Summer 5-18-2012

Against the Linguistic Analogy

Noel B. Martin
Georgia State University

AGAINST THE LINGUISTIC ANALOGY

by

NOEL B. MARTIN

Under the Direction of Eddy Nahmias

ABSTRACT

Recently it has been proposed that humans possess an innate, domain-specific moral faculty, and that this faculty might be fruitfully understood by drawing a close analogy with nativist theories in linguistics. This Linguistic Analogy (LA) hypothesizes that humans share a universal moral grammar. In this paper I argue that this conception is deeply flawed. After profiling a recent and appealing account of universal moral grammar, I suggest that recent empirical findings reveal a significant flaw, which takes the form of a dilemma: either there is something wrong with the moral grammar model because we do not actually possess the innate contents (rules, principles, and concepts) it says we have, or the moral grammar model is simply the wrong model of moral cognition. In light of this dilemma, I conclude we ought to be skeptical that the Linguistic Analogy can adequately serve as a general account of moral cognition.

INDEX WORDS: Moral cognition, Innateness, Linguistics, Computational theory, John Rawls, Noam Chomsky

AGAINST THE LINGUISTIC ANALOGY

by

NOEL B. MARTIN

A Thesis Submitted in Partial Fulfillment of Requirements for the Degree of

Master of Arts

in the College of Arts and Sciences

Georgia State University

2012

AGAINST THE LINGUISTIC ANALOGY

by

NOEL B. MARTIN

Committee Chair:	Eddy Nahmias
Committee:	Daniel Weiskopf
	Andrea Scarantino

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2012

To Sally

– Always

ACKNOWLEDGEMENTS

Thanks, first, to Eddy Nahmias –I am grateful for his guidance. Since the beginning, he has struck a fine (and rare) balance between criticism and support. My deep appreciation goes to Dan Weiskopf and Andrea Scarantino, who graciously agreed to serve on my committee, and for their valuable comments and discussions during this process. Thanks also to Michael Owren for modeling what a good empiricist looks like, whether or not his influence shows here. I owe a debt of gratitude to Neil van Leewen, and my colleagues Toby Amoss, Michael Johnson, Kathryn Joyce, and Anaïs Stenson, whose camaraderie and insights have all helped greatly with the ideas presented here. To my family, for their support throughout this process (and even before it), thanks is not enough.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

In concluding *The Theory of Moral Sentiments*, Adam Smith suggests that the goal of moral philosophy is to answer two questions:

> First, wherein does virtue consist? … And, secondly, by what power or faculty in the mind is it, that this character, whatever it be, is recommended to us? Or in other words, how and by what means does it come to pass, that the mind prefers one tenour of conduct to another, denominates the one right and the other wrong; considers the one as the object of approbation, honour, and reward, and the other of blame, censure, and punishment? (VII.i.2)

Smith's second question constitutes one of the core concerns of moral psychology: how and why do we humans make the moral judgments that we do? One way to answer such questions is to offer an account of the faculty or faculties that underlie moral cognition.

Moral anti-nativists argue that moral cognition is not innate. These anti-nativists propose that moral cognition relies on psychological mechanisms that are not specifically dedicated to making moral evaluations (e.g., Nichols, 2004; Sterelny, 2010). One might be an anti-nativist, then, by claiming that the ability to make moral judgments results from our emotional capacities (Prinz, 2007). Views of this kind imply that there are no moral rules or principles that humans possess merely in virtue of being human. Instead, all distinctively moral concepts, rules, and principles are learned and employed via cognitive systems that also perform (and evolved to perform) other functions.

Moral nativists, on the other hand, have suggested that our capacities to make moral judgments are subserved by innate psychological mechanisms dedicated to, or evolved for, moral cognition, and that the capacities these mechanisms support cannot be the result of experience alone. Instead, on the nativist view, there must be some natural, biological basis for moral cognition. One recent formulation of this position is the *Linguistic Analogy* (LA) (Dwyer, 2006, 2009; Dwyer et al., 2010; Harman, 2008; Hauser, 2006; Hauser et al., 2007; Mikhail, 2007, 2011). The LA engages with and extends the traditional rationalist notion that the moral law is "engraved in the mind" (e.g., Grotius, 1625; Kant, 1788; Leibniz, 1705). Inspired by Chomsky's (1965) theory of a Universal Grammar that is subserved by an innate linguistic faculty, the LA hypothesizes that

there is a Universal Moral Grammar subserved by an innate, possibly domain specific, human moral faculty.[1] Advocates of the LA propose that the best empirical findings support the case for the existence of this moral faculty and that it is empirically and explanatorily illuminating to presuppose a strong analogy between the mechanisms underlying human linguistic and moral capacities. Though there have been several recent formulations of the LA (Dwyer, 2006; Roedder & Harman, 2010; Hauser et al., 2007), in what follows I will specifically focus on the account developed by John Mikhail (2011).

In this thesis, I argue that we ought to reject Mikhail's view on the grounds that it fails to give a descriptively adequate account of moral judgment, thus failing to satisfy adequately its own explanatory *desideratum*. This failure is revealed by considering evidence from recent studies of moral judgment and the LA's inability to explain these findings successfully. Since the LA is unable to offer a successful response to the findings I present, it faces a dilemma: either the LA attributes the wrong contents (moral rules, principles, and concepts) to our moral grammar, or the moral grammar model is the wrong model of our moral faculty. There is a straightforward way in which the LA might be wrong about the contents of our moral grammar: by claiming that we innately possess deontic concepts, rules, or principles that we do not, in fact, possess. So, if the LA proposes that we each innately possess a deontic rule that forbids acts of type *F*, but we appear to have no such innate prohibition against *F*-ing, then the LA is wrong about the contents of our moral grammar.

On the other hand, there are a number of ways for Mikhail's moral grammar model to be the wrong model of our moral faculty. The most obvious way in which the moral grammar model might be the wrong model is if moral nativism is false, full stop. However, some brand of nativism might be true, while the moral grammar model is false. For example, morality might still be innate without being analogous to language. Or morality might be innate, and in some sense language like, without moral cognition having the central features that Mikhail's moral grammar model claims. I argue that, at a minimum, the last of these ways of being wrong applies to the LA.

---

[1] The notion of innateness is, itself, much disputed (see Griffiths, 2002; Samuels, 2007). I do not wish to take up this issue here. For the discussion to follow, I simply employ 'innate' (and cognates) in a manner consistent with the sense and usage found in the literature concerning the Linguistic Analogy.

Despite my criticism, I consider Mikhail's account a significant development in contemporary moral psychology. While the LA, like other modern theories, focuses on the elicitation and constitution of moral judgments, it also offers a number of virtues many of its rivals lack. First, Mikhail (2007) suggests that his account offers a fruitful approach to investigating the behavioral features of human morality, as well as the ontogenic, phylogenetic, and computational features of moral cognition. If the analogy between language and morality is a good one, then, moral psychologists have a ready framework from which to develop a rich account of moral learning and development, as well as its evolution. Another asset of Mikhail's model is that it offers a level of analysis unexplored by many recent and influential accounts of moral cognition (e.g., Greene et al., 2004; Haidt, 2001): computational analysis, which allows the LA to model (and generate hypotheses regarding) the mental representations and computations involved in the process of making moral judgments. In offering this level of analysis, Mikhail's LA provides a distinctively complex and detailed portrait of moral cognition. The LA possesses another distinct virtue in that, if it is correct, it can provide moral psychologists with a theoretic framework for understanding the innate limitations or constraints on the development of moral systems, just as Chomskyan accounts of language propose that there may be limitations on what types of languages humans can develop. An upshot of this virtue is that the LA can explain the convergence (and divergence) of moral systems. Thus, the LA, if correct, may also be able to provide moral philosophers with the tools for illuminating (and potentially resolving) moral disagreement. These virtues are noteworthy even if the LA is the wrong model of moral cognition. At minimum, I take it that future models would do well to consider the value of the LA's novel focus on computational analysis, as well as its ambitious attempt to detail which, if any, underlying principles are common to human moral judgments.

In the discussion to follow, I provide a brief portrait of the LA (Chapter 2). I then consider recent evidence that shows that the LA fails to offer an adequate response to broad patterns in human moral judgments including gender differences and partiality toward the young, kin, and in-group members (Chapter 3). I go on to articulate the deeper challenge that these data pose for the LA (Chapter 4), and conclude that

we ought to be skeptical of the very idea of an innate moral grammar, and the Linguistic Analogy on which it relies.

2. MORAL GRAMMAR: MODEL & CONTENTS

Recently, there has been a deepening interest in explaining moral cognition and moral judgment (Cushman, Young, & Greene, 2010; Doris, 2002; Haidt, 2001; Moll et al., 2005; Nichols, 2004; Prinz, 2007; Sripada & Stich, 2006). This interest includes the revival of a notion proposed by John Rawls: the Linguistic Analogy (Dwyer, 2006; Hauser, Cushman, & Young, 2008; Mikhail, 2000). In Section 9 of *A Theory of Justice*, Rawls (1971, pp. 46-53) suggests that his own account of the human moral sense is usefully understood as being akin to linguists' accounts of linguistic competence.[2] Though Rawls provided the inspiration for contemporary proponents of the LA, Noam Chomsky's (1965) work furnished its structure.[3]

As with generative linguistics, the LA's goal is to depict the mental system or innate knowledge underlying a distinctive human capacity. In particular, the LA hopes to provide a *descriptively adequate* account of the human moral faculty and its operations (e.g., judgments). As with much of the LA's conceptual apparatus, this key *desideratum* is borrowed from linguistics.

2.1. Descriptive Adequacy & the Goals of a Grammar

'Descriptive adequacy' is a technical term used to indicate the relative quality or success of a theory as compared to competing theories operating at similar levels of description (Chomsky, 1965). A theory is descriptively adequate if it meets the following criteria:

---

[2] In what follows, I use 'linguistics' and 'generative linguistics' to refer to accounts of the human linguistic faculty inspired by and in harmony with Chomsky's (1965) account.

[3] While Mikhail references Chomsky's work, and generative linguistics, broadly, the account developed in Chomsky's *Aspects of a Theory of Syntax* (1965) provides the fundamental starting point of the LA. Those familiar with the development of his views will rightly resist the following discussion's treatment of Chomsky's work as simplistic, since there are important theoretical differences between, say, Chomsky's account as developed in *Aspects*, and his *Lectures on Government & Binding* (1981), or *The Minimalist Program* (1995). Mikhail's simplification of Chomsky's view, however, need not be vicious, since the LA Mikhail develops is meant to highlight a potentially fruitful research program identified by John Rawls; Chomsky's work is intended to provide the framework and conceptual apparatus involved in this research program. Thus, even if there are significant shortcomings to Chomsky's theory as developed in the *Aspects* (1965), it is possible that Mikhail's LA can avoid such difficulties.

1. It can account for more observed data than other theories.
2. It produces correct predictions.
3. It succeeds in correctly describing its target.[4]

What would it mean for the LA to satisfy these three criteria? In this case, satisfying the first criterion means that the LA's moral grammar model is consistent with, and accurately describes, more moral judgments than other accounts of moral cognition. Meeting the second criterion is straightforward –here judgments the LA predicts we will make must match judgments we do make. To satisfy the third criteria, the LA must correctly describe the mental operations (i.e., the computations) involved in the familiar process of taking in a stimulus as input and producing a moral judgment as an output.

Mikhail (2007) suggests that the LA is descriptively adequate where other theories of moral cognition are not largely because it is sensitive to *computational theory* (Marr, 1982), a level of analysis underemphasized or unmentioned in other theories of moral cognition. This computational analysis considers the mental processes, or "conversion rules", that convert a given stimulus (e.g., a particular moral dilemma) into a structural description –the mental representation of the stimulus' acts and omissions– to which moral (or "deontic") rules will be applied, and the deontic rules implicated in the rendering of moral judgments (Mikhail, 2007, p. 145).

For the moment, however, the fine points of Mikhail's computational analysis are not important. Instead, it is simply worth keeping in mind that even Mikhail's critics, such as Greene, acknowledge that computational theory has been largely ignored in studies of moral cognition (Greene, 2008). Additionally, it is important to note that for Mikhail's computational analysis to be counted as a descriptive virtue of the LA it must accurately analyze the human moral faculty that constitutes its target. If there is no universal moral grammar, then Mikhail's computational analysis merely models the properties of a fictional system. If this is so, then the LA's computational analysis is plausibly insensitive to the broader data set that should constitute its target, and will only minimally or accidentally contribute to the first two criteria for descriptive adequacy –

[4] The 'target' in Mikhail's case, is the human moral faculty "in its mature or steady state" (Mikhail, 2011, p. 22). This last criterion is largely an inference motivated by the success of a theory with respect to the first two criteria.

accounting for more data and producing correct predictions– and fail altogether with respect to the third (i.e., accurate target description).

2.2. The Moral Grammar Model

Two arguments, both "inferences to the best explanation" (Harman, 1965), constitute the foundations the Linguistic Analogy's model of moral grammar (Mikhail 2011, p. 17):

1. *The Argument for Moral Grammar* – the observed properties of our moral judgments are best explained on the assumption that our minds possess a moral grammar.
2. *The Argument from the Poverty of the Moral Stimulus* – the speed and ease with which children acquire "a moral sense" (a moral grammar) is best explained by the existence of an innate genetic program devoted to this acquisition.

It is worth noting that for Rawls (1971) and Mikhail (2011), the Argument for Moral Grammar is logically prior to the Argument from the Poverty of the Moral Stimulus, since the latter succeeds in explaining the acquisition and development of moral grammar only if, in fact, there is a moral grammar that is possessed universally by normal adults.[5]

But what are the properties of our moral judgments for which, according to the Argument for Moral Grammar, a universal moral grammar is the best explanation? This argument follows the structure of the more familiar arguments for linguistic grammar.[6] First and foremost, Rawls (1971) observed that normal adults are prepared to make a "potentially infinite number and variety of [moral] judgments" (p. 46), despite limited exposure to human actions and action contexts (e.g., social arrangements and institutions). Similar to

[5] In light of the logical priority of the Argument for Moral Grammar, my emphasis (and Mikhail's) here is with it, rather than the POS arguments. However, there may be good reason to find this emphasis objectionable, since poverty of stimulus concerns provided much of the initial motivation for developing generative theories in linguistics (Chomsky, 1965).

[6] A natural worry for the LA is whether or not Chomsky's account can serve as a viable basis for drawing an analogy in moral psychology. Addressing this concern is a complex affair, and adequately doing so is beyond the scope of my remarks here – but the question behind this concern is potentially significant and worth making explicit: if Chomsky's program in generative linguistics is indefensible, won't the LA simply inherit those problems? The answer to this question might be affirmative, but depends very much on the details. If, for example, it turns out that Daniel Everett's (2005) work on Pirahã does undermine the Chomskyan claims to linguistic universals, this would clearly not spell the demise of the LA. Yet, if generative linguistics suffers from deep, theoretical problems, it seems likely that the LA –in appropriating its theoretical framework from linguistic theory– may face similar problems.

the linguistic case, then, the question is how to explain a normal individual's ability, and preparedness, to project (*sensu* Goodman, 1983, p. 85) from her modest range of experiences to a "potentially infinite" number of novel cases about which she can render moral verdicts. Further, normal individuals' moral judgments are (in many cases) stable and predictable. To Mikhail and Rawls, it seems plausible that this stability and predictability reflects the existence of a system of deontic rules or principles (i.e., a moral grammar), which serves as a basis for these judgments. For instance, even if I have never heard of elder abuse (or similar forms of abuse), once I am presented with an example of it, I am readily able to make the intuitive judgment that such conduct is immoral. Also, my intuitive judgment that elder abuse is immoral is stable in that it is unlikely to change if I am presented with other examples of elder abuse. According to the LA, then, the best explanation for these observed features of human moral judgments (e.g., stability and boundless productivity) is that each person's mind "contains a moral grammar" (Mikhail, 2011, p. 88). Therefore, the LA concludes, normal individuals' minds do, in fact, contain a moral grammar.

This moral grammar is defined as a cognitive system, comprised of "a complex and largely unconscious system of moral rules, concepts, and principles" (Mikhail, 2011, p. 16). According to the moral grammar model, this system generates the observed human capacity to offer a potentially inexhaustible variety of deontic judgments –those judgments in which the status of 'morally permissible', 'morally obligatory', or 'morally forbidden' is assigned to an action.[7] Though these rules can be consciously represented, the LA claims that in many, if not most cases, these rules are unconsciously or tacitly represented, while playing an operative, causal role in production of moral judgments. This suggestion highlights a further feature of the LA: the distinction between *express* and *operative* principles. According to the LA, in many instances operative principles –that is, the principles responsible for one's moral judgment– may not be available when one makes a moral judgment. And, similarly, express principles (those one can access and invoke) may not be the principles causally involved in one's moral judgments. This distinction has additional explanatory value in that

[7] Mikhail does not discuss two remaining traditional deontic categories: "supererogatory" and "indifferent". While the LA is presented primarily as an account of "moral" judgment, to be precise, his discussion focuses on *deontic* judgments. With this in mind, the scope of my discussion follows Mikhail's.

it helps to account for commonly observed phenomena surrounding moral judgments, including confabulation and moral dumbfounding.[8]

A chief goal of the LA is making these operative rules and principles explicit. Since the LA claims that human moral grammar is a system of predominantly unconscious principles, concepts, and rules, a natural concern is how they are to be discovered. To address this concern, the moral grammar model again borrows from linguistics, invoking the *competence-performance distinction* (Mikhail, 2011, p. 52). Just as one's linguistic competence is the idealized set of unconscious linguistic (e.g., grammatical) knowledge that serves as the basis for one's linguistic performance (i.e., real world language use, interpretation, and evaluation), one's moral competence is the idealized body or system of implicit moral knowledge that serves as the basis for one's moral performance (i.e., real world deontic judgments). Knowledge, here, is not meant to suggest anything like a set of justified true beliefs; rather, the term refers to a potentially domain-specific, implicit system in the human mind-brain.[9] This implicit system, then, is one's moral grammar, which is revealed by one's moral competence.

2.2.1. Competence, Performance, & Considered Judgments

Borrowing from *A Theory of Justice*, Mikhail identifies competent judgments with Rawlsian *considered judgments* (1971). According to the LA, considered judgments putatively reflect our underlying moral competence – i.e., these intuitive judgments are most likely to reveal our moral competence without "distortion" (Mikhail, 2011, p. 53). As Mikhail (2011) suggests, considered judgments are the "categorical data a descriptively adequate moral grammar must explain" (p. 110).[10]

According to Mikhail (2011, p. 83) considered judgments are immediate and spontaneous, stable, stringent, impartial, and certain (Rawls, 1950, p 45). A judgment is *spontaneous*, just in case it is made directly in

---

[8] For more on these topics, see Cushman et al. (2006) and Haidt et al. (2004) respectively.
[9] This knowledge, in a sense, subserves our 'moral know-how'. And this system is 'implicit', in that its operations are often opaque to us, and appear to happen automatically –i.e., one is exposed to a stimulus, such as news of pervasive government corruption, and one makes a moral evaluation without intending to or without deliberately doing so.
[10] These considered judgments, Mikhail notes, are not to be confused with "considered judgments in reflective equilibrium" (Rawls, 1971, p. 51) as the latter involve two features lacked by the former: knowing the principles to which a given judgment conforms, and these principles' source and of those principles' derivation" (Mikhail 2011, p 99).

light of inspecting or considering a situation or event, rather than being pre-planned, imitative, or the result of mere chance –'intuitive' may be another way to characterize this feature. *Stable* judgments are those that an individual retains over time, and are also shared by a class of individuals. Considered judgments are *stringent* in the sense that they are largely immune to revision, even in the face of reflection and argument.[11] On Rawls' (1950) view, these judgments are also *impartial* in that they are not made as the result of a desire for personal gain or intense "emotional duress" (p. 53). Also, these judgments cannot be the result of inadvertently or willfully ignoring the interests of those the judgment concerns. So, Rawls implies that for a judgment to be impartial is for that judgment to give due weight to the interests relevant to that judgment. *Certainty*, here, is understood as a psychological alignment toward a judgment – namely, an intuitive sense about the truth of a judgment, a sense that typically withstands subsequent reflection.

Any judgment possessing these features counts, on Rawls' (1950) view, as a considered judgment. Mikhail agrees, and proposes that competent moral judgments have these same features. More important, then, judgments possessing these features provide insight into human moral grammar, in that these judgments reflect the (relatively) undistorted workings of human moral competence. Conversely, judgments that lack these features do not qualify as considered judgments, and will likely reflect performance errors in moral judgment. Given that the LA is a theory of moral competence, advocates of the LA are not chiefly concerned with judgments involving performance errors, since such judgments putatively reflect the *distorted* workings of the human moral faculty.

Ideal cases for generating a theory of our underlying competence, then, are those cases in which an individual's performance will reflect her competence. In developing a model of moral cognition –that is, a theory of the rules and principles of which human moral grammar is constituted– the LA crucially relies on deontic judgments that are most likely to reflect normal individuals' moral competence. For cases of this kind, it is critical that one's performance is not affected by irrelevant conditions such as distractions, limitations in memory, or shifts in attention or interest, since such conditions are likely to lead to judgments that are

[11] Interestingly, Rawls (and Mikhail, one suspects) would agree that, through reflective equilibrium, it is possible that one would relinquish some considered judgments, despite their stringency.

inconsistent with one's competent verdicts (Chomsky, 1965). Similarly, Mikhail proposes that factors such as emotional duress, self-interest, confusion, and even physical discomfort can all negatively affect one's moral performance. For example, if one is asked to make a moral judgment about a scenario, but the scenario is presented in a way that is extremely complicated, it is likely that one's judgment will not cohere with one's judgment about a simpler case that is morally equivalent (i.e., the context, actions and omissions are identical, as are the agents, patients, intentions, and outcomes are identical). Similarly, my competent judgment may be that a particular type of social or political arrangement is patently unjust; however, if it turns out that in a particular context I stand to gain a great deal from this arrangement, I am less likely to judge that this arrangement is unjust. Instances in which such factors affect one's real world judgments, such that these judgments do not reflect one's underlying competence, are deemed errors in performance.

On Mikhail's view, developing a theory of moral competence (i.e., his theory of moral grammar) is logically prior to an account of moral performance. Hence, for Mikhail, until the moral grammar model under discussion here is complete, the LA cannot (and need not) offer a theory of moral performance. This commitment thus compels defenders of the LA to privilege a particular set of moral judgments as reflective of our moral competence, and to reject others as the byproduct of performance errors.

I have not yet said anything about why Rawlsian considered judgments might plausibly reflect the operations of our innate moral faculty. Why should we think this? Mikhail again points to the analogy between morality and linguistics. If competent grammaticality judgments and competent moral judgments are properly analogous, then the properties possessed by the former (*qua* competent judgments) will also be shared by the latter. He observes that competent grammaticality judgments and Rawlsian considered judgments do, in fact, have multiple, significant properties in common. First, the grammaticality judgments that constitute the base data set of competent judgments in linguistics share the primary features that Rawls and Mikhail identify in their considered judgments: spontaneity, immediacy, impartiality, stability, and stringency. Both judgment types have other notable features in common. For example, competent grammaticality judgments and considered moral judgments are "highly predictable" (Mikhail, 2011, p. 83), since both reflect and accord with widely observed judgment trends. They are both characterized as feeling

'objective' or certain, rather than being perceived as a the result of subjective preference. Lastly, these judgments are also intuitive, in the sense that they are not causally produced by consciously applying rules or principles of morality or grammar (Mikhail, 2011, p. 82). Instead, the moral grammar model hypothesizes that there are three elements behind the human moral faculty's outputs: conversion rules, structural descriptions, and deontic rules.

When a person is presented with a "fact pattern stimulus" –a particular case or moral dilemma– innate conversion rules convert the stimulus into a structural description (Figure 1).

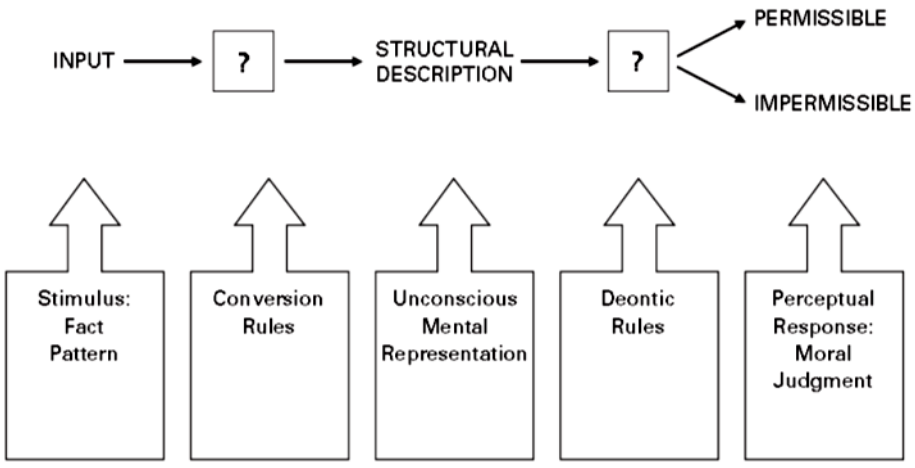


Figure 1. The LA's Expanded Perceptual Model of Moral Judgment[12]

This structural description reflects the structure of acts and omissions (and causal relations, etc.) of the stimulus. Once a stimulus has been converted to a structural description, innate deontic rules are unconsciously applied to that structural description, producing a moral judgment.

Having described these basic elements, the LA is now able to offer a more detailed account of the sequence of unconscious steps usually involved in producing moral judgments (see Appendix A). On Mikhail's view, when one is presented with a moral stimulus (e.g., a moral dilemma) the conversion rules first produce a representation of the temporal structure of the dilemma. This structure is a simple accounting of

[12] From Mikhail, 2007 and 2011.

the stimulus' sequence of events, actions, and omissions. The conversion rules next give rise to the representation of the stimulus' causal structure. When one represents this causal structure, one represents the relations between objects, agents, actions, states of affairs, the effects they produce, and the impact of those effects on the objects and patients described in the stimulus (Mikhail, 2011, p. 172). Following the conversion of a stimulus into its temporal and causal structure, the LA claims that the conversion rules produce a moral structure that represents the good and bad effects involved in the stimulus. Here, harmful effects are represented as 'bad', while positive effects are represented 'good' –bad effects that are negated (prevented) are also deemed 'good'. After converting a stimulus to into its moral structure, its intentional structure follows. This structuring concerns the means, ends, and side effects of the actions isolated in the moral structure. As a last step, the conversion rules culminate with the representation of the deontic structure of the stimulus. This representation builds on the temporal, causal, moral, and intentional structure of the stimulus, adding in the relevant innate deontic rules, principles, and concepts.

On the LA, the deontic rules of our moral grammar include prohibitions of intentional homicide and battery, the Rescue Principle, and the Principle of Double Effect (PDE).[13] These partially represent the contents of our moral grammar. The prohibition of intentional homicide forbids willfully killing another person. A prohibition of intentional battery "forbids purposefully or knowingly causing harmful or offensive contact with another or otherwise invading her physical integrity without her consent" (Mikhail, 2011, p. 117). The PDE is the principle according to which it can be permissible for one to act in a way normally prohibited if that act produces both good and bad effects under four conditions: 1. the forbidden act is not "directly intended" as such; 2. the good effects outweigh the bad; 3. one intends the good effects, but not the bad; 4. one has no "morally preferable" alternatives (Mikhail, 2011, pp. 117-118). For example, one might claim that the PDE allows for military action that results in collateral damage –i.e., unintended damage to non-military targets (persons and property). The Rescue Principle prohibits one from not preventing serious harms (including deaths) that can easily be prevented, provided the harm can be prevented "without violating other

[13] Mikhail discusses other innate rules, but, since battery and homicide prohibitions, as well as the Principle of Double Effect and Rescue Principle, are most central to his account, I focus on these particular rules in what follows.

fundamental moral precepts" and without serious risk to one's safety (Mikhail, 2011, p. 117). This principle is exemplified by the common response to Peter Singer's (1972) example of a bystander who comes upon a child drowning in a shallow pond; the bystander's clothing will be muddied (and perhaps ruined) if she saves the child, but there is no further risk to the bystander.[14]

Minimally, then, these rules constitute the contents of our universal moral grammar – at least as regards the contents of our innate deontic rules. Mikhail argues for this claim, in part, by considering what best explains a broad range of ordinary judgments about familiar philosophical thought experiments. I now turn to this argument, the evidence for it, and its role in Mikhail's broader account of the innate rules, principles, and concepts that constitute our moral grammar's contents. Once the contents of Mikhail's model are clear, my plan is to suggest that there is important evidence for which the LA, in its present form, cannot account. If this suggestion is correct then, I will argue, we have good reason to think that either the LA is wrong about the contents of our moral competence, or it is the wrong model of our moral endowment.

2.3. Our Moral Grammar's Contents

Recall that the LA is fundamentally a model of our innate moral competence, and that Mikhail (borrowing from Rawls) claims that our considered judgments are plausibly competent ones. So, the data for which the LA must account are our considered judgments. Mikhail cites a recently acquired set of considered judgments in support of his model of moral competence, its contents, and the conclusion that humans universally possess a moral grammar (Mikhail, 2000, 2002; Mikhail, Sorrentino, & Spelke, 1998). This evidence is taken from multiple studies of folk intuitions about a range of "Trolley Problem" thought experiments (Fischer & Ravizza, 1992; Foot, 1967; Thomson, 1986). Traditionally, these problems involve scenarios in which a runaway train will kill five people if not stopped or diverted. Subjects are then typically presented with two non-ideal choices. Mikhail and colleagues (Mikhail, 2000, 2002; Mikhail, Sorrentino, & Spelke, 1998) employed a total of twelve variations on these traditional trolley problems, in order to gather a

[14] Mikhail indicates that other rules are likely a part of our moral grammar, though for the most part he does not discuss these rules in detail. For example, he suggests that we may possess a catastrophe prevention rule (a "supreme emergency exemption") which might permit us to intentionally harm another person if a large number lives could be saved by doing so.

detailed enough data set from which to begin modeling the computational structure of moral cognition and its innate contents. In particular, the various trolley variants provide a basis for positing a modest set of deontic rules, concepts, and principles that partially constitute the contents of the human moral faculty. While Mikhail presented twelve different scenarios to subjects, I focus on the four that most crucially support his account of the contents of our moral grammar.[15]

2.3.1. Four Trolley Cases

In one case (*Bystander*), subjects were asked whether or not it would be morally permissible for Hank to throw a switch, diverting the path of the train so that it will kill one person, instead of five. In another case (*Footbridge*), subjects were asked about the moral permissibility of Ian's stopping the train before it kills five people by pushing a large person off of a footbridge that passes over the train tracks. Though Ian's pushing the large person from the footbridge would stop the train, it will also kill the large person.

Mikhail labels another related pair of cases *Loop Track* and *Man-in-Front*. In the former, a runaway train will kill five men if it is not diverted. Ned, a bystander can temporarily divert the train onto a sidetrack that eventually loops back onto the main track. If Ned diverts the train onto the side track, it will strike a man who is standing there, killing him, but slowing the train enough to allow the five men on the main track to escape. The *Man-in-Front* case also involves five men on the main track, and a side track that eventually loops back to the main track. A bystander, Oscar, can divert the train onto the side track, where it will strike a heavy object, slowing the train. There is a man standing in front of the heavy object, and –though it is the heavy object that will slow the train enough to allow the five men on the main track to escape– Oscar's throwing the switch will kill the man on the side track. The majority of subjects (nearly 90%) judged that it is morally permissible to for Hank to throw the switch in the *Bystander* case, but nearly 90% of subjects judge that it is impermissible for Ian to push the stranger in the *Footbridge* case (Mikhail, 2002; 2007). For *Man-in-Front*, a narrow majority (62%) judged it morally permissible for Oscar to throw the switch, while a narrow majority (52%) judged it impermissible for Ned to throw the switch.

[15] See Appendix A for the scenarios as presented in Mikhail and colleagues' studies.

For Mikhail, not only do the responses to these trolley cases reveal the operations of our moral faculty, they can be used to discover the contents of our moral grammar. Specifically, by considering structurally similar cases that elicit differing moral responses we might discover the concepts, rules, and principles that are partly constitutive of our universal moral grammar. Since the outcomes, in terms of lives saved and lost, in *Bystander* and *Footbridge* are identical (if Hank throws the switch and Ian pushes the large stranger) the LA, like any account of moral judgment, must explain the striking asymmetry between our judgments about the two cases. The more modest differences between judgments about Oscar and Ned similarly warrant explanation. If the LA is the correct model of the human moral faculty, then these asymmetries are best explained by the rules, principles, and concepts that partially constitute our moral grammar. With this provision in mind, what explains the difference between judgments about the *Bystander* and *Footbridge* cases, and judgments about *Loop Track* and *Man-in-Front*, is that we innately possess deontic rules and principles that apply to one case, but not the other.

As I have mentioned, Mikhail claims that we innately possess a deontic principle against intentional battery (i.e., performing battery as a means to some other end), and the PDE. By appealing to these principles, the LA can readily explain the differences in permissibility judgments when comparing Hank and Ian, and also Ned and Oscar. To save the five people on the main track in *Footbridge*, Ian intentionally commits three acts of battery (touching the large person without consent, causing him to strike the ground after falling from the footbridge, and causing the man to be struck by the train) and knowingly produces the side effect of killing the large person as a means of intentionally producing the end result of saving five people.[16] Thus, to save the five Ian violates the PDE, as well as our rule against intentional battery. If Ian does not act by pushing the large person from the footbridge, he does not save the five, but he does not violate the PDE either. In throwing the switch Hank in *Bystander* knowingly commits an act of battery (though his committing battery, as such, is not his intended end) and an act of homicide as a side effect of preventing five deaths. According to the LA, Hank's behavior, in contrast to Ian's, is deemed permissible

[16] Mikhail uses this language, even though it may seem unusual to refer to killing the large person a 'side effect'. Mikhail's point, I take it, is that the death of the large person is not, itself, intended, nor is his death strictly speaking a means to Ian's intended end of saving the five on the tracks.

since it conforms to the PDE, the Rescue Principle, and it does not violate our innate rule against intentional battery.

The asymmetry between Oscar and Ned is explained in a similar way. In throwing the switch, both intentionally produce the good result of preventing five deaths. However, Ned intentionally commits battery as a means to slow the train and save the five, since to save the five requires the train slow as a result of striking the man on the strike track. Also, Ned knowingly commits homicide as a side effect of intending the train to strike the man on the side track – the point being that Ned doesn't want the man's death, but he does intend for the train to strike the man. Oscar, on the other hand, merely commits battery and homicide as side effects of slowing the train, since he intends for the train to strike the heavy object on the side track, and it just so happens that there is no way to achieve this intention without also having the train strike the man on the side track. So, the LA explains that Oscar's switch throwing is deemed morally permissible by most people, since it violates no innate principles or rules, but, since Ned violates the PDE and our innate rule against intentional battery, his switch throwing is judged impermissible by most people.[17]

It is important to note that, according to the LA, the deontic judgments provided by the majority in these cases constitute considered judgments, in the Rawlsian sense. So, since Ned's throwing the switch and Ian's pushing the large person are judged morally impermissible by the majority of respondents, these modal responses are, on Mikhail's view, the considered judgments. The majority judgments that Hank and also Oscar are morally permitted to throw the switch also count as considered judgments.

Mikhail (2011) highlights this commitment by making "the simplifying assumption that these judgments [the majority modal responses] are considered judgments in Rawls' sense" (p. 110). Additionally, as with all considered judgments, these possess the key features (spontaneity, impartiality, stability, stringency, and certainty) required of such judgments. So, Mikhail proposes that considered judgments are competent judgments, as are the majority modal responses, and that these judgments reflect the operations and contents

[17] It might seem that since these cases (Ned and Oscar's) are complex, and only subtly different, that the LA can easily explain the observed response patterns (e.g., the semi-split decision in *Loop Track*) as resulting from the complexity of, and subtle differences between, the two cases. In §4 I critically assess this explanation, as well as Mikhail's claim that 'most people' judge Ned's switch throwing to be impermissible and the significance of this (narrow) majority.

of our moral grammar. Conversely, the minority modal responses are rejected as errors in performance and do not, according to the LA, accurately reflect our innate moral endowment.[18] In §3, I go on to discuss how the LA responds to cases in which the majority response is only slightly larger than the minority response, and argue that, ultimately, this type of response presents a problem for the LA. Specifically, it reveals that the LA is unable to satisfy one of its key *desiderata*: descriptive adequacy.

2.4. Summary

I will now argue that the LA is undermined by recent empirical work on moral judgments, since it fails to offer a satisfactory response to three lines of evidence that plausibly constitute part of our innate moral endowment. The failure to deal with this evidence successfully, I suggest, is the result of the LA's untenable account of moral competence. Before detailing this evidence, it is worth keeping in mind the fundamental structure of the LA as I have just described it. The LA concludes that we each possess an innate system of deontic principles and rules, and that this system is implied by the observed features of moral judgments.[19] This system of principles and rules is characterized by a theory of moral competence, and our considered judgments, in turn, reveal the undistorted operation of our moral competence. Mikhail (2011) examines a set of judgments about trolley problem variants, and claims that these judgments qualify as considered judgments. This data set of considered judgments provides the LA with a starting point for determining the specific moral rules and principles that are operative in the causal production of normal individuals' judgments. Having developed an account of the rules and principles implicated in considered judgments about trolley cases, the LA can then begin to predict the competent judgments of normal individuals, at least in those cases sufficiently similar (i.e., cases possessing the same set of value neutral facts) to the trolley problem data set.

---

[18] 'Modal response' is meant to refer to statements concerning how things ought to be. The response that pushing the large person in *Footbridge* is forbidden (i.e., you ought not push the large person), is the majority modal response, whereas the judgment that pushing the large person is not forbidden is the minority modal response.

[19] At least one virtue of our moral faculty (and its rules) being innate, on Mikhail's view, is that it might help to address Poverty of Stimulus worries –worries that empiricist accounts cannot possibly be correct since the 'moral' stimuli in our developmental environment is cannot explain how children can use and understand and use moral concepts and make moral judgments.

3. EVIDENCE AGAINST THE ANALOGY

Having explained the LA's moral grammar model, and the putative contents of our moral grammar, I will argue that, though a descriptively adequate account of moral cognition may be possible, the LA cannot satisfy this *desideratum*. In particular, when one considers the evidence for which any general model of moral cognition must account, one finds that the LA's theoretical commitments to a Rawlsian account of considered judgments, as well as its way of drawing the competence-performance distinction, produces a descriptively inadequate model of moral cognition. The evidence I will now consider puts pressure on Mikhail's account of moral competence, and, as a result, his model more broadly. Since the LA critically relies on the distinction between competent and erroneous judgments, if its account of competence is untenable in light of the considerations I raise, then the LA, itself, may prove untenable as well.

This chapter proceeds in three parts. I first consider the significance of the minority modal responses to trolley cases, then move on to the differences in women and men's moral judgments, and conclude by considering our patterns of preference toward our kin, the young, and social in-group members. I also discuss why the LA fails to offer a satisfactory response to these lines of evidence. In the next chapter I go on to highlight why the LA must respond to this evidence, and whether or not the LA can succeed in doing so while preserving its structure as a model of moral cognition and its account of moral competence. I conclude that it cannot.

3.1. Modal Minorities

Recall the *Loop Track* and *Man-in-Front* cases involving Ned and Oscar. 52% of subjects judge that it is morally *impermissible* for Ned to throw the switch, killing the man on the side track as a means to slow the runaway train and save the five people on the main track (Mikhail, 2002). In contrast, 62% judge that it is *permissible* for Oscar to throw the switch, saving the five people on the main track and killing the man on the side track as a side effect of diverting the train to slow it (Mikhail, 2002).

Recall, also, Mikhail's "simplifying assumption" that the majority modal responses are Rawlsian considered judgments, and thus constitute the data for which the LA, as a theory of moral competence, must account (Mikhail, 2011, p.110). Clearly, then, on Mikhail's view, the LA does not accept modal minorities, nor does it need to. So, despite the slim margin between the 52% who judge Ned's switch throwing *impermissible* and the 48% who judge Ned's switch throwing *permissible*, the LA is committed to rejecting the latter as revealing anything significant regarding the operations or contents of the human moral faculty.

Interestingly, the same will also be true for many of the judgments offered by any person who consistently responds with paradigmatic utilitarian or deontological judgments about these trolley cases. Imagine an individual, call him "Pete", who judges that that Ian's pushing the large man from the footbridge is permissible, and that throwing the switch is permissible for Hank, Oscar, and Ned in *Bystander*, *Man-in-Front*, and *Loop Track*, respectively. Based on the LA's commitment to majority responses as considered judgments, fully half of Pete's judgments would be considered performance errors and ignored by the LA. The same is true for another imagined person, call him "Manny", who judges that it is morally impermissible for Hank, Oscar, and Ned to divert the runaway train, and also judges that it is impermissible for Ian to push the large man in *Footbridge*. As with Pete, the LA happily embraces half of Manny's judgments as competent, but dismisses his remaining judgments as performance errors. In both cases the defender of the LA is committed to the claim that when Manny and Pete's judgments diverge from the majority, those judgments cannot be the result of the proper functioning of the human moral faculty regardless of whether or not Pete and Manny would (or could) justify their judgments across these four cases in terms of a moral theory they each hold.

Though I have focused on the minority modal responses as represented by Pete and Manny's imaginary judgments, as well as the judgments that Ned's switch throwing is permissible and that Oscar's is impermissible, the general message should be clear: these modal minorities constitute potentially significant moral judgment data for which any theory of moral cognition must account. Yet, the LA effectively ignores these data by assuming that the majority response just is the competent or considered response (and the minority response is not), even in cases where subjects' judgments are almost evenly split. Though I take it

that this message is generally noteworthy, it is particularly important given the role played by the majority judgments on the *Loop Track* and *Man-in-Front* cases in supporting Mikhail's claims that the PDE is among the innate deontic principles that partly constitutes our moral grammar. Since people's judgments about *Loop Track* and *Man-in-Front* only support the LA if Mikhail's view of competence is correct, should it turn out that the minority modal responses undermine Mikhail's view of competence, these responses will, in turn, undermine his claims regarding our moral grammar's contents. As I will argue in §4, these responses do in fact call Mikhail's view of moral competence into question, and thus undermine his broader account of moral cognition.

3.2. Gendered Judgments

A second line of evidence that the LA fails to offer an adequate response comes from data on gender and moral judgments. In the last twenty years multiple studies have found differences in moral judgment across genders (Gilligan & Attanucci, 1988; Petrinovich et al., 1993; Petrinovich & O'Neill, 1996; Zamzow & Nichols, 2009). Petrinovich and colleagues (1993; 1996) employed trolley variants modeled on those discussed in Fischer & Ravizza (1992). They observed that women found throwing the switch in *Bystander*, saving five people and killing one, significantly more permissible when the questionnaire materials employed the term "Save" rather than "Kill" in describing the possible outcomes of acting or refraining from action. Men's responses, under the same conditions, varied less often, and did not vary significantly.

Zamzow & Nichols (2009) also found gender differences in a study involving *Bystander* case variants. In this study, male and female subjects were asked to consider trolley cases structurally similar to *Bystander*, and to evaluate the moral acceptability of throwing the switch, killing one and saving five, depending on the description of the person on the side track. In one condition, the person on the side track was either described as a stranger or as a 12 year old child. In another condition, this person was described as either the subject's brother or the subject's sister. Males expressed greater agreement than did females that throwing the switch was 'morally acceptable' in the sister and 12 year old child conditions. Conversely, females expressed

greater agreement than did males that throwing the switch is morally acceptable in the brother and stranger conditions (Zamzow & Nichols, 2009).[20]

Even Mikhail's own research (2002) on moral judgment is consistent with this pattern of gender differences. Mikhail (2002) reported significant differences in male and female permissibility judgments regarding the *Bystander* case. While men, on average, rate switching the track, saving five people and killing one, permissible 85% of the time, Mikhail found that 60% of female subjects rated switching as permissible. Though the sample size in this study is modest, these gender differences are not only statistically significant, they match a pattern found repeatedly in other studies (Gilligan & Attanucci, 1988; Zamzow & Nichols, 2009): though a substantial majority of males find diverting the trolley permissible in the standard *Bystander* case, women systematically judge throwing the switch less permissible (though switch-throwing remains the majority response among women).

Though the above mentioned studies (with the exception of Mikhail 2002) did not employ trolley cases identical to those used by Mikhail, there is a clear pattern of systematic differences in moral judgment across genders. This evidence is modest but mounting, and, at present, the LA is largely insensitive to it. Rather, the LA currently reduces the complexity of the evidence by averaging men and women's responses together, and selecting the majority modal response as the competent response, thus constituting the response the LA must account for. Conversely, regardless of gender, the minority responses are labeled performance errors and rejected as data irrelevant to a theory of moral competence. This rejection, however, effectively eliminates a significant proportion of women's responses to specific cases (e.g., 40% of them in Mikhail's *Bystander* case), and ignores robust variability across genders. If we concede that this variability does not matter, then Mikhail's account of our moral grammar's contents fits the data nicely, but, as I will argue in Chapter 4, there is reason to resist this concession.

At this point, it is worth noting that the need to explain these data is not unique to the LA. Any sufficiently general theory of moral cognition must also explain the differences in judgments across genders.

---

[20] Subjects were asked to imagine themselves in the role of Bystander, and rate their agreement with the following question (among others): "Is it morally acceptable for me to pull the switch?" Subjects were also asked what they thought they *would* do, if in the described situation – there were no significant gender differences in the answers this further question received.

This is especially true of the LA, since its goal is to offer a descriptively adequate account of *human* moral competence. The LA does purport to be an account of women's moral competence and men's moral competence. Thus, when gender judgments diverge significantly, the LA must successfully explain these differences such that it preserves a tenable and unified account of the human moral endowment.

At first glance, it appears that the proponent of the LA can explain these gender differences as the result of gender-specific performance errors. One strategy would be to claim that men are prone to ignore relevant facts when faced with cases concerning the prospect of saving others, or, conversely, perhaps women more often pay attention to irrelevant factors when considering these kinds of cases. In the next chapter I consider this strategy, and what becomes of the LA if it is taken. I conclude that taking this strategy is unpromising, since it fundamentally relies on the assumption that majority judgments are competent judgments. As I will show, this assumption turns out to be deeply problematic for the LA, supporting my conclusion that the LA is descriptively inadequate and cannot be rescued from this inadequacy by explaining away the evidence presented here. Before turning to these arguments in detail, I will consider one last line of evidence.

3.3. Patterns of Preference

In addition to the data on gender differences in moral judgment, there is a remaining line of evidence for which the LA fails to offer an adequate response. Consider the following variant on the familiar *Bystander* case. A bystander, Jayne, may choose to throw the switch or to refrain from doing so. As usual, throwing the switch saves five and kills one. Imagine the five persons on the main track are undescribed strangers to Jayne, whereas the individual on the side track is an elderly person. Is Jayne morally forbidden from throwing the switch? What if the person on the side track is a toddler instead? Further, imagine variants in which five undescribed strangers are on the main track, and the person on the side track is Jayne's romantic partner or Jayne's genetic relative (but not both!). Again, is it permissible for Jayne to refrain from throwing the switch under these conditions? Imagine Jayne in the *Footbridge* scenario. Imagine that she can push the large person from the footbridge in order to save the five people on the main track. If the five people are all children, is it

permissible for Jayne to push the large man from the footbridge? What if Jayne's romantic partner is among the five on the main track? Would it be permissible for Jayne to push the large man from the bridge to stop the train if all five people on the tracks were Jayne's relatives?

A recent study by Bleske-Rechek and colleagues (2010) tested participants using the *Bystander* variations just mentioned.[21] In brief, they found that, as compared to cases with an undescribed person on the side track, participants were significantly more likely to throw the switch, directing the trolley toward older persons or unrelated persons, and that participants were substantially less likely to throw the switch to direct the trolley toward the younger or more genetically related the person on the side track. Subjects were also substantially less likely to flip the switch if they were told to imagine their romantic partner on the side track.

As Paul Bloom (2011) observes, going to great lengths on behalf of one's own children, family, or members of one's community is prevalent. It is noteworthy that failures to go to similar lengths for strangers (or out-group) members are seldom deemed *moral* failings (cf. Singer, 1972). But the reverse is also true (Bloom, 2011). If, for example, I fail to help or care for my family in commonplace ways –failing to secure medical treatment for my children, say– this is widely viewed as a moral failing. Further, it seems plausible that normal (non-pathological) individuals who disagree that it is immoral not to care for one's kin are thought to be mistaken.

Viewed through the lens of evolutionary theory, though, these judgment tendencies are not particularly surprising. In fact, one would predict these very responses in light of our evolutionary heritage (de Waal, 2006), since it seems entirely plausible that our moral sensibilities first evolved in a way that singled out the young as worthy of protection, and distinguished kin or in-group members as worthy of prosocial treatment that was not extended to non-kin or out-group members.[22]

As with the differences in judgments across genders, the LA once more fails to successfully explain these variations. As structured, the LA's base data set is comprised of responses to trolley variants involving

---

[21] The conditions tested in this study partially overlap with, but extend those considered by Zamzow and Nichols (2009).
[22] For more, see Trivers (1971) on altruism, Sober & Wilson (1998) on prosociality, and de Waal (1996) for general discussion.

victims who are unnamed (presumably adult), male strangers. To be clear, the data I discuss in this section – patterns of kin, young, and in-group preferences in moral judgment– are not merely significant because the LA does not yet say much about them, it is because they have the potential to present a problem whether or not the advocate of the LA attempts to embrace them, or if she rejects them as the result of performance errors. As I argue in the next chapter, explaining these data is difficult for the LA, since it cannot coherently subsume these patterns of preference into its model, and it cannot offer a principled basis for rejecting these data. In either case, this difficulty further supports the claim that the LA's account of competence is problematic, as well as my conclusion that the LA is the wrong model of our moral endowment.

3.4. Summary

These three lines of evidence pose a similar sort of challenge to the LA. On the face of it, the challenge seems modest. I have simply claimed that the LA currently fails to offer an adequate response to the evidence presented to this point. Naturally, the defender of the LA has two available responses. She can make changes in order to accept this evidence, integrating it into her model's base data set, or she can offer a principled basis for rejecting it. However, I argue that the LA cannot successfully accept or reject these data. As a result, the advocates of the LA can either admit that their moral grammar model assigns the wrong contents (principles, rules, or concepts) to our moral faculty, or they can concede that the moral grammar model is simply the wrong model of our moral endowment.

4. THE ANALOGY'S DILEMMA

In the previous section I presented three lines of evidence, and suggested they each present the same sort of difficulty for the LA and its account of moral competence and performance errors. In this chapter, I detail the nature of this difficulty, and its implications for the LA. As I emphasized previously, the LA currently fails to provide adequate responses to a broad range of modal minority judgments, the differences in judgments across gender lines, and patterns of preference toward kin, social in-group members, and the young. I now consider each line of evidence in turn, and discuss the LA's prospects, depending on which

strategy –acceptance or rejection– it adopts. I then show how both responses are problematic for the LA's account of the contents of our moral grammar, as well as its depiction of moral competence.

4.1. Explaining Modal Minorities

In §3.1., I introduced what I've termed 'modal minority' responses. For the purposes of this discussion, these are the responses favored by the minority of subjects asked to offer permissibility or impermissibility judgments about the trolley problems Mikhail uses to develop his model of moral cognition. For example, the 48% of subjects who judge that Ned's switch throwing is *permissible* in the *Loop Track* case are offering the modal minority response, as compared to the 52% of subject who judge Ned's switch throwing *impermissible*. However, the notion of modal minority responses (MMR) can be extended to include the minority response to any experimental probe used for the purposes of modeling our moral competence.

Just as with the other lines of evidence, the LA's defenders face two options with respect to the MMR evidence: accept it as reflecting our innate moral competence and thus part of its data set, or reject the MMR as the result of performance errors thus excluding it as part of the data for which it must account. Currently, the LA, as structured, pursues the latter strategy. The reason for this choice is straightforward. If the advocates of the LA were to accept the MMR evidence as reflective of our innate moral competence, while affirming that the evidence it currently accepts (the modal majority responses) also reflects the undistorted operations of our innate moral faculty, the LA would be forced to attribute contradictory contents (rules or principles) to our innate moral faculty.

For example, if the LA accepts that the 48% of subjects who respond that Ned's throwing the switch is morally permissible are making competent judgments, and also accepts that the 52% of subjects who judge that Ned's switch throwing is impermissible judge competently, then the advocates of the LA must conclude that there both is and isn't an innate prohibition against intentional battery (i.e., performing battery as a means to some other end). Similarly, if the LA accepts as competent the responses of the 38% of subjects who judge that it is morally impermissible for Oscar to throw the switch in *Man-in-Front*, as well as the 62%

who respond that Oscar's switch throwing is permissible, then the LA must seemingly affirm that the PDE (or a similar principle) is tacitly accepted in our moral grammar, and also that it isn't.

So, by including the MMR and the modal *majority* responses, the LA cannot offer an internally consistent depiction of its own core data set, since that data set involves contradictory deontic judgments about identical cases. But, at minimum, for the idea of a universal human moral grammar to preserve its promise of descriptive adequacy, and to be worth modeling, there must be a principled and non-contradictory set of contents which partly constitute our moral grammar. If the Linguistic Analogy is left to conclude that we do possess an innate battery prohibition and that we also do not –which it must, if it accepts both the MMR and modal majority responses– the LA's account of the contents of our innate grammar is either vacuous, or requires giving up on the very notion of a universal moral grammar. On either conclusion, it isn't clear what explanatory or predictive value the LA retains as a model of moral cognition.

While this simplistic integration of the MMR data is unpromising, perhaps the LA has the resources for a more sophisticated and compelling response. It seems plausible that the advocate of the LA can defuse my worries in this section by offering the Plurality Response, which, in brief, proposes that since our moral grammar is complex and encompasses more than one rule, in some instances different innate rules may be responsible for different modal responses, and so this disagreement is not the result of performance errors. Not only does the Plurality Response suggests that each person innately possesses multiple deontic rules and that there are going to be cases to which more than one deontic rule is applicable, it also suggests that two people's competent judgments about such cases may disagree, since one person's judgment reflects one innate rule, and another's judgment reflects a different rule.

So, for example, if Pete responds that it is morally permissible for Ned to throw the switch in *Loop Track*, and Manny judges that it isn't, perhaps it is because the Rescue Principle is driving Pete's response, while the prohibition against intentional battery is driving Manny's response. Or, to be more precise, the Rescue Principle has more weight for Pete's decision making, while the intentional battery prohibition is weightier for Manny. Thus, we may be able to conclude that both Manny and Pete have made competent judgments about the *Loop Track* case. It bears mentioning that, in this instance, neither judgment needs to

count as an error, since the difference in judgment can still be the result of Manny and Pete both structurally representing the same case in the same way, and appropriately applying an innate deontic rule to that structural representation.

To a first approximation, then, the Plurality Response seems to answer easily the concerns that the MMR data raise for the LA. If so, then the MMR data do not challenge the contents the LA attributes to our moral faculty, and these data do not motivate the conclusion that the moral grammar model is mistaken. However, there are substantive reasons to think that this response is unavailable to proponents of the LA, one of which comes directly from the Rawlsian commitments of Mikhail's account.[23]

Though Mikhail (2011) does not discuss the Plurality Response, he very briefly considers how it might be that, when one offers a competent judgment about a case, one innate deontic rule determines one's judgment, even though others are applicable. Once more, he calls on Rawls, specifically invoking the notion of *lexical priority* (Mikhail, 2011, p. 146-152). In *A Theory of Justice*, Rawls (1971, p. 40-45) uses this notion to address the challenges involved in ranking, or assigning the relative weight to, moral principles. In lexically ordering our principles, the elements that have priority 'trump' those that do not. This ordering is one "which requires us to satisfy the first principle in the ordering before we can move on to the second, the second before we consider the third, and so on" (Rawls, 1971, p. 42). For example, Mikhail suggests that the innate deontic rule against intentional homicide is lexically prior to the innate prohibition of intentional battery (2011, p. 152). Similarly, the prohibition of intentional battery is lexically prior to the Rescue Principle. So, using this ordering, we can explain how the innate rule against intentional battery gives rise to the judgment that Ian's pushing the large person from the footbridge is forbidden, even though the Rescue Principle

[23] One might think that phenomenological considerations will help distinguish which cases of disagreement the Plurality Response can help resolve. For example, it might seem that the Plurality Response especially applies to cases about which we have conflicting intuitions but are prompted or forced to make judgments. This specification may initially seem helpful to the LA, since the Plurality Response might then explain the phenomenology of being conflicted about a case, and, under this specification, the familiar experience of having conflicting intuitions provides a possible criteria for distinguishing cases in which the Plurality Response might be useful for explaining disagreement from cases in which disagreement is best explained by performances errors. Unfortunately, cases in which one experiences clearly conflicting intuitions cannot provide the LA with the means to distinguish competent judgments from performances errors. These intuitions will not qualify as competent judgments because they will not count as considered judgments in Rawls' sense. For example, conflicting intuitions, by their very nature, are not accompanied by the certainty, stability, or stringency that attends Rawlsian considered judgments –all of which are (conjointly) features common to considered judgments.

relevantly applies to the case, and would seem to oblige Ian to take action to save the five –since the intentional battery prohibition is prior to the Rescue Principle, our competent judgment follows the intentional battery prohibition in this case. Mikhail's claims that the lexical priority of our deontic rules explains which rules will determine our competent judgments in any particular case, and that indicates that our innate set of deontic rules is partly constituted by the structure of lexical priority of these rules.[24]

Returning to the Pete and Manny example, we can see why the proponent of the LA is unlikely to respond to the modal minority response data by adopting the Plurality Response. First, accepting the Plurality Response destabilizes the LA's account of our innate deontic rule set. Remember that the Plurality Response claims that Manny and Pete disagree about the *Loop Track* case because each is applying a different innate deontic rule to that case, implying that they possess deontic rule sets that are different in lexical order. So, while in Manny's rule set the prohibition of intentional battery is prior to the Rescue Principle and throwing the switch is impermissible, perhaps in Pete's, the reverse is true, then the Rescue Principle has priority. But notice that if the Rescue Principle has priority over the rule against intentional battery, then Pete implicitly no longer accepts the PDE, since favoring the Rescue Principle in the *Loop Track* case means that one can intentionally case a bad effect in order to produce a (greater) good effect. This is clearly at odds with the PDE's core commitment to the claim that producing a bad effect is permissible, provided that the bad effects are mere side effects, and are outweighed by the good effects. The train striking Ned in *Loop Track* is not a mere side effect, it is the means by which the good effect (saving the five on the main track) is achieved.

Not only does the Plurality Response destabilize the LA's account of our moral grammar's contents, but it also is incompatible with Mikhail's account of the lexical priority of our deontic rules. Since the lexical priority of our deontic rules is partly what composes the contents of our moral grammar, to attribute different lexical orderings to different individuals (an implication of the Plurality Response) is to effectively claim that

[24] Admittedly this is an oversimplification. Lexical priority may not be behind the order of our set of deontic principles. For example, as Rawls (1971) suggests, perhaps there is an overarching (meta)principle that determines the ordering of our set of moral principles. Regardless of the mechanism, however, the outcome is the same: our innate set of moral principles, rules, and concepts, has a structure of priority, with some (e.g., the prohibition of intentional homicide) clearly trumping others (e.g., the prohibition of intentional battery), and this structure of priority determines which judgments about a case are the competent judgments.

two individuals have different moral grammars, which implies that there are multiple human moral competencies. The proponent of the LA will reject this consequence, and so reject the Plurality Response, since the LA is centrally committed to there being a universal human moral faculty (and competence). Imagining for a moment that the LA could accept the Plurality Response without relinquishing its central theoretical commitments, an added problem for the LA is that as a methodological matter it would then have to develop a principled means for determining which cases are due to performance errors and which due to individuals' having differently ordered sets of principles. This worry brings us to the second reason for thinking the Plurality Response is not an available response for the proponent of the LA.

Accepting the Plurality Response puts pressure on the very idea of a distinction between competence and performance, such that, if the MMR and majority responses can both reflect competent judgments, then the competence-performance distinction collapses. In brief, the ability for the LA to develop an account of our innate moral endowment (and the rules that partly comprise this endowment) relies on the assumption that those studying moral cognition will be able to effectively distinguish competent judgments from performance errors. If it is possible that for any instance of disagreement between Manny and Pete, both have made competent judgments, then the LA must offer some criteria for distinguishing which cases of disagreement are due to performance errors and which aren't. The strategy that Mikhail (2011) alludes to for distinguishing cases of this kind involves observing which principles have lexical priority, and attributing errors to those judgments that do not cohere with the principles that have priority. But, since the Plurality Response effectively denies that there is a common lexical ordering scheme for our judgments, then lexical ordering cannot be used to distinguish competent judgments from errors –at least not if the proponent of the LA accepts the Plurality Response.

At this point it might seem promising for the defender of the LA to suggest instead that the MMR data are best explained by the fact that subjects are representing the facts about Ned's (or Oscar's) case differently than those offering the majority response. This suggestion, however, will not allow the LA to accept the minority responses as competent, since claiming that the subjects who offer the minority response are representing, say, *Loop Track* differently than the majority is simply to claim that the members of the

minority are making a performance error, since for two people to make competent judgments about the same moral stimulus, they must not only render the same moral judgment, they must also mentally represent that stimulus (including its value neutral facts) in the same way. Thus, if Maude offers a competent moral judgment about a case, and Jeff offers a different judgment about that case, it may be true that the difference between their judgments is best explained by Jeff's representing the case differently than Maude, but this just is to say, according to the LA, that Jeff's judgment is the result of performance error(s).

So, since accepting the MMR *and* the majority responses as competent is unavailable to the LA, perhaps its current approach (rejecting the MMR as performance errors) is more promising. The defender of the LA can simply argue that she needn't consider the MMR, since such responses constitute performance errors, and thus are not part of the relevant data for which her theory of moral competence must account. Specifically, the LA's defenders can take up the suggestion currently under discussion: the MMR data result from subject's structurally representing the fact patterns of cases such as *Loop Track* and *Man-in-Front* in an atypical (and incorrect) way. Though at first glance this seems like a fine proposal, it is ultimately problematic for the LA.

First, I take it that this proposal is especially enticing here because the *Loop Track* and *Man-in-Front* cases seem significantly more complex than (for example) *Footbridge* and *Bystander*. Perhaps it is the complexity of the former cases that results in the significant numbers of atypical fact pattern representation, which in turn explains the high incidence of performance errors when subjects are asked to judge Ned and Oscar's possible behaviors. Once more, the analogy with linguistics provides a useful illustration. When linguists ask adults to make grammaticality judgments about particularly complex probes (e.g., sentences with multiple center embedding), they expect to find more errors in performance as compared to simpler probes. As with *Loop Track* and *Man-in-Front*, the increased performance errors in judgments about structurally complex probes are plausibly due to some subjects representing the probes incorrectly. Interestingly, though, this analogy works against the LA, rather than for it. Consider a linguistic probe that is analogous to Mikhail's *Loop Track* in that it elicits a 52%-48% split among subjects. In the face of a response pattern of this kind, the linguist is unlikely to take Mikhail's approach and label the majority judgment the competent one. Instead, the

linguist will reject the probe as being too complex to shed light on the features of our linguistic competence, or the contents of our linguistic grammar, though the probe may reveal a possible complexity threshold for sucessful linguistic performance. If the LA's defenders wish to follow the linguist's lead, they would be well-advised to do away with *Loop Track* and *Man-in-Front*, since both elicit split-decisions that are plausibly due to the cases being too complex for subjects to mentally represent correctly. Unfortunately for Mikhail, though, doing away with these two cases undermines the key evidential support for the LA's account of the contents of our moral grammar –especially as it relates to the PDE. Given this cost, it seems unlikely that the defenders of the LA would want to explain the minority response (as performances errors) as primarily resulting from the complexity of cases like *Loop Track* and *Man-in-Front*.

There is a second reason to think that the advocates of the LA would not want to explain the MMR data as resulting from subjects' incorrect structural representations of the relevant stimuli. While this explanation may be helpful to Mikhail, it has the potential to work against his case, since in at least some cases "competent judgments" might also be the result of atypical, divergent, or incorrect structural representations too? Recall that the LA simply assumes that when two people's judgments about a case agree, and those judgments match the considered judgment about the case, then neither person has made an error in performance. This means that both responses are driven by the same structural description, as well as the application of the same deontic rules. But it seems entirely plausible that two people could possess different structural descriptions of a stimulus, and apply different deontic rules or principles to their structural descriptions such that their deontic judgments can coincide. In this case, perhaps only one, or neither, person is making a judgment that reflects our innate moral endowment, yet Mikhail's account concludes, on the basis of their agreement, that they must be caused by the same (or sufficiently similar) structural representations. The possibility that moral verdicts might agree even though they are the result of different deontic rules and structural descriptions further serves to undermine the assumption that majority judgments are considered judgments. So, if it is likely that the MMR data are the result of errors in some subjects' mental representations, it seems similarly likely that at least some of the majority responses might also be the result of subjects making errors by representing the cases incorrectly and applying the wrong deontic rules to their

representations. If Mikhail's base data set of putatively competent judgments is partly comprised of judgments that are the result of performance errors, then his account of competence will be based on the wrong data –wrong, at least, by his lights.

Even if neither of the preceding two objections is successful, I do not think that the LA can justifiably reject the MMR data. Remember that Mikhail rejects the minority responses as likely resulting from performance errors. But what principled reason does the LA offer on behalf of this proposal? Why think that the majority response is the considered response and that the MMR involve performance errors? Mikhail's (2011) argument is contained (in full) in the following statement: "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'" (p. 110).

A provisional worry for this assumption is that it reveals the LA's distinction between competence and performance, as regards actual judgments, to be unprincipled. Remember that the LA first makes the distinction between competent judgments and judgments that result from performance errors by drawing on features common to Rawlsian considered judgments and competent linguistic judgments. Initially, Mikhail proposes that any judgment that is immediate, stable, stringent, certain and impartial, is a considered judgment. Yet, Mikhail has further criteria in mind. Considered judgments are also those favored by the majority. Now this criterion might seem arbitrary, but Mikhail suggests that linguists share it. So, *prima facie*, the analogy between linguistics and moral psychology motivates what might otherwise be an arbitrary criterion for distinguishing judgments that are competent from those that aren't. One might wonder whether or not linguists actually share this assumption. Mikhail (2011, p. 232-238) seems to suggest that this assumption is shared, but, if he is mistaken and this assumption is not shared, then it simply serves to strengthen my claim that the LA's assumption that the majority judgment is the competent judgment lacks adequate justification.

Granting, for the moment, Mikhail's suggestion that this assumption is shared, there still appears to be an important disanalogy between linguistics and moral philosophy – one which undermines the justification for the assumption that majority judgments are the considered judgments. In linguistics the

assumption that majority responses typically reveal competent judgments is only motivated by the presence of generally broad agreement among the acceptability judgments of a language's native speakers, and also broad agreement among linguists' grammaticality judgments. These two types of agreement can independently, or conjointly, justify the assumption that majority responses reveal competence.[25] In moral theory, there are few such convenient majorities. Further, there is marked disagreement even among a number of trolley cases (see Figure 2) critical for establishing the innate principles that the LA ascribes to our moral grammar.

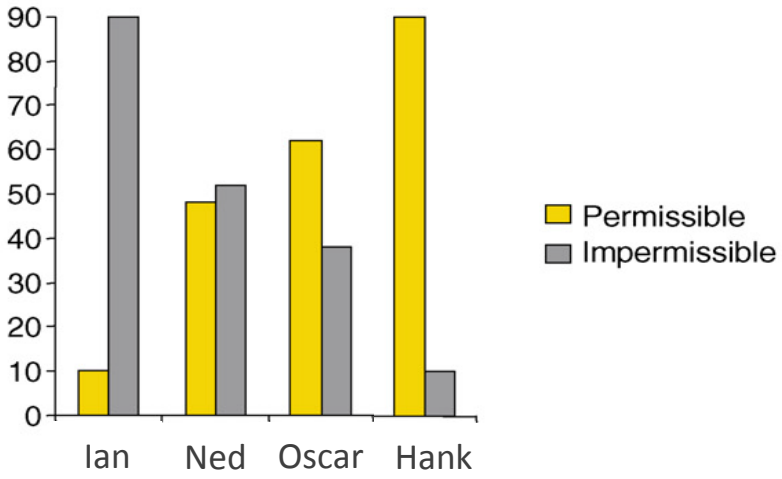


Figure 2. Rates of Permissibility/Impermissibility Judgments[26]

If this disanalogy is as significant as I've suggested, Mikhail has no compelling justification for assuming that, in every case, majority responses reflect our innate moral competence. If the LA defenders' claim that majority judgments are competent judgments lacks justification, then the LA's rejection of the MMR data seems similarly unjustified. If this rejection is unjustified, then there appears to be no principled way for the LA to distinguish between the permissibility and impermissibility judgments about Ned and Oscar's cases (*Loop Track*, and *Man-in-front*, respectively) –remember that 52% of people judge that Ned's

[25] Note that linguists also grammaticality judgments aren't the sole basis for linguistic theorizing (Mikhail, 2011, p. 235). Accounts of linguistic competence draw on data from studies of the brain, and studies of individuals with specific deficits. The LA is unlikely to draw on data of this sort, though, since such evidence is subordinate to the LA's account of moral competence, which is methodologically and logically prior in moral theorizing (according to Mikhail).
[26] Modified from Mikhail, 2007.

switch throwing (which involves using the person on the side track as a means of slowing the train in order to save five lives) is *impermissible*, while 62% of people judge that Oscar's switch throwing (which produces the side effect of killing the person on the side track as a result of directing the runaway train to strike the large object on the side track such that the train slows, saving five lives) is *permissible*.

Lacking a way to distinguish between the responses to these cases has significant implications for the LA's account of our moral grammar's contents, since the majority responses to these cases largely underwrite the plausibility of Mikhail's claim that the PDE is a constituent of our universal moral grammar. Ned's throwing the switch involves causing harm (death) to the man on the side track as a *means* to prevent the greater harm of five deaths on the main track. Oscar's throwing the switch also harms the man on the side track, but does so as the *side effect* of producing the good end of saving the five on the main track. The fact that Oscar's switching is judged permissible while Ned's isn't, and that the only factor that seems to distinguish the cases is whether or not the death of the man on the side track is a side effect or a means, is well-explained by the LA's claim that we possess the PDE as an innate deontic principle. Put another way, if the competent judgment is that Ned's switching is impermissible, and that Oscar's switching is permissible, then our innate moral principles nicely conform to the PDE. But notice that this is only true on this particular permutation (i.e., the competent judgments are that Ned's switching is impermissible and Oscar's is permissible). If, as I've suggested, there is no compelling, principled reason for the LA to exclude the MMR data, then the defender of the LA is unjustified in maintaining that these are the competent judgments. If these are not the competent judgments, then our considered judgments do not conform to the PDE, nor are they explained by claiming that the PDE is among our innate deontic principles. This outcome will additionally problematize other elements that the LA attributes to our innate system of deontic rules and principles, such as the prohibition against intentional battery, even if some cases (i.e., *Bystander* & *Footbridge*) still support such attributions.

For example, imagine a permutation (*Table 1)* in which the considered (competent) judgment is that Ned's switch throwing is permissible, and also that it is morally permissible for Oscar to throw the switch.

Table 1. Matrix of Permissibility/Impermissibility Judgments in *Loop Track* and *Man-in-Front*

Switching Status	Ned Permissible	Ned Impermissible
Oscar Permissible	-Intentional battery permitted -Intentional homicide permitted(?) -Harms as means and side effects permitted -PDE: Not Applicable	-Intentional battery forbidden -Intentional homicide forbidden -Harms as means forbidden, as side effect permitted -PDE: Supported
Oscar Impermissible	-Intentional battery permitted -Knowing battery forbidden -Harms as means permitted, as side effects forbidden -PDE: Not Applicable	-Intentional battery forbidden -Knowing battery forbidden -Harms forbidden as means and as side effects -PDE: Not Applicable

On this permutation it is plausible that there is no innate prohibition against intentional battery and (possibly) intentional homicide. Additionally, on this permutation, negative effects (the death of the man on the side track) are apparently permissible as means and also as side effects, which is clearly in violation of the PDE.[27]

Moving beyond the individual judgments about particular cases, consider how many people Mikhail would count as judging competently across the four cases I have been discussing (*Bystander*, *Man-in-Front*, *Loop Track*, and *Footbridge*). Though this has yet to be studied directly, one can look to the responses Mikhail has found (Table 2) and extrapolate likely patterns therefrom.

Table 2. Competent Judgments About Four Trolley Cases

Case	**Competent (Considered) Judgment**	**% Competent**	**% Incompetent**
Bystander (Hank)	Switching Permissible	90	10
Man-in-Front (Oscar)	Switching Permissible	62	38
Loop Track (Ned)	Switching Impermissible	52	48
Footbridge (Ian)	Pushing Impermissible	90	10

First, consider Pattern 1 (see Table 3). According to Mikhail's data, approximately 10% of subjects will judge (erroneously) that Ian's pushing the large man from the footbridge is morally permissible, killing

[27] Further reason for skepticism regarding the LA theory of competence comes from Cushman and Schwitzgebel (2011), who have found that trolley cases are significantly affected by which case is presented first, but they also found that experience with moral reflection (i.e., being a professional philosopher) had not notable mitigating impact on these order effects.

the large man but saving five men. It is intuitively plausible that one who judges that Ian's pushing is permissible will also judge that it is permissible for Ned, Oscar, and Hank to throw the switch, saving five but killing one person.

Table 3. Matrix of Competent/Incompetent Judgment Patterns

		Case Judgment – Switch Throwing/Pushing Permissible			
Pattern	**Percent Competent or Incompetent**	*Bystander*	*Man-in-Front*	*Loop Track*	*Footbridge*
1	10% Incompetent	Y	Y	Y	Y
2	38% Incompetent	Y	Y	Y	N
3	14% Competent	Y	Y	N	N
4	28% Incompetent	Y	N	N	N
5	10% Incompetent	N	N	N	N

Second, under Pattern 2, we find someone who does not think that Ian's pushing the large person is permitted, but judges that it is morally permissible for Ned to throw the switch. Once more it seems likely that any person who judges that Ned is morally permitted to throw the switch will also think that Oscar and Hank are permitted to do so. According to Mikhail's data, approximately 38% of respondents will conform to this pattern of errors (since 48% judge Ned's switching permissible, and 10% of this group have already been covered under Pattern 1).

Looking now at Pattern 5, we see a rather different pattern than the preceding two. Here, subjects judge that Hank is not permitted to throw the switch in *Bystander*. Those who judge accordingly are also likely to judge that Oscar and Ned are not permitted to throw the switch, and that Ian is not permitted to push the large man from the footbridge. Mikhail's findings suggest that about 10% of subjects will follow this pattern of performance errors.

Consider the remaining options. Under Pattern 4, subjects think that it is morally impermissible for Oscar to throw the switch. It is eminently plausible that subjects who judge this way will also judge that it is impermissible for Ned to throw the switch, and also that it is impermissible for Ian to push the large person from the Footbridge. If this prediction is correct, then 28% of people will match this pattern of performance

errors (beginning with the background 38% of subjects who judge that Oscar's throwing the switch to be impermissible, and subtracting the 10% already considered under performance error Pattern 5).[28]

In Pattern 3, which exactly matches Mikhail's account of our competent judgments, subjects judge that it is permissible for both Hank and Oscar to throw the switch, and that it is impermissible for Ian to push the large person from the footbridge and for Ned to throw the switch in *Loop Track*. In light of Mikhail's findings, it appears that only about 14% of subjects (subtracting the sum of the previous percentages from the possible number of subjects) will match this competent pattern across the four cases under consideration. In other words, approximately 86%, according to Mikhail's data are apt to make incompetent judgments about one or more of the four cases.

It is surprising to find that the LA's account of competence is such that people are very unlikely to count judge even a small set of test probes competently. This result presents the LA's defenders with a dilemma of sorts. Either some people are competent in judging some cases and not others (i.e., they are partially competent), or Mikhail cannot dismiss the minority responses without some further principled reason for doing so. Though the former option may be promising, I do not know how fruitful it might prove for the LA's defenders, especially since their attempt to model our innate moral rules presupposes that under favorable conditions (e.g., when faced with relatively simple moral stimuli) subjects' responses will likely reflect their innate moral competence, and the deontic rules and principles that partially constitute their moral grammar. If it turns out that most people are only partially competent, or, rather, that most people only make competent judgments about some of Mikhail's test probes, it is either because moral competence is more elusive than the LA presupposes, or that the test probes are of the wrong sort for revealing our innate competence. In either case, the LA must again explain why its rejection of the MMR data is justified, and, hence, why its way of drawing the competence-performance distinction is defensible.

---

[28] It might be that I am incorrect in assuming that subjects will be consistent in their judgments (e.g., that if one judges Ian's pushing the large man to be permissible then one will judge Ned's throwing the switch to be permissible too). If so, though, the LA's task is even more problematic than initially anticipated, since, as I've suggested previously, Mikhail presupposes that our moral judgments are systematic enough to be usefully modeled. If they are not, then the LA's central goal may not be achievable from the outset.

As I have hoped to show, the MMR evidence is problematic for the moral grammar model regardless of how the advocate of the LA chooses to respond to it. If the LA accommodates the evidence, it appears (at best) to attribute the wrong contents to our moral faculty. In contrast, by rejecting this evidence, the LA gives us reason to think that its account of moral competence is untenable, and that it is, ultimately, the wrong model of our moral endowment. Regardless of which option the advocate of the LA selects, we are left to conclude that the LA is descriptively inadequate. As I will now show, the LA fares no better with respect to the data regarding gender differences in judgments.

4.2. Reconsidering Gendered Judgments

Recall the discussion in §3.2. regarding gender differences in judgments about trolley cases. As I suggested, the LA does not currently succeed in dealing with these differences. Indeed, I suggest that the advocate of the LA is incapable of accepting the gender difference data by integrating it into her model of universal moral grammar, since, in accepting the gender difference data as reflective of our moral competence, the she is either forced to commit to an internally inconsistent or *ad hoc* account of the contents of our moral grammar or to give up on the idea of a *universal* human moral grammar.

Consider, for example, the differences between men's and women's judgments about the *Bystander* case, keeping in mind that women are less likely than men to judge that Hank's switch throwing is permissible. One option available to the LA is to accept the differences between women and men's judgments as part of our moral *competence*. As a result of this acceptance, it is possible that the LA would have to commit to the claim that women and men possess different innate deontic rules.[29]

For example, based on the evidence that women are less likely to judge throwing the switch permissible in *Bystander,* the LA might claim that women are less likely to accept the PDE (or are more likely to accept some other rule). Or, perhaps women possess an innate rule against intentional battery *and* mere knowing battery. On either condition, the contents of our universal moral grammar are contradictory –

[29] Remember that these conversion rules are what allow individuals to convert a perceived stimulus into a description of the temporal, causal, deontic, and intentional structure of that stimulus (for more see §2.1.1.).

contradictory in the sense that the LA is left to claim that an innate prohibition against knowing battery is a part of the universal moral grammar and also that this rule is not a part of the human moral grammar. There are two notable problems with this possibility. First, the possibility of our moral grammar possessing contradictory contents is a problem since the LA presupposes that our innate system of moral rules and principles is at least minimally coherent.[30] Second, avoiding the previous problem, one might suggest that men's moral grammar has a certain set of rules, and women's contains another set (and, presumably, there is substantial overlap between these sets). The chief problem with this strategy is that it amounts to a concession that there is no universal human moral grammar, and that we should divide our account of the moral faculty along gender lines. This division between men's rules and women's rules is a problem, since part of what motivates the LA is that it promises to offer a unified account of the moral faculty possessed by all humans.

That said, if Mikhail pursues this strategy, he is offering a theory of moral *competencies*, rather than universal human moral competence. Such a theory might be promising. In fact, it is possible that Mikhail would say that there are moral competencies, so that there is a male moral competence and a female moral competence, but this is not how he develops his theory (a theory of the human moral faculty), and this would be at odds with his Rawlsian commitments, especially since the moral grammar the LA is concerned with is supposed to be a system of deontic rules, principles, and concepts that is universal to human beings and is possessed by all normal adults.

In contrast to the previous strategy, the proponent of the LA might claim that men and women possess the same innate deontic rules, but apply different conversion rules to the same stimuli.[31] To put this claim more plainly: when exposed to the same trolley case, men and women represent the causal, temporal,

---

[30] Coherence of some minimal variety is a necessary condition for Mikhail's approach to modeling moral cognition and its hopes of bearing descriptively adequate fruit. To get a sense of why this is so, imagine trying to develop (as Mikhail does) the common conversion rules, that produce the common structural representations that, when combined with our innate deontic rules, give rise to our shared deontic verdicts *without* presupposing minimal coherence (among individuals and within individuals) across this process.

[31] Notice that the suggestion is that women and men implicitly *apply* different conversion rules to the same stimuli. Were we to claim that they simply *possess* different conversion rules *simpliciter*, then we would effectively be claiming that men and women have different moral competencies. Though I consider this claim later in this section, it is importantly different from the one I am examining here.

moral or intentional structure of the case differently.[32] If this is the case, it should not be surprising to find that they offer different deontic judgments about the case. This strategy seems promising for allowing the LA to explain away the observed differences in judgments across genders. However, taking this line reveals deeper challenges facing the LA and its account of competence.

Consider men and women's judgments about the *Loop Track* case (Mikhail, 2002a), where 45% of women and 54% of men claim that it is permissible for Ned to throw the switch, causing the train to hit and kill the man on the side track, but giving the five men on the main track time to escape. If we explain the data here by suggesting that men and women possess the same deontic rules, but perhaps possess different conversion rules, we are obliged to say that (since the majority modal response is taken to be the considered judgment) women must be applying a different conversion rule to the case than men are, such that they are mentally representing the case differently. This suggestion seems helpful to Mikhail, since it allows the LA to explain men and women's different judgments as the result of non-moral factors –a familiar and potentially fruitful approach to addressing moral disagreement.

Conveniently, in order to test how and where women's and men's mental representations of these cases differ, Mikhail (2007) has proposed that subjects be queried about the temporal, causal, moral, and intentional structures of the cases they are asked to judge. By strategically using 'by' or 'in order to' to link components of a case, Mikhail claims, we can recover some of the structure of individuals' representations. For example, if subjects respond in the affirmative to the question "Did Hank cause the train to turn by throwing the switch?", then we can infer, in part, the causal structure of subjects' representation of the *Bystander* case. Specifically, we can infer that subjects represent Hank's behavior as causing the train to turn, and that the switch throwing played a role in this process. The moral structure of these cases can be found by asking "In the *Bystander* case was a bad effect produced by the train running over and killing the person on the side track?" or "In the *Bystander* case was a bad effect prevented by the train being switched?" If, for example, subjects answer in the affirmative to the first question, then their representation of the moral structure of the case includes a negative valence to the death of the person on the side track. Similarly, we can arrive at a

[32] Recall that the moral structure involves the assignation of the status of 'good' or 'bad' to specific acts or outcomes.

picture of subjects' representation of the intentional structure of the case by asking "Did Hank throw the switch in order to turn the train?" Here, we can infer that subjects who answer in the affirmative are representing Hank's switch throwing as intentional, and that this switching was a means to achieve the intended effect of turning the train.

While subjects have not, to my knowledge, ever been systematically queried in the way Mikhail proposes, I take it that were we to ask questions of this sort to men and women as regards trolley cases where we find differences in judgments along gender lines, it would be unlikely that women's and men's responses would differ in the face of such queries. Thus, we have reason to be skeptical that men and women do differ as regards their representations of cases' temporal, causal, moral, or intentional structures.[33] Admittedly, my suspicion here is defeasible in light of further evidence, however, even if such future evidence does discover that there are consistent differences between men's and women's structural representations of these cases, these differences would have to track the differences in men's and women's moral verdicts in order for this evidence to provide aid and comfort to the LA's defenders.

An alternative option for the LA is to propose that men and women possess the same conversion rules and innate deontic rules and principles, but that they apply their deontic rules or principles differentially. There is a *prima facie* difficulty with claiming that men and women apply their deontic rules differentially, since this strategy effectively amounts to the suggestion that in the cases of gender disagreement, women or men are simply making performance errors –which amounts to a rejection of the gender difference evidence. To see why, recall the LA's depiction of how we go from stimulus to moral judgment. If two people are exposed to the same stimulus, and possess the same conversion rules, then both will produce the same structural description of the stimulus. If, in turn, both people possess the same structural descriptions and the same deontic rules or principles, then, both will automatically and unconsciously apply the same rule to the structural description, unless one person commits a performance error.

---

[33] One might complain that this test (i.e., using the above sorts of queries) are not adequate for determining how people mentally represent the structure of these trolley cases (or any other moral dilemmas, for that matter). I must express some sympathy for this complaint, but what matters is that there is some way to test how subjects actually structure these moral stimuli. Regardless of the test used, my suspicion is that once subjects' structural representations are studied, we will find that the observed data (e.g., gender difference data) is not best explained by subjects' structurally representing stimuli differently.

But let us grant that the LA's defenders can propose that gender differences are explained by instances of the differential application of deontic rules without effectively calling these instances performance errors. Even so, this claim is deeply problematic for the LA. Simply stated, this proposal forces the LA to declare that men and women possess the same moral competence, but that sometimes one gender applies a deontic rule to a case in a way that is at odds with the general trends for humans or for that gender. Rather than being self-contradictory, the LA now appears *ad hoc*, since the LA lacks the tools to explain why in most cases it asserts that we should see convergence in competent judgments, but here and there a judgment might still be competent while being the result of the non-standard application of a rule or principle.

I take it that accepting the gender difference data, and integrating such findings into its account of moral competence appears unpromising for the LA. But perhaps I've simply ignored the simplest explanation available to the LA as regards gender differences, since it seems that the LA can readily respond to these gender differences by explaining their origin. After all, any biologically-informed theory of moral judgment might explain the gender differences by suggesting that since men and women are biologically different, and psychologically different, it is unsurprising that there might be divergences in moral judgments between genders. Differences in the way women and men are socialized also plausibly contribute to these divergences. Thus, these gender differences may be the result of a dimorphic moral competence, or two moral competencies (one male, one female). But, as we have already seen, this response is unavailable to the defender of the LA if she wishes to hold to the notion that she is developing a *universal* moral grammar. But perhaps there is another way. Rather than adverting to dimorphic moral competences, the LA's defenders may still preserve a universal account of moral grammar by claiming that there are stereotypical male performance errors and female performance errors. What of this possibility? Perhaps the LA can explain away the gender difference data by invoking the competence-performance distinction.

Unfortunately, things are no better for the LA if it chooses to reject the gender difference data as the result of performance errors, even though this seems like a *prima facie* plausible strategy. Again, due to biological, psychological, or social factors, it might be that one gender is more apt to make performance

errors when considering particular cases. The problem with this reply is that the LA again owes us an explanation of which gender it is that is more apt to make these performance errors, and why. Further, it appears, at least provisionally, that the LA is committed to claiming that women are more prone to performance errors. This is due to Mikhail's "simplifying assumption" that the majority responses represent considered judgments, and therefore competent judgments. Yet women's and men's judgments both appear to possess the features required for counting as considered judgments, since they are seemingly stable, stringent, immediate, impartial, and certain (Mikhail, 2002a). If both genders' judgments possess the features that all considered, and thus competent, judgments possess, then it is particularly puzzling as to why Mikhail accepts only some judgments with these features.

Additionally, there are two other problems with endorsing the claim that women are significantly more prone to performance errors than men are: first, the LA still has offered no explanation as to why women make such errors more often; second (and more importantly), the plausibility of this claim rests merely on Mikhail's "simplifying assumption" that the prevailing modal responses reflect considered judgments. Not only does Mikhail offer no justification for this claim (aside from, perhaps, trading on its intuitive appeal), but this claim appears particularly strained when one considers the cases in which one response is favored by an especially narrow majority.[34] For example, in Mikhail (2002a) approximately 54% of men and 45% of women judged Ned's switch throwing (in *Loop Track*) permissible. In this case it is far from obvious that women's judgments are more likely to be the result of performance error than are men's, especially since it is also the case that approximately 55% of women judged Ned's switch throwing impermissible, as did 46% of men. As I argued in the previous section, it isn't at all clear that there is a principled way to determine which modal judgments are in the majority, much less who is in error, since, in order to do this, one must have a clear sense of what the norm or correct judgment is –unsurprisingly, in the realm of moral judgment, this is a particularly vexed question.

[34] It might be that the analogy with linguistics can justify Mikhail's assumption. However, the reason this assumption appears justifiable for linguists (i.e., that the majority response reflects the competent response) is that there is broad agreement on which grammaticality judgments are correct –however, there is no such agreement among our moral judgments. Thus, the justification available to linguists is plausibly unavailable to the LA's defenders. I consider this point further in §4.3

In short, my suggestion here is that rejecting the gender difference evidence is unprincipled, and provides us reason to think the LA is the wrong model of moral cognition. On the other hand, attempting to accommodate this evidence is similarly problematic for the LA, since doing so produces a contradictory or *ad hoc* account of our moral faculty's contents, motivating the conclusion that the moral grammar model ascribes the wrong contents to our moral faculty. As we will see in the next section, the worry is similar for a third line of evidence: the data regarding our judgments that favor kin, in-group members, and the young (henceforth, KIY data).

4.3. Accounting for Preferential Patterns

In §3.3 I discussed the trends in moral judgments consistent with KIY preferences. Since our innate moral sense is presumably evolved, one might think that the LA can and should grant that the features of our moral faculty will reflect the conditions and constraints under which that faculty evolved. So, the contents of our moral grammar should reasonably conform to the manifest KIY preferences reflected in studies of the sort conducted by Bleske-Rechek and colleagues (2010). But can the LA successfully accept these findings as reflective of our innate moral competence?

One possible way for the LA to accept these patterns of preference is to say that we possess distinctive conversion rules for cases involving kin, the young, or members of our social group. If this is so, then perhaps when we represent cases involving the young, kin, or in-group members in a way that is different than we do when cases involve unnamed actors. This suggestion seems promising, since these rules convert a stimulus into a representation of the stimulus' temporal, causal, deontic, and intentional structure.

If the LA's advocates take this line, it is important to understand how this might work for actual cases within the existing moral grammar model. Imagine again the *Loop Track* case featuring Ned, and one featuring Jayne. In Jayne's case there is an unnamed stranger on the side track, and there are five young children on the main track. If Jayne throws the switch, the train will strike the stranger and slow down prior to rejoining the main track, giving the five children time to escape. Compare Jayne's case to Ned's, in which all the persons in the *Loop Track* scenario are undescribed adult men. Though a modest majority of subjects

studied judge that Ned's throwing the switch is impermissible, it is extremely likely –if Bleske-Recheck et al. (2010), de Waal (2006), and Bloom (2011) are correct– that many more of us will find Jayne's switching permissible.

If judgments about these cases can differ, yet reflect our innate moral competence, then it must be that some structural or representational elements do not apply to cases involving strangers, but do apply to cases involving kin, in-group members, or the young. To discover whether or not this line has promise, as well as what these elements might be, I consider the stages of our structural representations in turn.

On Mikhail's view, the temporal structure that results from our conversion rules produces a simple accounting of the sequence of events, actions, and omissions, such that the identities of the agents or patients is irrelevant (see Appendix B, (c)). When we represent the causal structure of a case, we simply represent the relations between objects, agents, actions, states of affairs, the effects they produce, and the impact of those effects on the objects and patients described in the stimulus (see Appendix B, (d)). Again, this value-neutral structuring is and should be unaffected by the identities of agents and patients involved (cf. Knobe, 2010).

Following the conversion of a stimulus into its temporal and causal structure, the LA claims that the moral structure subsequently represents the good and bad effects involved in the stimulus (see Appendix B, (e)). Here, harmful effects are represented as 'bad', while positive effects are represented 'good' –bad effects that are negated (prevented) are also deemed 'good'. Once more, this structuring appears insensitive to the specific features of the agents and patients involved, instead evaluating the effects of actions and omissions by applying simple valences where appropriate. It may be possible that the presence of kin, the young, or in-group members might affect the number and location of good or bad effects, but it isn't clear how, since, as Mikhail proposes, the moral structure is applied to the effects of actions and omissions, rather than whom these effects befall. [35] So, for example, the moral structure involves an effect befalling a 'patient' as the result of the actions of an 'agent'. Individual identities do not enter into the calculus as structured by Mikhail. Perhaps, though, individual identities can enter into this structuring. Thus, the effect of killing a social in-

[35] This is not to suggest that 'good' and 'bad' cannot be graded notions on Mikhail's account. He allows, for example, that death is worse than bodily injury, which are both worse than property destruction, and that five deaths are worse than one. Yet, as I suggest, this grading is insensitive to the identities of (or demographic facts about) the agents and patients concerned.

group member is worse than killing a stranger, as is allowing harm to befall a genetic relative as opposed to a stranger. However, this way of measuring harms implies that not all persons are of equal moral worth, an implication, I go on to suggest, that Mikhail will not accept.

Returning to the conversion process, the LA proposes that, after converting a stimulus to into its moral structure, its intentional structure follows (see Appendix B, (f)). This structuring concerns the means, ends, and side effects of the actions isolated in the moral structure. For example, the 'bad effect' of throwing the switch in *Loop Track* is the train striking the stranger on the side track as a means to bring about the intended end of preventing the deaths of the five people on the tracks. Similarly, in *Bystander*, the 'good effect' of throwing the switch is a means to the intended end of preventing the five deaths, while the death of the person on the side track is a 'bad effect' that is a side effect of preventing the five deaths. But this representation fails to include any of the information that, on its own, would explain why a competent moral judgment about, say, Ned's case would differ from Jayne's. Nowhere in the intentional structure is there space for the relevance of the agents or patients identities. So, again, this layer of the structuring process holds little hope in helping the LA integrate the KIY preference data.

As a last step, the conversion rules culminate with the representation of the deontic structure of the stimulus (see Appendix B, (g)). This representation builds on the temporal, causal, moral, and intentional structure of the stimulus, adding in the innate deontic rules, principles, and concepts (e.g., "battery" or the PDE). Here, I think, there might be room for the LA to accommodate demonstrable preferences for the young, kin, and in-group members in our moral judgments. If so, then Mikhail may be able to respond to the KIY data by claiming that our conversion rules yield different structural representations depending on the identities of the agents and patients in a scenario. As Mikhail (2011) suggests, it is at this stage where "the comparative moral worth of the principal and collateral objects, must also be calculated and incorporated into these evaluations" (p. 173). If we grant that persons can be among these principal and collateral objects, then perhaps we have found the point at which our conversion rules produce varying outputs. So, if KIY have more moral worth than unnamed adult strangers, then the LA has a principled and internally consistent way to explain why competent moral judgments about structurally identical moral dilemmas can vary, even when

the only difference between moral dilemmas is the presence or absence of kin, the young, or in-group members.

Though this explanation provides Mikhail with a way to account for the preferences clearly manifested by the KIY data, we have good reason to think that the LA would not opt for this approach – namely, it is unlikely that it will choose to assign more moral worth to KIY than to strangers. One reason that the LA, as structured, is unlikely to propose that it is a feature of our innate moral competence to judge that KIY have more moral worth than strangers is that the account of moral competence under discussion here is fundamentally inspired by Rawls' *A Theory of Justice*. Not only is impartiality a feature of considered judgments in Rawls' sense, but Mikhail's account of moral competence gives every appearance of presupposing that the human moral faculty is and must be fair or impartial.[36] Here, Mikhail might object that all impartiality requires, on a conventional reading of the term, is that morally equivalent situations be treated the same way. Notice, though, how Rawls (1950) depicts impartiality:

> [A] man may be impartial and care a great deal. [His impartiality] depends on the cause of his concern, whether it is some private interest of his (gain in wealth and the like), or whether it is his concern for the achievement of a just resolution of conflicts. A person who wants to be just, is more likely to be impartial than one who does not. … Those interests which disqualify a judgment are only those which are readily admitted to work for unfairness. (pp. 54-55)

I take it that the crucially relevant notions here are "private interest", and judgments being disqualified if they work toward unfairness. First, "private interests", I take it, are the very interests operative in KIY judgments. For example, Jayne's interest in not harming his genetic relative (or, say, his romantic partner) is a private interest he holds. Or, more broadly, there is a "private interest" held in common between Jayne and his relative, but this particular interest does not extend to bystanders, or the five men on the main track. Second, KIY preferring judgments are paradigmatically unfair. So, inasmuch as judgments that do not promote fairness do not count as considered judgments, KIY judgments will not make the cut.

---

[36] Yet again, it is not critical for my case that Mikhail correctly interprets Rawls. What matters is that Mikhail commits to the view that he interprets Rawls to have, and that I am arguing against this view.

This inclination is revealed by Mikhail's response to an objection he considers while discussing the critiques of Rawls offered by Brandt (1979), Kagan (1989), and Lyons' (1975).[37] According to this "prejudice objection" it is plausible that the human moral faculty might actually be prejudiced or prejudicial (as regards race, religion, or nationality), and Rawls' account simply ignores this possibility – instead assuming that the moral faculty is entirely fair. The long history of human racism, genocide, bigotry, and violent nationalism lends some plausibility to this objection, as do the findings furnished by the Implicit Association Test (Greenwald et al., 2009). If the human moral faculty is innately biased or prejudiced in this way, then Rawls' account –and thus Mikhail's, too– makes significant, false assumptions about our innate moral endowment. Were this true, then the LA would almost certainly be descriptively inadequate, since it would fail to seek out its proper target for empirical study, illegitimately excluding a range of important data for which it must account.

In reply to the objection that our moral faculty might be prejudiced, Mikhail's response follows the same lines as Rawls'. Mikhail claims that if a moral judgment is merely the result of prejudice or bias, then it cannot count as a considered judgment, and, therefore, will not reflect the innate moral competence. To believe otherwise, he suggests, is to confuse mere prejudice with moral judgment, and is a distortion of both notions as ordinarily conceived (Mikhail, 2011, p. 262). To support this last distinction, Mikhail goes on to claim that, as with any theory in cognitive science, he is entitled to begin his research program from a set of reasonable assumptions. Further, as Rawls and Mikhail have objected, it is possible that the prejudice objection fundamentally relies on the stipulation that there is no such thing as a 'considered judgment' (since the objection claims that there is no meaningful difference between a biased or prejudicial judgments and considered judgments). If the prejudice objection does deny the distinction between considered and biased judgments, then Mikhail (2011) claims that the objection relies on a premature rejection of the LA's legitimate starting assumption that it can draw a principled distinction between prejudice and considered judgment, and

[37] Variations on this objection can be found in Daniels (1979), and Williams (1985).

that the former involve judgments "in which the moral capacities are distorted in some ways" (p. 259), while the latter are not similarly distorted.[38]

To illustrate this point, Mikhail presents his trolley cases again, but this time specifying the race, religion, or nationality of the persons on the tracks –e.g., a *Bystander* case in which there are five black people on the main track and a white person on the side track, or five Americans on the main track and one Asian on the side track. Mikhail suggests these sorts of details are irrelevant to moral judgment, and that considered moral judgments would not be affected by them.

He goes on to imagine presenting these scenarios (with details about race or nationality) to a group of children, and also presenting structurally identical scenarios, with no biasing details, to another group of children. If one found that the two groups of children produce differing verdicts, there are three possible explanations (Mikhail, 2011, pp. 259-261):

i. The children have the same innate moral principles, and the differences are due to performance errors, etc.
ii. The children do not have the same innate moral principles (thus the disagreement among response groups may be attributable to different moral competencies).
iii. The children do not possess any innate moral principles.

Though Mikhail favors explanation (i.), he grants that determining which explanation (i, ii, or iii) is correct will be an empirical matter, so he appears open to the possibility that our innate moral faculty is less impartial than he assumes for the purposes of theorizing.

In light of Mikhail's treatment of the prejudice objection, it is plausible that he would similarly reject the significance of KIY factors, as found in the evidence discussed above –labeling them as biases, rather than accepting that they can be accounted for by innately assigning less moral worth to strangers than to kin, in-group members, and the young.

[38] My claim isn't that there is no principled distinction between competent and erroneous moral judgments, but that this distinction is problematic when it demands that judgments incongruous with Mikhail's data set and model be labeled as performance errors. Since many judgments seem to closely resemble considered judgments, yet the LA rejects them as erroneous, it may be that the LA cannot capably distinguish cases where it is mistaken from cases where individuals make performance errors. Thus, the LA might be unfalsifiable, since it will label its predictive failures as performance errors.

Another reason to think Mikhail will not feel pressured to accept (as reflective of our moral competence) the KIY preferences is that the LA, as a theory of competence is prior to the sorts of considerations (e.g., evolutionary) that make KIY preferences compelling constituents of the human moral faculty. Specifically, Mikhail (2011) claims that a theory of moral competence is "logically prior" to questions about the acquisition and implementation of our moral faculty (p. 82). It is also putatively prior to questions about the evolution and neuroscience of moral cognition, meaning a clear and well-developed account of moral competence must be in place before one moves on to consider the evolutionary and neuroscientific dimensions of moral cognition.

The suggestion seems, on first consideration, to be a good one, since it is problematic to develop an evolutionary account of a particular trait unless one has a clear picture of what that trait is. Also, numerous disputes over moral nativism are partly due to disagreement over what features are possessed by the trait in question (i.e., the human moral faculty).[39] In another sense, though, this strategy is profoundly problematic, since it means that Mikhail does not, and need not, develop an account of our moral competence that is sensitive to evolutionary concerns and constraints. Those features of a completed theory of human moral cognition will come later, after the account of moral competence is relatively settled. As a result, Mikhail chooses to pass over evolutionary concerns, including the plausible proposal that the human moral endowment evolved in a way that was fundamentally linked to our social relations with kin or our social groups and in which reciprocity in social relations was critically relevant (de Waal, 2006; Trivers, 1971). In light of the LA's commitment to the logical priority of the theory of moral competence, I take it that Mikhail will feel little pressure to let such considerations (i.e., KIY) determine the LA's account of the moral worth of principle and collateral objects. If this is correct, then Mikhail's remaining option is to reject moral judgments consistent with KIY preferences as the result of performance errors.

Rejecting KIY preferring judgments as performance errors commits the LA to the view that these preferences reflect the distorted operations of our innate moral faculty. I suggest that this rejection is problematic for two reasons. First, as mentioned above, it is exceedingly plausible that our innate moral

[39] See Joyce (forthcoming) for an excellent discussion of this issue.

endowment is an evolved faculty, or constituted by such faculties. Among the most compelling hypotheses regarding the evolution of our moral capacities is that our moral faculty evolved in harmony with, or as an extension of, our phylogenetic ancestors' capacities and tendencies for cooperation and prosociality (Brosnan, 2010, forthcoming; de Waal & Lanting, 1997; Jaeggi et al., 2010; Sober & Wilson, 1998). If this hypothesis is even approximately correct, then it is likely that this faculty will reflect the contexts under which it evolved. In these contexts it is almost certain that our ancestors' cooperation and prosocial behavior was initially limited to kin, and social in-group members, so it is unsurprising that out moral faculty would reflect these limitations. Similarly, if concern for young conspecifics is generally adaptive among primates, then it is unsurprising to find our innate moral endowment in agreement with this adaptive tendency. If our innate moral faculty has evolved in a manner that is scaffolded by or extends our innate KIY preferences, then the LA's rejection of moral judgments that reflect these preferences is a mistake –a mistake, at least, inasmuch as Mikhail is committed to developing a descriptively adequate account of our moral faculty. Second, the KIY preferring judgments possess all or nearly all the features common to the considered judgments that Mikhail accepts as competent judgments: they are stable, stringent, immediate, and certain. If they possess all the features of considered judgments, then, as we have seen, KIY preferring judgments are a part of the "categorical data a descriptively adequate moral grammar must explain" (Mikhail, 2011, p. 110). Clearly, the defender of the LA cannot coherently offer a principled basis for granting that KIY preferring judgments are also impartial while simultaneously rejecting the KIY preference data, since to do so would be granting that such judgments possess the features sufficient for being considered judgments but still aren't considered judgments.

Instead, the LA's advocates might claim that KIY preferring judgments are not impartial, and that judgments that aren't impartial cannot be considered judgments. Of course, KIY preferences are not impartial in the everyday sense of the term. However, consider how Rawls (1950) understands impartiality as regards considered judgments. Considered judgments are impartial in that they are not the result of a desire for personal gain or intense "emotional duress" (Rawls, 1950, p. 53). Also, impartial judgments give due weight to the interests relevant to that judgment, and cannot be the result of inadvertently or willfully

ignoring the interests of those the judgment concerns. While I take it that KIY judgments are impartial inasmuch as they do not conspicuously result from a desire for personal gain or emotional duress, it is unclear whether or not they result from inadvertently (or willfully) ignoring interests relevant to the judgment. In virtue of the above discussion regarding the LA rejecting the distinctive moral worth of KIY as compared to strangers, it is likely that the LA's defenders would locate the *partiality* of KIY preferring judgments here: these judgments ignore interests (i.e., all persons are of equal moral worth/weight) that are relevant to any moral judgment. By ignoring these interests, KIY preferring judgments aren't impartial, and, therefore, do not count as considered judgments. Since any judgment that isn't a considered judgment isn't a competent judgment, then the KIY preferring judgments must count as performance errors, and so can be rejected as data for which the LA –as a theory of moral competence– need not account.

This reasoning is fine as far as it goes. However, Mikhail grants that it is an empirical matter as to whether or not his assumption that Rawlsian considered judgments reflect competent judgments –taken on for the purposes of developing a theory– is a good one. In light of evolutionary considerations and the KIY preference evidence, perhaps there are good, empirically well-motivated grounds for relinquishing this commitment to considered judgments as impartial judgments. It is also worth noting that giving up on this particular commitment is well advised on intuitive grounds. If only impartial judgments are competent judgments that reflect the undistorted operations of our innate moral faculty, and impartial judgments are perfectly fair (i.e., all individuals involved in moral judgments are of equal moral weight, and their interests should be considered equally), then the judgments that result from our innate moral faculty must be perfectly fair too. While we might demand that a justifiable moral system be perfectly fair or impartial, it is far from obvious that we should assume that our moral competence will also be perfectly fair or that the selective pressures that gave rise to our species distinctive capacities would have produced a completely impartial innate moral endowment. My parting suggestion here is that the hypothesis that KIY preferences are somehow a part of our innate moral endowment is more plausible and empirically better supported than is the LA's apparent assumption that our moral faculty is perfectly impartial.

If the above discussion is correct, then we have at least two reasons to think that the LA's rejection of the KIY preference data cannot be both principled and unproblematic. First, it is extremely plausible that our KIY preferences are a part of the contents of our innate moral faculty, particularly since it is likely that these preferences evolved prior to, and were instrumental in the evolution of our moral faculty. Second, the KIY preferring judgments appear to possess the features (stability, stringency, immediacy, and certainty) possessed by Rawlsian considered judgments, and, thus, competent judgments. If this is the case, then the LA's rejection of the KIY preference data is unprincipled. But, if there is any feature KIY judgments lack that considered judgments possess it is impartiality. However, the theoretical justification Mikhail provides for assuming that the human moral faculty may be perfectly impartial is empirically undermined by the very KIY data under discussion here. Additionally, as we have seen there is reason to think that the LA also cannot successfully account for the KIY preferring judgments by accepting these judgments as competent. So, the LA faces a dilemma as regards the KIY preference evidence: rejecting this evidence is problematic and suggests that the LA is the wrong model of our innate moral endowment, while accepting this evidence is also problematic for the LA, since it suggests that the LA attributes the wrong contents to our moral grammar.

4.4. Revisiting the Analogy's Dilemma

As I have shown, the LA cannot successfully account for the evidence first introduced in §3. Since this evidence is a part of the evidence that any descriptively adequate theory of moral cognition must consider, I conclude that the LA is descriptively inadequate. This descriptive inadequacy is either the result of the moral grammar model being the wrong model of our moral endowment or attributing the wrong contents (e.g., rules and principles) to our moral faculty. And, while the moral grammar model could be rescued from descriptive inadequacy by successfully accounting for the evidence I have presented, any attempt to embrace this evidence produces an internally inconsistent account of our moral grammar's contents, and any attempt to reject this evidence fundamentally relies on the LA's untenable and implausible account of moral competence, thus undermining the moral grammar model altogether.

Remember that for the LA to be descriptively adequate requires that it successfully accounts for more data than other theories of moral cognition, produces correct predictions as regards moral judgments, and accurately describes its target – the human moral faculty. Though I have not emphasized this second requirement, at this point it should be clear that the LA won't successfully satisfy the first and third. Specifically, since the LA is incapable of successfully accounting for a range of data that any theory of moral cognition must consider, it is likely that the LA does not accurately describe its target. While I haven't directly compared the LA to other theories of moral cognition, the only condition under which the LA will count as descriptively adequate is if it turns out that all other theories of moral cognition are even more incapable than the LA in accounting for the relevant data. It strikes me as profoundly implausible that the LA, as Mikhail presents it, will emerge as the "best of the worst" accounts of moral cognition. However, even if it were the case that the LA fares no worse than other theories of moral cognition, we have reason to be skeptical that the LA accurately describes its target. Lastly, in light of the LA's inability to offer an adequate response to the data I have discussed, as well as the justifiable skepticism I've expressed as regards its success in describing the human moral faculty, it is reasonable to conclude that the LA is unlikely to produce correct predictions about the general data which a general account of moral cognition must consider.

4.5. The Analogy's Aftermath

As I have observed, the LA fundamentally rests on an untenable account of moral competence. This concern extends well beyond my discussion of the LA, however, since philosophers and psychologists alike often invoke some version of the competence-performance distinction in studying moral judgment (e.g., Greene, 2008; Sinnott-Armstrong, 2010). Any theory of moral cognition that hopes to distinguish competent judgments from performance errors faces challenges similar in kind (if not in detail) to those I have raised against the LA. Since the competence-performance distinction is fundamentally normative, it is a live possibility that in drawing this distinction a theorist might embed assumptions in her account of competence such that her account of competence incompatible with the data for which her theory must account.

This, I suspect, is precisely where Mikhail's model begins to go astray: by assuming that competent judgments are impartial, and by assuming that majority (considered) judgments are competent judgments *at the very outset*, the LA proves to be fundamentally at odds with the relevant, available data. Avoiding this pitfall is especially challenging in moral psychology. In other domains of inquiry, theorists typically inform their accounts of competence (and performance errors) by considering the relevant norms. However, as I indicated previously, the 'norms' are far less clear cut (and often fraught) in moral psychology, such that they may not prove reliable guides to moral competence. Norms are, of course, not the only means by which competence and performance are distinguished in cognitive science. Once one has a developed theory of the cognitive architecture in which a faculty is embedded, one can also begin to distinguish examples of competence from errors in performance. In moral psychology, we arguably lack a sufficiently developed account of the cognitive architecture in which our moral faculty is embedded. If this suggestion is correct, then attempts to distinguish moral competence from performance errors in moral judgment may be premature. At present, then, it may be advisable to refrain from making substantive assumptions regarding moral competence, instead focusing keenly on the cognitive architecture that subserves moral cognition, and, perhaps, attempting to discover which moral 'norms', if any, might helpfully inform future accounts of moral competence.

5. CONCLUSION

The Linguistic Analogy offers us a novel and ambitious way to explore human moral cognition. It is particularly laudable is its engagement with computational theory (e.g., detailing the structural descriptions involved in the process of making moral judgments). With its conceptual sophistication and rigor, the LA raises the bar for any account of moral cognition that aspires to descriptive adequacy. However, the LA fails to give a satisfactory account of significant and broad classes of moral judgments and this failure poses an insoluble dilemma for the LA. Absent a solution to this dilemma, there is little reason to accept Mikhail's Linguistic Analogy. As I have tried to show, no such solution is forthcoming. But that needn't prevent us from looking to the LA for valuable lessons about modeling moral cognition, even if those lessons are largely cautionary.

REFERENCES

Bleske-Rechek, A., Nelson, L. A., Baker, J. P., Remiker, M. W., & Brandt, S. J. (2010). Moral decisions in the Trolley Problem: People save five over one unless the one is young, genetically related, or a romantic partner. *Journal of Social, Evolutionary, and Cultural Psychology, 4,* 115-127.

Bloom, P. (2011).Moral nativism and moral psychology. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil.* Washington, DC: American Psychological Association.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (2005) Universals of human nature. *Psychotherapy & Psychosomatics*, 74, 263–268.

Cushman, F. A., & Greene, J. D. (2011). Finding faults: How moral dilemmas illuminate cognitive structure. In J. Decety & J. T. Cacioppo (Eds.), *The handbook of social neuroscience*. New York, NY: Oxford University Press.

Cushman, F., Young, L., & Greene, J. D. (2010). Multi-system moral psychology. In J. Doris (Ed.), *The moral psychology handbook* (pp. 47-71). New York, NY: Oxford University Press.

de Waal, F. B. M. (1996). *Good natured: The origins of right and wrong in humans and other animals*. Cambridge, MA: Harvard University Press.

de Waal, F. B. M. (2006). *Primates and philosophers: How morality evolved*. Princeton, NJ: Princeton University Press.

de Waal, F. B. M., & Lanting, F. (1997). *Bonobo: The forgotten ape*. Berkeley, CA: University of California Press.

Doris, J. (2002). *Lack of character*. Cambridge, UK: Cambridge.

Dworkin, R. (1973). The original position. In N. Daniels (Ed.) (1989) *Reading Rawls: Critical studies on Rawls' "A Theory of Justice"* (pp. 16-53). Stanford, CA: Stanford University Press.

Dwyer, S. (1999). Moral competence. In K . Murasugi & R . Stainton (Eds.), *Philosophy and Linguistics* (pp. 169-187). Boulder, CO: Westview Press.

Dwyer, S. (2006). How good is the linguistic analogy? In P. Carruthers, S. Laurence & S. Stich (Eds.) *The innate mind vol. II: Culture and cognition* (pp. 237-256). New York, NY: Oxford University Press.

Dwyer, S. (2008). How not to argue that morality isn't innate: Comments on Prinz. In W. Sinnott-Armstrong (Ed.), *Moral psychology, vol. 1, the evolution of morality: Adaptations and innateness* (pp. 407-418). Cambridge, MA: MIT Press.

Dwyer, S. (2009). Moral dumbfounding and the linguistic analogy: Methodological implications for the study of moral judgment. *Mind & Language, 24*(3), 274–296.

Dwyer, S., Heubner, B., and Hauser, M. (2010). The linguistic analogy: Motivations, results and speculations. *Topics in Cognitive Science*, 2, 486-510.

Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã. *Current Anthropology, 46,* 621-647.

Fischer, J. M., & Ravizza, M. (Eds.). (1992). *Ethics: problems and principles*. Fort Worth, TX: Harcourt Brace Jovanovich.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. In J. M. Fischer & M. Ravizza (Eds.) (1992), *Ethics: Problems and Principles* (pp. 60-67). Fort Worth: Harcourt Brace Jovanovich.

Gilligan, C. & Attanucci, J. (1988). Two moral orientations: Gender differences and similarities. *Merrill-Palmer Quarterly, 34,* 223-237.

Goodman, N. (1983/1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Greene, J. D. (2008) The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.) *Moral psychology, vol. 3, the neuroscience of morality: Emotion, disease, and development (*pp. 35-79). Cambridge: MIT Press.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364-371.

Greene, J. D. et al. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44,* 389-400.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). Understanding and using the implicit association test III: Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17-41.

Griffiths, P. E. (2002). What is innateness? *The Monist, 85*(1), 70-85.

Grotius, H. (1925) *On the law of war and peace.* (Kelsey, F. W., Trans.) Oxford, UK: Clarendon Press. (Original work published 1625).

Haidt, J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108,* 814–34.

Harman, G. (1965). Inference to the best explanation. *Philosophical Review*, *74,* 88-95.

Harman, G. (2008). Using a linguistic analogy to study morality. In W. Sinnott-Armstrong (Ed.) *Moral psychology, vol. 3, the neuroscience of morality: Emotion, disease, and development (*pp. 345–351). Cambridge, MA: MIT Press.

Hauser, M. D. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. New York, NY: Harper Collins.

Hauser, M. D., Cushman, F., Young, L., Jin, R. K., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind and Language, 22,* 1-21.

Hauser, M. D., Cushman, F., & Young, L. (2008). Reviving Rawls' linguistic analaogy. In W. Sinnott-Armstrong (Ed.) *Moral psychology, vol. 2 the cognitive science of morality: Intuition and diversity* (pp. 107-143). Cambridge, MA: MIT Press.

Kant, I. (1993). *Critique of practical reason*. (Beck, L. W., Trans.) New York, NY: Macmillan. (Original work published 1788).

Leibniz, G. W. (1981). *New essays on human understanding*. (Remnant, P. & J. Bennet, Eds.) Cambridge, UK: Cambridge University Press. (Original work published 1705).

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

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"*. Ph.D. dissertation in Philosophy. Ithaca, NY: Cornell University.

Mikhail, J. (2002). Aspects of the theory of moral cognition: Investigating intuitive knowledge of the prohibitions of intentional battery and the principle of double effect. *Georgetown University Law Center Public Law & Legal Theory Working Paper No. 762385*.

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Science*, 11, 143–152.

Mikhail, J. (2008). Moral cognition and computational theory. In W. Sinnott-Armstrong (Ed.) *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development* (pp. 81-92). Cambridge: MIT Press.

Mikhail, J. (2009). Moral grammar and intuitive jurisprudence: A formal model of unconscious moral and legal knowledge. *Psychology of Learning and Motivation*, 50, 27-100.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge: Cambridge University Press.

Mikhail, J., Sorrentino, C., & Spelke, E. (1998). Toward a universal moral grammar. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. (p. 1250). Mahwah, NJ: Lawrence Erlbaum Associates.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). Opinion: the neural basis of human moral cognition. *Nature Reviews Neuroscience, 6,* 799–809.

Nagel, T. (1973). Rawls on justice. In N. Daniels (Ed.) (1989), *Reading Rawls: Critical Studies on Rawls' "A Theory of Justice"* (pp. 1-16). Stanford, CA: Stanford University Press.

Nichols, S. (2004). *Sentimental rules*. New York, NY: Oxford University Press.

Nichols, S. (2005). Innateness and moral psychology. In. P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Structure and content.* (pp. 353-369). New York, NY: Oxford University Press.

Petrinovich, L., O'Neill, P., Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality and Social Psychology, 64*(3), 467-478.

Petrinovich, L., O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology, 17,* 145-171.

Prinz, J. (2007). *The emotional construction of morals*. New York, NY: Oxford University Press.

Rawls, J. (1950). *A study in the grounds of ethical knowledge: Considered with reference to judgments on the moral worth of character*. Ph.D. dissertation in Philosophy. Princeton, NJ: Princeton University.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Roedder, E. & G. Harman. (2010). Linguistics and moral theory. In. J. Doris (Ed.), *The moral psychology handbook*. (pp. 273-296). New York, NY: Oxford University Press.

Samuels, R. (2007). Is innateness a confused notion? In. P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind, vol. 3: Foundations and the future* (pp. 17-36). New York, NY: Oxford University Press.

Schwitzgebel, E. & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language, 27,* 135-153.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs, 1*(3), 229-243.

Singer, P. (1974). Sidgwick and reflective equilibrium, *Monist, 58,* 400-517.

Sinnott-Armstrong, A. (2010). Moral intuitionism meets empirical psychology. In T. Nadelhoffer, E. Nahmias, & S. Nichols (Eds.), *Moral psychology: Historical and contemporary readings* (pp. 373-387). New York, NY: Wiley-Blackwell.

Sripada, C. & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind, vol. 2: Culture and cognition* (pp. 280-301). New York, NY: Oxford University Press.

Sterelny, K. (2010). Moral nativism: A sceptical response. *Mind & Language*, 25(3), 279–297.

Thomson, J. J. (1985). The trolley problem. In J. M. Fischer & M. Ravizza (Eds.), *Ethics: problems and principles* (pp. 67-79). Fort Worth, TX: Harcourt Brace Jovanovich.

Thomson, J. J. (1986). *Rights, restitution, and risk*. Cambridge, MA: Harvard University Press.

Zamzow, J. & Nichols, S. (2009). Variations in ethical intuitions. *Philosophical Issues*, *19,* 368-388.

APPENDICES

Appendix A

Four Trolley Cases (from Mikhail 2000, 2002a, 2007, 2011)

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?

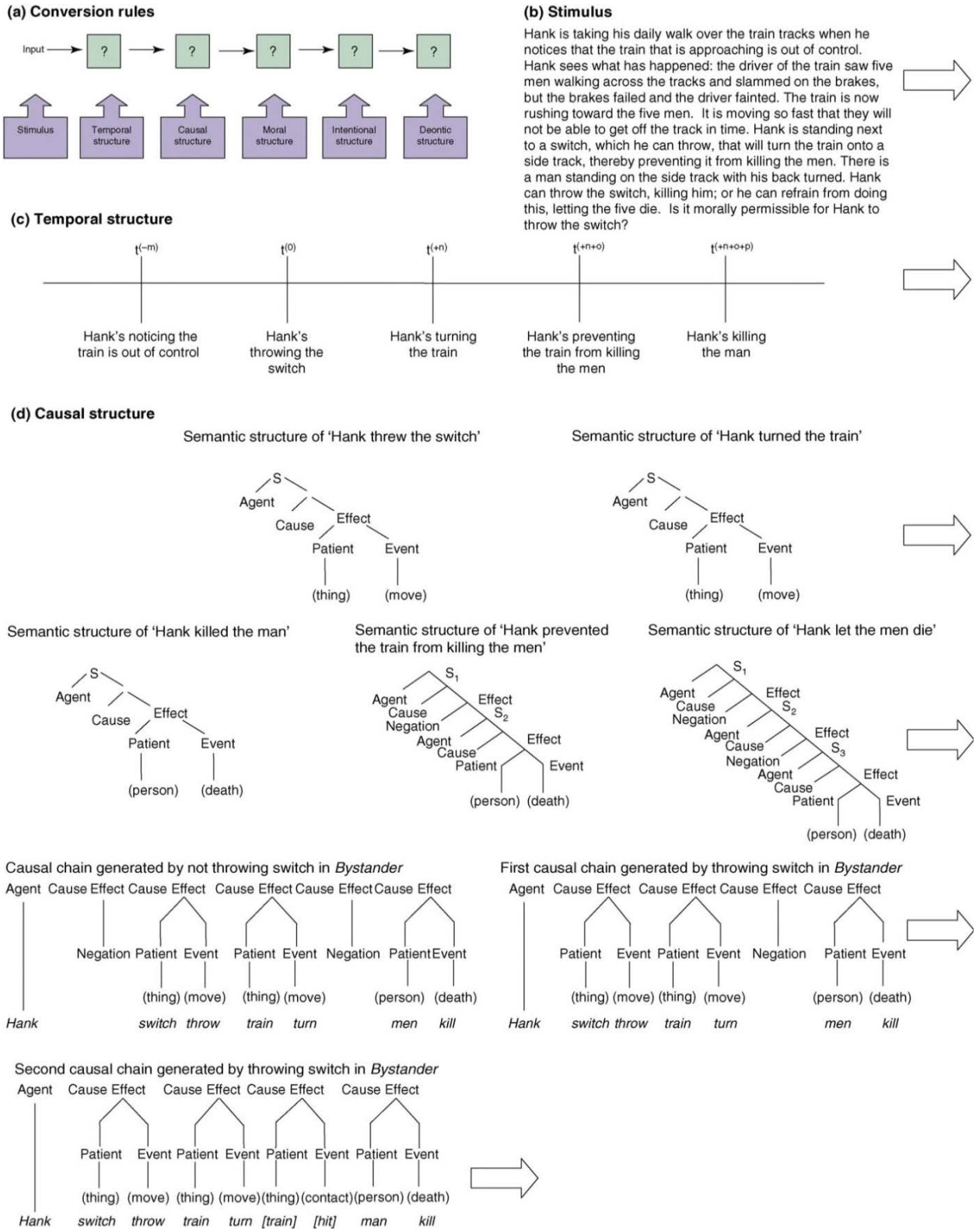
Footbridge: Ian is standing next to a heavy object, which he can throw onto the track in the path of the train, thereby preventing it from killing the men. The heavy object is a man, standing next to Ian with his back turned. Ian can throw the man, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Ian to throw the man?

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?

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?

Appendix B

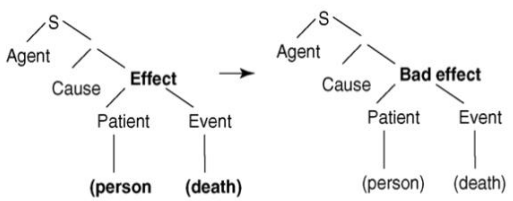
Computing Structural Descriptions (from Mikhail, 2007)



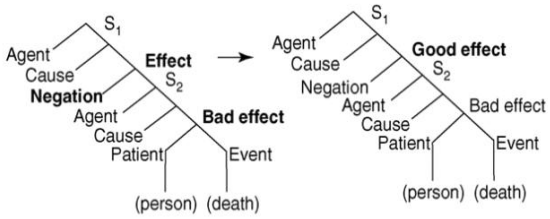
Computing Structural Descriptions (continued)

(e) Moral structure

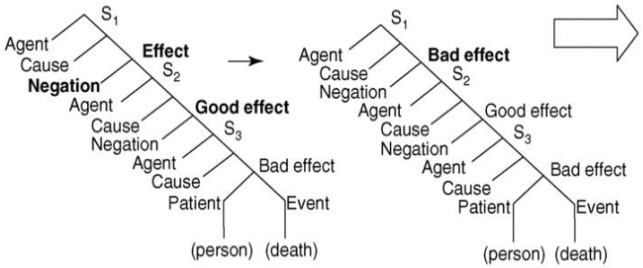
Moral transformation of 'Hank killed the man'



Moral transformation of 'Hank prevented the train from killing the men'

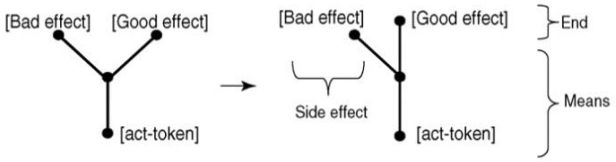


Moral transformation of 'Hank let the men die'

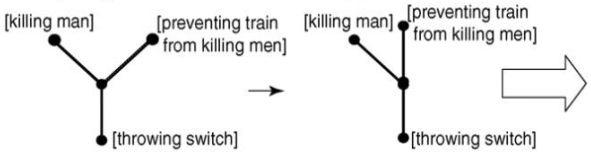


(f) Intentional Structure

Computing intentional structure of act with good and bad effects



Computing intentional structure in trolley problems



(g) Deontic structure

Partial derivation of representation of battery in *Footbridge*

1. [Ian's throwing the man at $t^{(0)}$]^{C}	
2. [Ian's throwing the man at $t^{(0)}$]	Given
3. [Ian throws the man at $t^{(0)}$]	2; Linguistic transformation
4. [Ian throws the man at $t^{(0)}$] ⊃ [Ian touches the man at $t^{(0)}$]	Analytic
5. [Ian touches the man at $t^{(0)}$]	3,4; Modus ponens
6. [The man has not expressly consented to be touched at $t^{(0)}$]	Given
7. [Ian throws the man at $t^{(0)}$] ⊃ [Ian kills the man at $t^{(+n)}$]	Given
8. [[Ian throws the man at $t^{(0)}$] ⊃ [Ian kills the man at $t^{(+n)}$]] ⊃ [the man would not consent to being touched at $t^{(0)}$, if asked]	Self-preservation principle
9. [The man would not consent to be touched at $t^{(0)}$, if asked]	7,8; Modus ponens
10. [Ian touches the man without his express or implied consent at $t^{(0)}$]	5,6,9
11. [Ian touches the man without his express or implied consent at $t^{(0)}$] ⊃ [Ian commits battery at $t^{(0)}$]	Definition of battery
12. [Ian commits battery at $t^{(0)}$]	10,11; Modus ponens
13. [Ian's committing battery at $t^{(0)}$]	Linguistic transformation