

# Notre Dame Philosophical Reviews

---

2012.02.04

Author

**John Mikhail**

## Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment

Published: February 01, 2012

**John Mikhail, *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*, Cambridge University Press, 2011, 406pp., \$90.00 (hbk), ISBN 9780521855785**

**Reviewed by Mark Phelan, Lawrence University**

In *Elements of Moral Cognition*, John Mikhail clarifies and attempts to vindicate John Rawls' linguistic analogy, according to which moral cognition is usefully modeled on Chomsky's account of linguistic cognition. In the first part of the book, Mikhail explicates key aspects of Chomsky's theory of language, shows how these have analogues in moral theory, and demonstrates Rawls' awareness of the isomorphism via key quotes from early works. In drawing out these analogies with linguistics, Mikhail suggests a new framework for moral theorizing. In the second part, Mikhail attempts to ground the empirical significance of this new framework by showing how it allows for a provisional description of "the mature individual's system of moral knowledge," and thereby explains a number of commonsense moral intuitions. In the third part, Mikhail shows how the new framework allows for forceful responses to some early criticisms of Rawls' linguistic analogy.

This book is both enlightening and frustrating. It is incredibly well-informed and, consequently, incredibly dense. Press reviews for the book -- from Noam Chomsky, Gilbert Harman, and Frans De Waal -- accurately note the conceptual depth, careful execution, and great erudition with which Mikhail's central claims are developed. The text includes twenty-three epigraphs and only ten chapters. I worry that many readers will find the book overwhelming and tedious, wonder whether defending an analogy Rawls drew early in his career is worth all the effort, and abandon it midway. To abandon or ignore this book on that basis would be a mistake. As Mikhail suggests, the defense is just a cover story (11). For the bulk of readers, the worth of this book will lie not in its destination but in some of the many stops it makes along the way. The explication of Chomsky's linguistic theory is (compared to many other presentations) concise, clear, and accurate, and also interlaced with careful analogues to moral theory. The provisional account of moral grammar is similarly insightful and honest -- even if (as I'll explain) ultimately unsatisfying. These aspects of the book make it worth its admittedly steep price of admission (both monetary and intellectual), at least for scholars working in related areas. Here I will discuss the first two parts of Mikhail's book, leaving an analysis of the third section, containing responses to Rawls' critics, to those more informed about the historical debate.

The first part of *Elements of Moral Cognition* consists of three chapters, comprising an introduction and outline of the book, an elucidation of key concepts in Chomsky's linguistic theory, and a careful analysis of section nine of a *Theory of Justice*, demonstrating how Rawls modeled his theory of moral cognition on Chomsky's theory of Universal Grammar. Mikhail does a laudable job of clarifying key Chomskian terminology, which may be foreign to many ethicists. Essentially, Mikhail claims that Rawls "conceives of the subject matter of moral theory . . . as a mental object that is internal, intensional, individual, and ideal" (64). Thus, Rawls holds that this internal, mental object -- this moral sense -- is an *I-morality*, to be distinguished from an *E-morality* which is, "external to and independent of the mind/brain" (64). According to Mikhail, "Rawls' early writings contain the germs of a scientific theory of moral cognition that far surpasses the work of psychologists like Jean Piaget (1965/1932) and Lawrence Kohlberg (1981, 1984) in terms of depth, coherence, and analytical rigor" (11). Perhaps this will constitute the most controversial theme for contemporary ethicists -- that Rawls' early work, including *A Theory of Justice*, is primarily moral psychology, not normative ethics. (In part three, Mikhail's responses to challenges to Rawls' work also rest on this reconceptualization.)

Much of *Elements of Moral Cognition* is devoted to providing a provisional description of the moral sense, and, *eo ipso*, an answer to one question (an analogue to a fundamental question in Chomsky's framework): "What constitutes moral knowledge?" (15) According to Mikhail, for reasons which become clearer throughout the first and second parts of the book,

it seems reasonable to suppose . . . that the normal individual's moral knowledge consists in part in her possession of what I will call a *moral grammar*, a complex and largely unconscious system of moral rules, concepts, and principles that generates and relates mental representations of various types (16).

The reasons for this supposition rest on analogies between our linguistic and moral capacities. The key reason for supposing that a language user has a complex and largely unconscious system of grammatical rules, concepts, and principles that generates and relates linguistic representations rests on the *argument for linguistic grammar*. This argument starts with the observation that competent users of a natural language can construct an infinitely large set of novel utterances that are immediately acceptable to other speakers of the language. At the same time, such competent language users can assess whether an unbounded set of novel utterances produced by others are well- or ill-formed. Language users, possessed of such infinite capacities, are nonetheless burdened with finite minds. Thus, the argument runs, each language user's mind "must contain a recipe or program of some sort -- a grammar -- that can build, out of a finite list of words and phrase patterns, the unlimited set of expressions she is able to produce, understand, and interpret" (46).

Mikhail suggests that a perfectly analogous argument (alluded to by Rawls in a single sentence in *A Theory of Justice*) leads us to suppose a moral grammar.<sup>[1]</sup> After all, an individual possessed of an adequate moral sense is able to judge a limitless set of novel acts permissible or impermissible. On the assumption that such moral judges are not possessed of infinite minds, we should presume a grammar for moral assessment.

To suppose that there is such a moral grammar is not to provide even a provisional description of it. If, as an I-morality allows, the moral sense differs for different people, one may legitimately wonder whether a generalized description is possible. This is why idealization is important. As Mikhail puts it, "Any serious attempt to delineate the structure of a cognitive system must seek reasonable ways to reduce the complexity of that system and to identify its fundamental properties" (54). According to Mikhail, Rawls follows Chomsky in supposing several kinds of intra- and inter-personal idealization. Rawls accepts the linguist's distinction between operative and express principles by supposing that the correct targets of moral theorizing, "are those principles that are actually operative in . . . [the] exercise of moral judgment," not the ethical principles individuals avow, which cannot be assumed to explain actual judgments (19). Rawls' approach mirrors the linguist's performance-competence distinction in positing that, in theorizing about moral grammar, we should not rely on all actual judgments of moral permissibility, which may often be distorted by factors external to the moral cognitive system. Rather we should rely only on considered moral judgments, which are supposed to be reflective of moral competency. Finally, Rawls avoids the problem of moral diversity by supposing that the principles of moral sense are "either approximately the same for persons whose judgments are in reflective equilibrium, or if not, that their judgments divide along a few main lines represented by the family of traditional doctrines" (56, cf. Rawls 1971, p. 50). Thus, the method of reflective equilibrium is itself conceived as an important license for moral psychology. Ultimately Mikhail expresses optimism regarding the prospects for a universal moral grammar, claiming greater agreement in natural moral systems compared to natural languages (57). However, considerations of individual variation (to be introduced later), as opposed to intercultural variation, could be construed as threats to this claim.

In the second part of *Elements of Moral Cognition*, Mikhail turns from considerations of the practice and possibility of moral theorizing to a provisional description of moral grammar. The three chapters of this section comprise an explanation of why and in what sense ordinary moral assessments of various trolley cases constitute considered moral judgments, an initial specification of moral grammar based on these assessments, and a formalization of the mental operations this moral grammar presupposes.

Trolley problems are simple vignettes in which an agent faces a binary choice between two non-ideal options. In a typical case an agent must decide whether to perform an action (e.g., throw a switch) that will result in some negative effect (e.g., a trolley changes to a side track and kills an innocent person) in order to achieve some positive objective (e.g., five people on the main track are saved). Over several years, Mikhail and other researchers asked experimental participants to assess whether it was

morally permissible for the agents in such cases to perform the relevant actions. Mikhail argues that participants' responses to these prompts constitute considered judgments in virtue of their sharing several key features with the grammaticality judgments that form the basis for linguistic theorizing. Assessments of trolley cases are "immediate and spontaneous," yet also, "stable, stringent, and highly predictable" (83). They are made with a high degree of certitude, "often experienced as being 'objective' rather than a matter of subjective taste or preference" (83). Yet, they are also intuitive, in that, "they are not made by a conscious application of moral rules or principles" (82). Indeed, as participants' justifications of responses to trolley problems reveal, moral assessments of trolley cases do not conform to the moral principles experimental participants avow.

Mikhail emphasizes the impossibility of explaining the pattern of permissibility assessments for trolley cases simply on the basis of information supplied in the vignettes. He identifies three elements of moral psychology that a satisfactory account of moral grammar must specify: deontic rules for moral judgment, a lexicon of structural descriptions, and conversion rules that transform verbal or visual stimuli into structural descriptions in moral grammar's proprietary lexicon. Mikhail posits several tacitly accepted deontic rules, including prohibitions of intentional battery and homicide, the Rescue Principle, and the Principle of Double Effect. The application of these deontic rules presupposes a structural lexicon that includes mental state descriptions. For example, the prohibition on intentional battery requires an assessment of the intentional states of the agent performing the battery. Such descriptions are seldom explicitly supplied in the case of verbal or visual stimuli, but their absence does not bar moral assessment. We must generate assessments of an agent's mental states from data supplied by the story or scene. Mikhail suggests provisional decision procedures for this purpose.

With these elements of moral cognition in place, Mikhail offers his description of how our moral sense effects permissibility judgments. We make such judgments by unconsciously working through a number of steps. First we construct a serial temporal structure for the events constituting the given stimulus. We then identify the causal structure of these events, using temporal information and other background knowledge to construct cause-effect chains for relevant acts and omissions. Next, "these causal structures are transformed into richer representations that encode good and bad effects," using specific rewrite rules (172). Then an overall intentional structure must be imposed on the representation, based on a presumption of innocence, according to which, "unless contrary evidence is given or implied, one should assume that S is a person of good will who pursues good and avoids evil" (173). A deontic structure is then imposed on the stimulus by identifying different events as batteries, homicides and other legal wrongs, and then, on the basis of information supplied in previous steps, as means, ends, or side effects. "Finally, once accurate structural descriptions of a given act-token representation and its alternatives . . . are generated, the correct deontic rules must be applied to these descriptions to yield a considered judgment" (173).

In the case of Ian's pushing a large man off a footbridge to slow an oncoming trolley in order to save five workers, these transformations generate a structural description containing three bad instances of intentional battery (respectively caused by intentional instances of touching a man, throwing a man off a bridge, and causing a train to hit a man) committed as a means to the good end of saving the five workers' lives. Thus, Ian's action is impermissible according to the prohibition on intentional battery, as experimental participants overwhelmingly agree. In the case of Hank, a bystander who saves five workers by throwing a switch and directing a train onto a sidetrack where it kills one man, these transformations generate a structural description containing the two bad side effects of committing a battery and a homicide, along with a pair of neutral means (throwing a switch, turning a train) and the good end of saving five workers. Thus, Hank's action is permissible according to the principle of double effect, which allows otherwise prohibited acts provided they are "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" (149). Again, experimental participants overwhelmingly confirm this verdict. Thus, the provisional description of moral grammar successfully explains these (and other) moral assessments.

*Elements of Moral Cognition* is full of exciting ideas. Nonetheless, I think it has several important shortcomings. One is its failure to directly address certain dissimilarities between permissibility assessments on the one hand and the grammaticality assessments that form the basis for linguistic theorizing on the other. Typically, the raw data that form the bases for theory construction in linguistics are the linguists' own grammaticality assessments for sentences containing the linguistic expressions of interest. This is a reasonable approach in linguistics because grammaticality assessments within a natural language are commonly shared. In fact, when asked in the context of an experimental study, ordinary speakers generally overwhelmingly agree with the grammaticality judgments of both one another and the linguists. Thus, there is little variance

within assessments of grammaticality for speakers of a natural language. However, as Mikhail's discussion of the trolley cases makes clear, this is not the case with permissibility judgments. For many trolley cases, only a narrow majority of experimental participants claim that the agent's action is permissible as opposed to impermissible (or vice-versa). Why is the situation so different with permissibility judgments as opposed to grammaticality judgments, if the systems that explain these judgments are purported to be so similar? This is an important question that Mikhail does not take up in his book (though he has discussed it in other papers).

In light of this omission, it is troubling that -- since the formulation of principles for moral grammar requires determinate permissibility judgments -- Mikhail takes the dubious step of deciding to "focus on the modal responses themselves and make the simplifying assumption that these judgments are considered judgments in Rawls' sense" (110). This is particularly suspect in cases for which only a very small majority of participants favored one response over the other. Without a more detailed explanation of individual variance, we can't conclude that the slightly more common response is the considered one. However, treating the slightly more common response as the considered one allows for considered judgments that conform very well to Mikhail's account of moral grammar.

For example, in one case participants are asked to consider Ned, who sees an out of control train careening towards five workers. Ned can throw a switch to temporarily direct the train onto a sidetrack where a heavy man is trapped. If the train hits the heavy man, it will kill him, but it will also slow the train enough to give the five workers time to escape the track before the train rejoins it. If Ned does not throw the switch, the five workers will not have time to escape the track and they will be killed. Here a scant majority of participants -- 52% -- think it *impermissible* for Ned to throw the switch.<sup>[2]</sup> In another case, Oscar faces a similar choice. Oscar can throw a switch to temporarily direct the train onto a sidetrack where a man is trapped in front of a large object. Temporarily changing the train's course to the sidetrack will kill the man, but it will also cause the train to hit the large object, thus slowing it enough to give the five workers time to escape. Here 62% of participants think it *permissible* for Oscar to throw the switch. Given his emphasis on the mode, Mikhail treats the slightly more common *impermissible* response as the considered judgment of the Ned case and the slightly more common *permissible* response as the considered judgment of the Oscar case, though the patterns of responses for these cases are very similar to one another and dissimilar from other cases assigned the same considered judgment. What's more, the difference between modal-*cum*-considered judgments for these two cases does important work for Mikhail, since treating Ned's action as impermissible and Oscar's as permissible aligns with Mikhail's emphasis on the principle of double effect. However, absent a full response to the variability issue, it is unclear whether we should interpret the pattern of responses as Mikhail does.

Another shortcoming of the book is that the formalizations for some of the mental operations presupposed by the moral grammar, though admittedly provisional, are not even provisionally acceptable for purposes of assessing the proposed moral grammar's capacity to explain its initial data: permissibility assessments for the twelve trolley cases. For example, moral grammar requires us to assess when an agent knowingly generates one event by means of another. Mikhail posits a provisional definition for such assessments, according to which for any two acts, S V-ing and S U-ing, S V-ing will knowingly generate S U-ing if and only if the two acts are not equivalent, the first act occurs before the second, and the first act results in the second (130). Of course, this definition would not be acceptable in numerous everyday cases, as when Sam breaks the neighbor's window by hitting a tee-ball for the first time ever. Sam's breaking the window by hitting the ball fits the provisional definition. Nonetheless, we won't think Sam's hitting the ball results in Sam's knowingly breaking the window, because we all recognize that Sam had no idea he could break the window by hitting the ball. Mikhail avoids this difficulty by simply assuming that each agent in his trolley vignettes, "both causes and knows the stipulated effects of his actions" (132).

But what is at issue is how ordinary moral judges understand these effects. To suppose that experimental participants always think the agents in the vignettes know the relevant effects of their actions casts the data in a favorable light for Mikhail's thesis, but in some cases the supposition is unwarranted. Whether or not ordinary moral judges assume that an agent knew about the effects of his actions is important to whether such judges decide that the agent knowingly generated the effects of his actions, which is important to whether such judges decide the agent intentionally generated the effects of his actions, which, in the case of intentional battery for example, is important to whether such judges decide the agent's action is permissible. If we simply concede that for all of his trolley cases all experimental participants assume that the agent both causes and knows the relevant effects of his action, Mikhail's model fits the data very well. On that assumption, Ian's pushing a large man off a footbridge to

slow an oncoming trolley will be seen as containing three bad instances of intentional battery, three separate effects respectively caused by touching a man, throwing a man off a bridge, and causing a train to hit a man. Whereas Victor's throwing a switch to drop a large man off of a footbridge to slow a train and save five workers will be seen as containing only two bad instances of intentional battery, respectively caused by dropping a man off a bridge and causing a train to hit a man. This fits well with the pattern of 10% permissibility judgments for Ian compared to 37% for Victor -- fewer batteries could somehow make Victor's action permissible in the mind of many more participants. However, it is *prima facie* implausible to suppose that ordinary moral judges even tacitly accept that Ian knows that by touching the large man he is causing some other effect, namely, the battery of the man. If they do not accept that, then Ian commits no more intentional batteries than Victor, and the asymmetry between these cases remains a mystery for moral grammar. To accurately assess how well the moral grammar hypothesis fits the data, we clearly need an account of how ordinary people assess known effects, among other things.

Finally, a substantial shortcoming of the book is its failure to take up any extended discussion of the question of explanatory adequacy: How is moral knowledge acquired? For Chomsky, the poverty of the stimulus argument leads to the conclusion that large aspects of linguistic grammar must be innate. But parallel reasoning regarding moral grammar is not assessed in Mikhail's book. Mikhail explains this omission, in part by claiming that the question of what our moral knowledge consists in is logically prior to the question of how moral knowledge is acquired (82). But such logical priority -- as opposed to, say, methodological priority -- does not clearly obtain. In linguistics one can appreciate the poverty of the stimulus argument before one has a good sense of what linguistic knowledge consists in. And knowing how moral knowledge is acquired could potentially inform our study of what moral knowledge is, by allowing for a range of developmental studies.

These are real shortcomings of the book. But Mikhail is upfront about the fact that his is not a complete picture (178). He is exceedingly careful and clearly knows what else the program requires. I look forward to seeing his future work on the topic. Despite its limitations, readers will learn a lot from *Elements of Moral Cognition*.

#### *Works Cited*

Mikhail, John (2007), "Universal Moral Grammar: Theory, Evidence, and the Future," *Trends in Cognitive Sciences*, Vol. 11, No. 4 April 2007, pp. 143-152.

Rawls, John (1971/1999), *A Theory of Justice*, Cambridge, MA: Harvard University Press.

---

[1] The relevant sentence may leave some readers underwhelmed. It is the second of this pair from Rawls (1977/1999): "Clearly this moral capacity is extraordinarily complex. To see this it suffices to note the potentially infinite number and variety of judgments that we are prepared to make" (41).

[2] Distributions of responses are drawn from Mikhail (2007), since they are not all reported in the book.