (as illustrated by the version of *Bystander* depicted here and again in Box 4). **(b)–(c)** First, one must identify the relevant action descriptions in the stimulus and order them serially according to their temporal properties. **(d)** Second, one must determine the causal structure of the relevant acts and omissions by (i) interpreting the relevant causative constructions (e.g. 'Hank threw the switch' or 'Hank killed the man') in terms of their underlying semantic structures, and (ii) combining those structures into ordered sequences of causes and effects ('causal chains'), supplying missing information where necessary, such as the bracketed effect of causing the train to hit the man, which is not represented in the stimulus. **(e)** Third, one must compute the moral structure of the relevant acts and omissions by applying the following rewrite rules to the causal structures in (d): (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 of these operations, these causal structures are transformed into richer representations that encode good and bad effects. **(f)** Fourth, one must determine the intentional structure of the relevant acts and omissions by assuming (absent conflicting evidence) that the agent's end or goal is to achieve the good effect, but not the bad effect. Note that some operation of this general type must be postulated to explain how the brain computes ends, means and side effects, because the stimulus itself contains no direct evidence of these properties. In the operation that is depicted in (f), the arrow is a rewrite rule that converts a representation of an act token with both good and bad effects into a more complex representation that encodes ends, means and side effects; the general rule is depicted on the left and its application to the trolley problems is depicted on the right. **(g)** Fifth, because the foregoing steps are necessary but not sufficient to support the relevant intuitions, one must supply additional moral (specifically, deontic) structure to these descriptions. For example, one must derive a representation of battery in *Footbridge* by inferring, on the basis of direct and circumstantial evidence, that (i) the agent must touch the man in order to throw him onto the track, and (ii) the man would not consent to being touched in this manner because of his desire for self-preservation (the 'self-preservation principle'). This derivation can be formalized using standard notation in action theory and deductive logic. Finally, one must locate this representation in the correct temporal, causal and intentional location in one's act tree, thereby identifying whether the battery is a means or a side effect [6].

simple in its own right, the overall length, complexity and abstract nature of these computations, along with their rapid, intuitive and at least partially inaccessible character [8–10,33,36,37], lends support to the hypothesis that they depend on innate, domain-specific algorithms. However, this argument is not conclusive [43–45], and further research is needed to clarify the relevant conceptual and evidentiary issues.

## Intuitive legal appraisal

An important alternative to the moral grammar hypothesis is defended by Greene and colleagues [8–10]. In their view, moral intuitions result from the complex interplay of at least two distinct processes: domain-specific, social–emotional responses that are inherited from our primate ancestors, and a uniquely human capacity for 'sophisticated abstract reasoning that can be applied to any subject matter' ([8], p. 519). However, while the authors' evolutionary rationale is compelling, their distinction between 'personal' and 'impersonal' harms seems far too crude to achieve descriptive adequacy. Ordinary legal casebooks – repositories of centuries of moral problems and the intuitions they elicit – are full of plausible counterexamples. By contrast, concepts like *battery*, *end*, *means* and *side effect* are computational formulas that have stood the test of time [30,46,47]. Not only can they predict human moral intuitions in a huge number and variety of cases, but they also can help to explain the variance one finds in unusual permutations of the trolley problem. (Box 4 sketches one such proposal, but both the data and hypothesis presented are preliminary.) Moreover, these concepts can be broken down into clear cognitive components, thereby providing links to other domains, such as theory of mind [48,49]. For example, our framework can be used to predict that individuals who have disorders such as autism or Asperger's syndrome might have difficulty distinguishing certain pairs of trolley problems, and to pinpoint the exact source of this difficulty [50]; likewise, the computations that are exhibited here may sharpen our understanding of a diverse range of neuropsychological phenomena, from psychopathy, sociopathy and various forms of brain damage [1,6] to the asymmetrical attribution of intentions underlying the so-called 'side-effect effect' [51]. Finally, Greene's conception of personal harm 'in terms of "me hurt you", and as delineating roughly those violations that a chimpanzee can appreciate' ([8], p. 519) seems to rest on the assumption that the psychological patterns that are associated with human deontological judgment are qualitatively similar to the thought processes of chimpanzees. Yet it seems clear that adequately specifying the kinds of harm that humans intuitively grasp requires a technical legal vocabulary [23–26,29–33,36–41,52–55],

### Box 4. Six trolley problems and their structural descriptions

**Initial circumstances**

[X] is taking his daily walk over the train tracks when he notices that the train that is approaching is out of control. [X] 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 towards the five men. It is moving so fast that they will not be able to get off the track in time.

**Six distinct fact patterns**

Differences across pairs are underlined.

(1a) ***Bystander***: 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?

(1b) ***Footbridge***: 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?

(2a) ***Loop track***: 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?

(2b) ***Man-in-front***: 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?

(3a) ***Drop man***: 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?

(3b) ***Collapse bridge***: 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 the 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?

**Preliminary findings**

Preliminary findings suggest that each pair yields different judgments, which in four of six conditions the personal–impersonal distinction (see below) cannot explain (Table I) [6].

'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' ([8], p. 519).

These results form a remarkably consistent pattern, with permissibility judgments increasing linearly across the six conditions (Figure Ia in this box, next page). The results can be tentatively explained by the properties of each act's structural description (Figure Ib). Acts appear more likely to be judged permissible in these circumstances as counts of battery that are committed as a means decrease from three (Ian) to two (Victor) to one (Ned), and as these violations become side effects (Oscar, Walter, Hank) and additional structural features come into play. In 'Oscar', whereas the agent's action plan (shown in Figure Ib) is to save the men by causing the train to hit the object but not the man, the actual result (not shown) is likely to involve hitting the man before hitting the object; hence, from an *ex post* perspective, the agent will commit a battery before and as a means of achieving his good end. Likewise, in 'Walter', one or more counts of battery must necessarily occur before the good end is achieved. By contrast, in 'Hank', battery is a side effect and occurs after the good end is achieved.

**Table I. Moral judgments in six trolley problem conditions**

| Case | Personal or impersonal | % Yes[a] |
|---|---|---|
| Hank | Impersonal | 90% |
| Ian | Personal | 10% |
| Ned | Impersonal | 48% |
| Oscar | Impersonal | 62% |
| Victor | Impersonal | 37% |
| Walter | Impersonal | 68% |

[a]Data from Ref. [6].

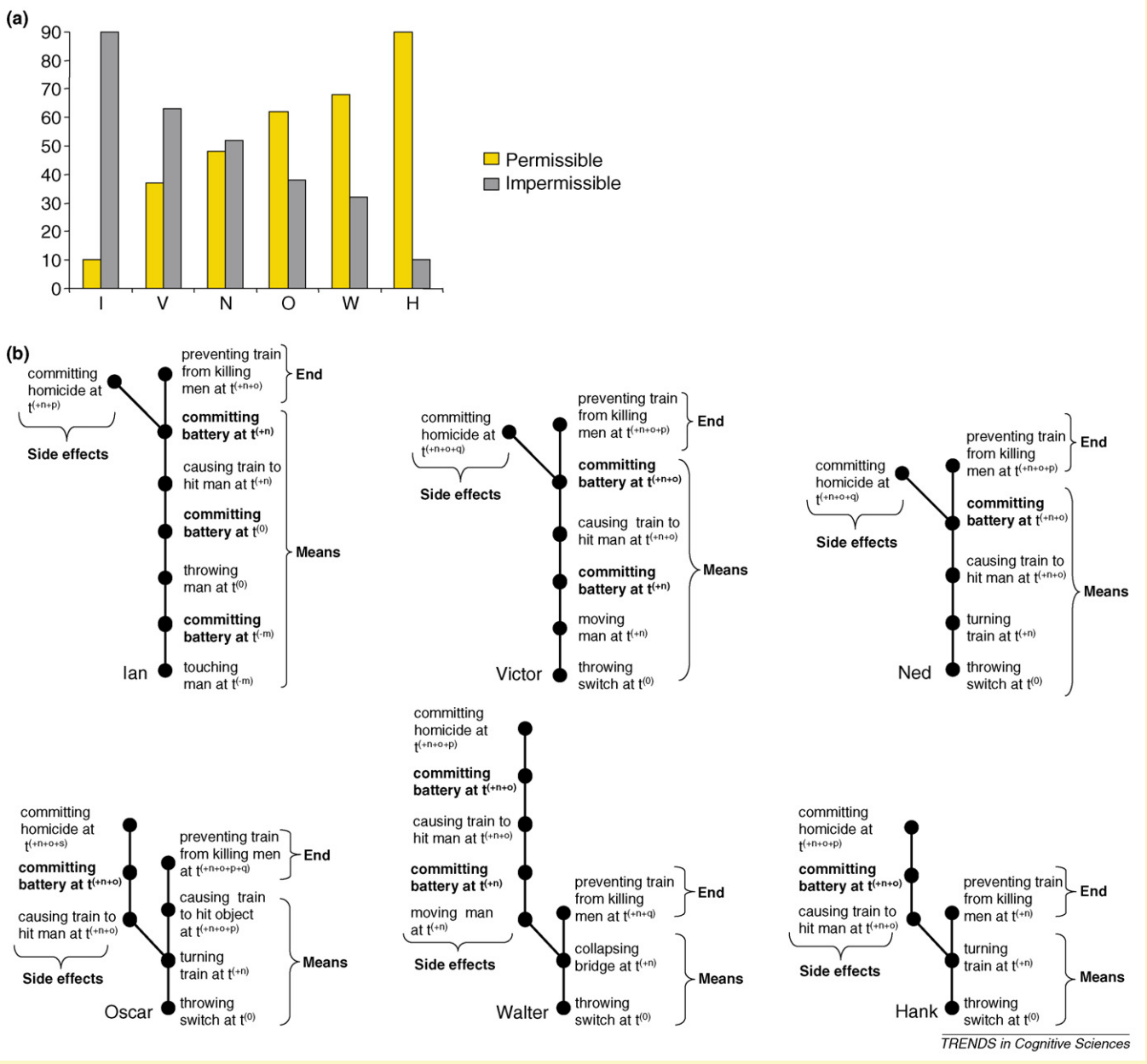


**Figure I (Box 4)**. Permissibility judgments **(a)** and structural descriptions **(b)** in six trolley problem conditions.

whereas the same is not true (or at least has not yet been shown) of our primate ancestors. The crucial issue is not whether moral intuitions are linked to emotions – clearly they are – but how to characterize the appraisal system that those intuitions presuppose and, in particular, whether that system incorporates elements of a sophisticated jurisprudence.

## Concluding remarks

Chomsky transformed linguistics and cognitive science by showing that ordinary language is susceptible to precise formal analysis and by rooting principles of UG in the human bioprogram. UMG holds out the prospect of doing the same for aspects of ordinary human moral cognition. The first step in the inquiry is to identify a class of considered judgments and a set of rules or principles from which they can be derived [7]. Initial efforts to explain trolley-problem intuitions within this framework suggest that individuals are intuitive lawyers who are capable of drawing intelligent distinctions between superficially similar cases, although their basis for doing so is often obscure. Future research on moral grammar should begin from this premise (Box 5), moving beyond the limited example of trolley problems and other doctrinally marginal 'dilemmas' to the core concepts of universal fields like torts, contracts and criminal law, which investigate the rules and representations that are implicit in common moral intuitions with unparalleled care and sophistication. Chomsky emphasized that rigorous formulation in linguistics is not merely a pointless technical exercise but an important diagnostic and heuristic tool, because only by pushing a precise but inadequate formulation to an unacceptable conclusion can one gain a better understanding of the relevant data and of the inadequacy of our existing

**Box 5. Questions for future research**

- How accurately do technical legal definitions of prohibited acts and valid defenses capture the structure of common moral intuitions?
- What mental representations are implied by common moral intuitions, and how does the brain recover these properties from the corresponding signal?
- What are the neural substrates and behavioral effects of legal concepts, such as the concurrence of act (actus reus) and mental state (mens rea), that link moral judgment with theory of mind?
- What are the moral grammars that children acquire and how diverse are they?
- What information is available in the child's environment with respect to this learning target?
- Is there a universal moral grammar and, if so, what are its properties?

attempts to explain them [56]. Likewise, Marr 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' ([22], p. 26). Cognitive scientists who take these ideas seriously and who seek to understand human moral cognition must devote more attention to developing computational theories of moral competence. Legal theory will have an important role in this process.

**Acknowledgements**

The author wishes to thank Noam Chomsky, Joshua Greene, Rebecca Saxe, Joshua Tenenbaum, the anonymous referees and the editor for helpful feedback on previous versions of this article. The article draws from two unpublished manuscripts: Rawls' linguistic analogy (J. Mikhail, PhD thesis, Cornell University, 2000) and Aspects of a theory of moral cognition (J. Mikhail, JD thesis, Stanford Law School, 2002).

**References**

1 Hauser, M.D. (2006) *Moral Minds: How Nature Designed our Universal Sense of Right and Wrong,* Harper Collins
2 Harman, G. (2000) *Explaining Value and Other Essays in Moral Philosophy,* Oxford University Press
3 Dwyer, S. (1999) Moral competence. In *Philosophy and Linguistics* (Stainton, R., ed.), pp. 169–190, Westview Press
4 Mahlmann, M. (1999) *Rationalismus in der Praktischen Theorie: Normentheorie und Praktiche Kompetenz,* Nomos Verlagsgesellschaft
5 Mikhail, J. *et al.* (1998) Toward a universal moral grammar. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (Gernsbacher, M.A. and Derry, S.J., eds), p. 1250, Lawrence Erlbaum Associates.
6 Mikhail, J. *Rawls' Linguistic Analogy*, Cambridge University Press (in press)
7 Rawls, J. (1971) *A Theory of Justice,* Harvard University Press
8 Greene, J. and Haidt, J. (2002) How (and where) does moral judgment work? *Trends Cogn. Sci.* 6, 517–523
9 Greene, J.D. *et al.* (2001) An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108
10 Greene, J. (2004) Cognitive neuroscience and the structure of the moral mind. In *The Innate Mind: Structure and Contents* (Carruthers, P. *et al.*, eds), pp. 338–352, Oxford University Press
11 Haidt, J. (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychol. Rev.* 108, 814–834
12 Moll, J. *et al.* (2005) The neural basis of human moral cognition. *Nat. Rev. Neurosci.* 6, 799–809
13 Sunstein, C.R. (2005) Moral heuristics. *Behav. Brain Sci.* 28, 531–573
14 Nichols, S. (2004) *Sentimental Rules: On the Natural Foundations of Moral Judgment,* Oxford University Press
15 Prinz, J. (2007) *The Emotional Construction of Morals,* Oxford University Press
16 Kohlberg, L. (1981) *Essays on Moral Development, Vol. 1: The Philosophy of Moral Development,* Harper and Row
17 Chomsky, N. (1964) *Current Issues in Linguistic Theory,* Mouton
18 Chomsky, N. (1986) *Knowledge of Language: Its Nature, Origin, and Use,* Praeger
19 Chomsky, N. (1995) *The Minimalist Program,* MIT Press
20 Jackendoff, R. (1994) *Patterns in the Mind: Language and Human Nature,* Basic Books
21 Gallistel, C.R. (1999) The replacement of general-purpose learning models with adaptively specialized learning modules, In *The Cognitive Neurosciences* (2nd edn) (Gazzaniga, M., ed.), pp. 1179–1191, Cambridge University Press
22 Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information,* W.H. Freeman
23 Nelson, S. (1980) Factors influencing young children's use of motives and outcomes as moral criteria. *Child Dev.* 51, 823–829
24 Smetana, J. (1983) Social cognitive development: domain distinctions and coordinations. *Dev. Rev.* 3, 131–147
25 Finkel, N. *et al.* (1997) Equal or proportional justice for accessories? Children's pearls of proportionate wisdom. *J. Appl. Dev. Psychol.* 18, 229–244
26 Chandler, M.J. *et al.* (2000) Beliefs about truth and beliefs about rightness. *Child Dev.* 71, 91–97
27 Bybee, J. and Fleischman, S., eds (1995) *Modality in Grammar and Discourse*, John Benjamins
28 Von Wright, G.H. (1951) Deontic logic. *Mind* 60, 1–15
29 Brown, D. (1991) *Human Universals,* McGraw-Hill
30 Mikhail, J. (2002) Law, science, and morality: a review of Richard Posner's 'The Problematics of Moral and Legal Theory'. *Stanford Law Rev.* 54, 1057–1127
31 Fletcher, G. (1998) *Basic Concepts of Criminal Law,* Oxford University Press
32 Green, S.P. (1998) The universal grammar of criminal law. *Mich. Law Rev.* 98, 2104–2125
33 Mikhail, J. The poverty of the moral stimulus. In *Moral Psychology, Vol. 1: Innateness and Adaptation* (Sinnott-Armstrong, W., ed.), MIT Press (in press)
34 Kamm, F. (1998) *Morality, Mortality, Vol. II: Rights, Duties, and Status,* Oxford University Press
35 Fischer, J.M. and Ravizza, M. (1992) *Ethics: Problems and Principles,* Harcourt Brace Jovanovich
36 Hauser, M. *et al.* (2007) A dissociation between moral judgments and justifications. *Mind Lang.* 22, 1–21
37 Cushman, F. *et al.* (2006) The role of conscious reasoning and intuition in moral judgments: testing three principles of harm. *Psychol. Sci.* 17, 1082–1089
38 Mikhail, J. Moral cognition and computational theory. In *Moral Psychology, Vol. 3: The Neuroscience of Morality* (Sinnott-Armstrong, W., ed.), MIT Press (in press)
39 Epstein, R. (2004) *Cases and Materials on Torts,* (8th edn), Aspen
40 Shapo, M.S. (2003) *Principles of Tort Law,* Thomson-West
41 Mikhail, J. (2005) Moral heuristics or moral competence? Reflections on Sunstein. *Behav. Brain Sci.* 28, 557–558
42 Fodor, J. (1985) Precis of the modularity of mind. *Behav. Brain Sci.* 8, 1–42
43 Nichols, S. (2005) Innateness and moral psychology. In *The Innate Mind: Structure and Contents* (Carruthers, P. *et al.*, eds), pp. 353–369, Oxford University Press
44 Sripada, C.S. and Stich, S. (2006) A framework for the psychology of norms. In *The Innate Mind: Culture and Cognition* (Carruthers, P. *et al.*, eds), pp. 280–301, Oxford University Press
45 Prinz, J. Is morality innate? In *Moral Psychology, Vol. 1: Innateness and Adaptation* (Sinnnot-Armstrong, W., ed.), MIT Press (in press)
46 Barnes, J. (1984) *The Complete Works of Aristotle,* Princeton University Press
47 Watson, A. (1985) *The Digest of Justinian,* University of Pennsylvania Press
48 Saxe, R. *et al.* (2004) Understanding other minds: linking developmental psychology and functional neuroimaging. *Annu. Rev. Psychol.* 55, 87–124
49 Knobe, J. (2005) Theory of mind and moral cognition: exploring the connections. *Trends Cogn. Sci.* 9, 357–359

50 Hauser, M. *et al*. Reviving Rawls' linguistic analogy: operative principles and the causal structure of moral actions. In *Moral Psychology, Vol. 2* (Sinnott-Armstrong, W., ed.), MIT Press (in press)
51 Leslie, A.M. *et al*. (2006) Acting intentionally and the side-effect effect: 'theory of mind' and moral judgment. *Psychol. Sci.* 17, 421–427
52 Mackie, J.L. (1977) *Ethics: Inventing Right and Wrong,* Penguin Books
53 Harman, G. (1978) *The Nature of Morality: An Introduction to Ethics,* Oxford University Press
54 Robinson, P.M. and Darley, J.M. (1995) *Justice, Liability, and Blame: Community Views and the Criminal Law,* Westview Press
55 Nader, L. (1997) *Law in Culture and Society,* University of California Press
56 Chomsky, N. (1957) *Syntactic Structures,* Mouton
57 Baker, M. (2001) *The Atoms of Language: The Mind's Hidden Rules of Grammar,* Basic Books
58 Nisbett, R. and Cohen, D. (1996) *Culture of Honor: The Psychology of Violence in the South,* Harper Collins
59 Shweder, R. *et al*. (1987) Culture and moral development. In *The Emergence of Morality in Young Children* (Kagan, J. and Lamb, S., eds), pp. 1–83, University of Chicago Press
60 Goldman, A. (1970) *A Theory of Human Action,* Princeton University Press