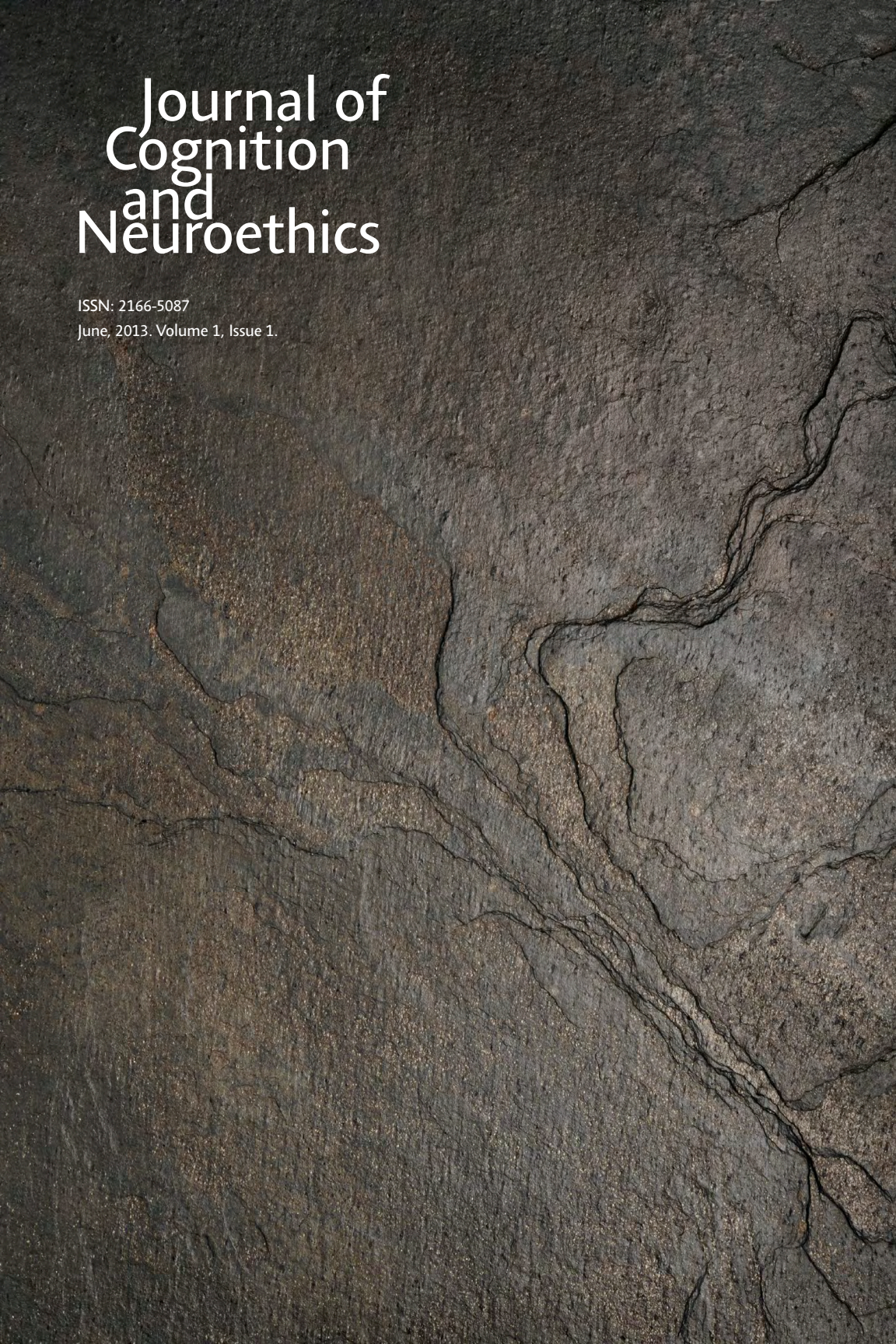
Journal of Cognition and Neuroethics

ISSN: 2166-5087
June, 2013. Volume 1, Issue 1.



Journal of Cognition and Neuroethics

Managing Editor
Jami L. Anderson

Production Editor
Zea Miller

Publication Details
Volume 1, Issue 1 was Digitally published June of 2013 from Flint, Michigan, under ISSN 2166-5087.

Center for Cognition and Neuroethics

The *Journal of Cognition and Neuroethics* is produced by the Center for Cognition and Neuroethics. For more on CCN or this journal, please visit cognethic.org.

University of Michigan-Flint
Philosophy Department
544 French Hall
303 East Kearsley Street
Flint, MI 48502-1950

Table of Contents

Introduction

Jami L. Anderson
University of Michigan-Flint

We are pleased to introduce the first issue of the *Journal of Cognition and Neuroethics* (JCN). The primary goal of JCN is to provide an open-access forum in which scholars from a wide variety of backgrounds, fields, professions and disciplines can contribute to discussions concerning cognition and neuroethics. Narrowly defined, cognition can be understood to refer to the mental processes that include attention, applying and accessing knowledge, memory, consciousness, self-awareness, producing and understanding language, all aspects of information processing, which includes learning, recalling, reasoning, problem solving, and decision making. More broadly defined, it can be understood to refer to any mental process, be it conscious, unconscious or artificial. Neuroethics has been defined as "the examination of what is right and wrong, good and bad about the treatment of, perfection of, or unwelcome invasion of and worrisome manipulation of the human brain."[1] These "treatments" and "intrusions" may arise from medical, neuro-pharmaceutical, psychological and psychiatric, or sociological research, as well as newly devised educational, healthcare, and legal policies and practices. Both cognition and neuroethics are relatively young fields, both cover a wide range of topics, and both are certain to remain of central importance for researchers, scholars and professionals for many years to come.

As technology advances and ever greater attention is directed at studying the human brain, far more questions are raised than are settled. Unfortunately, it is too often the case that discipline specific research is specialized and therefore difficult to understand (or even access) by those outside that field. The advisory board of the Center for Cognition and Neuroethics decided to create a space in which issues relevant to cognition and neuroethics will be discussed in a style and manner accessible to individuals from a variety of disciplines. Moreover, publishing JCN online means we are able to offer open access to its contents to scholars at any level, anywhere in the world where there is internet connection. Most importantly, we believe creating a space that rewards interdisciplinary discussions will foster both broader and deeper understandings of the issues discussed.

1. Safire, William. 2002. "Visions for a New Field of 'Neuroethics,'" Neuroethics: Mapping the Field Conference Proceedings, San Francisco, California, May 13–14.

Finally, to ensure that the review process JCN follows remains blind, our manuscript reviewers cannot be named but their time and efforts are very much appreciated. For more on the *Journal of Cognition and Neuroethics* or the Center for Cognition and Neuroethics, we invite you to visit our website at cognethic.org.

Jami L. Anderson
Managing Editor

Distinguishing Morality from Convention: Evidence for Nativism

Samuel R. Fullhart
Seattle Pacific University

Biography

Samuel Fullhart earned a B.A. in philosophy and economics and graduated with honors from Seattle Pacific University in 2013. He will be beginning law school next fall. His interests span a range of issues in philosophy of law, moral psychology, and ethics. He is particularly interested in (1) the nature of law and the relationship between law and morality; (2) theories of legal reasoning and adjudication; (3) moral reasoning as a species of practical reasoning; (4) topics at the intersection of moral reasoning and moral psychology; (5) the nature and extent of moral disagreement and the implications of disagreement for questions about the objectivity of morality; and, (6) the metaphysics of morality.

Acknowledgements

I would like to thank Leland Saunders and Thane Erickson for their detailed and incredibly helpful comments on several drafts of this paper. I would also like to thank an anonymous reviewer for the *Journal of Cognition and Neuroethics* for their helpful comments.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). June, 2013. Volume 1, Issue 1.

Citation

Fullhart, Samuel R. 2013. "Distinguishing Morality from Convention: Evidence for Nativism." *Journal of Cognition and Neuroethics* 1 (1): 1–37.

Abstract

Many moral psychologists, inspired by Noam Chomsky's work in linguistics, have attempted to offer Poverty of the Stimulus (PoS) arguments for moral nativism. In this essay, I consider one important line of evidence known as the "moral/conventional distinction," which nativists have used to support their PoS arguments. I argue that this distinction is not only real, but likely very abstract and sophisticated, such that the early recognition of this distinction in children across cultures constitutes an important line of evidence on which to ground a PoS argument for moral nativism. The paper is divided as follows: Section I briefly summarizes Chomsky's work and its integral role in the revival of nativism in cognitive science. Sections II and III review the large body of work, started by psychologist Elliot Turiel, on the moral/conventional distinction, examining Turiel's characterization of the distinction and presenting two possible models of the cognitive structures involved in moral judgments. Section IV considers a variety of criticisms of Turiel's view of the moral/conventional distinction. Section V argues that, though Turiel's view is vulnerable to several serious objections, there is still good reason to accept his claim that the moral/conventional distinction is a real distinction that children recognize from an early age. I suggest some ways in which Turiel's characterization should be modified in light of objections. Section VI presents several difficulties for empiricist explanations of the early acquisition of the moral/conventional distinction. These difficulties suggest that the distinction is the right kind of evidence for a PoS argument for moral nativism. Section VII concludes the paper.

Keywords: Moral Psychology, Nativism, Moral/Conventional Distinction

1. Nativism and the "Chomskian Revolution"

Noam Chomsky's pioneering work in linguistics in the 1950s and 60s has led to a nativist resurgence in the cognitive sciences. Before saying too much about Chomsky's work, let me first say a few things about nativism. Nativism is perhaps best understood when contrasted with empiricism. On the empiricist view, human beings enter the world as blank cognitive slates, equipped only with general purpose learning systems, which we then use to acquire all of our ideas, principles, concepts, knowledge, etc., through experience. Nativists deny that we begin life as blank slates and claim that we come equipped with *some* domain-specific cognitive structures.

Chomsky argued forcefully for a nativist view of language. He offered a radically new approach to linguistics, an approach which has been exported to many other areas of cognitive science, including moral psychology. Prior to Chomsky, if one was to ask what linguists study, the answer would have been, perhaps unsurprisingly, language. Linguists believed their job was to study languages, and it was thought that the way to do this was to study the linguistic behavior of speakers of these languages, looking for regularities, understanding the circumstances in which utterances are made, and observing the effects of such utterances.

Chomsky argued that this approach is entirely wrong because linguistic behavior is influenced by so many idiosyncratic factors that it would be futile to look for interesting regularities. He proposed instead that linguistics should be primarily concerned with studying the cognitive structures involved in linguistic behavior (e.g., speaking and understanding other speakers' utterances). A cognitive structure is something of an abstraction, but it is best thought of as a psychological process, or an organized set of psychological processes, structured to respond to various inputs and to produce certain outputs. It is also thought to contain (either representationally or structurally) information relating to the target domain, which Chomsky often loosely refers to as "knowledge." For example, a cognitive structure involved in language would include a set of psychological processes organized to respond to linguistic input (e.g., other speakers' utterances) and to produce linguistic output (e.g., the speaker's own utterances).

When Chomsky began his work, the received view in linguistics was that language acquisition is simply an induction problem, one of inferring from linguistic data what words mean and how they are strung together into sentences. On this picture, learning a language is simply a matter of applying general purpose learning mechanisms (e.g., mechanisms involved in induction) to the linguistic domain. However, Chomsky argued that some cognitive structures involved in language acquisition are "domain-specific." A structure is domain-specific if it is structured to respond solely to the input of a particular

domain and to produce output specific to that domain. For example, a structure designed specifically for language acquisition would only respond to linguistic data and not, perhaps, to non-human animal sounds.

Chomsky referred to the cognitive structure(s) exercised in the linguistic domain as the "Faculty of Language" (FL). It is difficult to say precisely how this faculty should be understood, but roughly, FL is whatever it is in the mind/brain[1] that enables humans to develop what Chomsky calls "linguistic competence."

Linguistic competence is best thought of as what a person knows when he or she knows his or her native language. It can be contrasted with linguistic "performance," i.e., the ways in which such knowledge is put to use in speaking and comprehending utterances. To illustrate the distinction, consider the sentence, "Bob thinks that Bill should stand up for himself." My knowledge that the reflexive pronoun "himself" must denote Bill is part of my linguistic competence as an English speaker, whereas my act of uttering this sentence would be a linguistic performance. Chomsky thought we could effectively study linguistic competence by gathering samples of native speakers' judgments about their languages, referred to as "acceptability judgments." In these studies, participants are presented with a variety of sentences and asked to judge which ones they think are grammatical and which are not.

Chomsky believed this information was important for three primary reasons. First, sampling acceptability judgments helps us study the degree of complexity of our grammars. If subtle differences in sentence structure seem to affect whether a sentence is judged to be grammatical, then we have reason to believe that the grammars participants are using in forming their judgments are highly complex. Second, by sampling the acceptability judgments of young children, we can see how early in childhood development grammars begin to form. Finally, we can study the acceptability judgments of speakers of a variety of languages to see how a speaker's linguistic background affects his or her grammatical judgments.

Although linguists have not reached a consensus on the strength of the nativist position, a large body of research suggests that: (1) individuals' grammars are highly complex and can only be specified in terms of abstract rules;[2] (2) grammars emerge early

1. I am using the term "mind/brain" so as to avoid any particular commitment to the relationship between the mind and the brain.

2. Linguists disagree about whether these rules are actually *in* the FL. At least in his early work, Chomsky seems to have thought that FL contains innately specified rules. Others, however, contend that while speakers' acceptability judgments exhibit regularities which are best described in terms of rules, the regularities are not

in the normal course of human development, around five or six; and, (3) the range of grammars found across the human species is highly constrained. Chomsky has argued that it is very unlikely that all normal children could develop such complex grammars in a short span of time using only general purpose learning systems and the linguistic data plausibly available to them. Given the linguistic input children receive, their output greatly exceeds what we would expect on empiricist assumptions. This kind of argument, that the output in a given domain exceeds the input given empiricist assumptions and that the output is therefore best explained by positing a domain-specific faculty, is referred to as a Poverty of the Stimulus (PoS) argument.

The influence of Chomsky's approach has extended far beyond linguistics. Some researchers in moral psychology have argued that there are significant parallels between language and morality and that a Chomskian approach to morality will prove fruitful. Furthermore, a number of theorists have argued that moral competence is best explained by positing an innate moral faculty.[3] For example, Dwyer (1999) makes use of research on the moral judgments of children to develop a PoS argument for moral nativism.

2. Evidence of an Innate Moral Faculty

It is not my intention here to try to assess how strong the analogy between language and morality is or to give any satisfyingly precise definition of what a "moral faculty" is supposed to be. When I say that the capacity to make moral judgments is "innate," I just mean that it is not acquired entirely from experience. Much more could be said about what is involved in this claim, but for the purposes of this paper, a broad characterization of the concept will suffice. Although conceptual questions are important in assessing the moral nativism debate, I think it is equally important to consider the lines of evidence that nativists have offered to support their claims, which will be my primary concern here.

Compared to linguists, moral psychologists have offered relatively few lines of evidence for moral nativism, but one potentially compelling piece of evidence that has been put forth relates to what is called the "moral/conventional" distinction. Although

in fact generated by rules contained in FL. These theorists argue that the regularities are rather the result of the structure of FL. I'll have a little more to say about this issue later with regard to moral rules.

3. There are sharp disagreements among moral psychologists about how "moral competence" should be understood, and whether and to what extent it is analogous to linguistic competence. For the purposes of this paper, I will assume that if the moral/conventional distinction (to be elaborated on in sections III and IV) is a real conceptual distinction, then the ability to distinguish between moral and conventional judgments is *part* of what is necessary for a human to be morally competent.

characterizations of this distinction can be traced all the way back to Aristotle, the formulation that is most relevant to the nativism debate is one which has been developed by an extensive research program started by the psychologist Elliot Turiel in the 1970s. Roughly, Turiel's claim is that at an early age, children consistently make judgments that fall into two distinct conceptual domains, the moral domain and the conventional domain. This pattern is thought to continue into adulthood and to be pan-cultural. While Turiel himself has not used this research to support nativist arguments, others such as Susan Dwyer (1999), John Mikhail (2000), and Gilbert Harman (2000) have argued that the moral/conventional distinction is not something that children could acquire from the moral "data" plausibly available to them. On their view, the moral/conventional distinction offers powerful evidence on which to ground a Poverty of the Moral Stimulus Argument.

In what follows, I will elaborate on the moral/conventional distinction as it has been developed by Turiel and his followers, and then consider a variety of criticisms it has received. I will argue that Turiel's characterization of the moral/conventional distinction presents several conceptual and empirical problems. However, I will develop a modified version of the Turiel characterization, one which I do not think is vulnerable to the objections raised against Tuirel's original formulation and which provides a better explanation of the evidence amassed by authors working within the Turiel tradition and authors who have been critical of it. In order for the moral/conventional distinction to provide support for nativism, it must be real,[4] and it must be such that it could not be acquired simply through general purpose learning systems and the evidence plausibly available to children. Although I do not claim to offer a fully satisfactory characterization of the distinction, I will argue that there are good reasons to believe the distinction is real, that it is highly abstract and sophisticated, and that it appears early in childhood development. For these reasons, I will conclude that the moral/conventional distinction provides the right kind of evidence with which to develop a Poverty of the Stimulus Argument for moral nativism.

3. Turiel and the Moral/Conventional Distinction

A. Children's Social Judgments

There is now a large body of research on what is referred to as the "moral/conventional

4. I will say more in section III C about what it could mean to say that the distinction is "real."

task."[5] This work has been motivated in large part by the ideas of psychologist Elliot Turiel, who proposed that our social judgments[6] fall into two distinct domains: the moral and the conventional. Many philosophers have argued that there is a distinction between morality and convention, but Turiel proposed that this distinction is something which all "normal" humans become sensitive to at an early age.

Although researchers have used the moral/conventional task for a variety of purposes, authors in the Turiel tradition have been particularly interested in studying two dimensions of judgment. First, they have examined what they call "criterion judgments." Criterion judgments can be thought of as participants' views about the status of their judgments. Turiel gives a list of some questions that have been used to elicit participants' criterion judgments: "whether the actions would be right or wrong in the absence of a rule or law, if the act would be all right if permitted by a person in authority (e.g. a teacher in a school context), whether an act would be all right if there were general agreement as to its acceptability, and whether the act would be all right if it were accepted by another group or culture" (Turiel 2000, 905). The second dimension concerns "justifications." To study this dimension of judgment, researchers ask participants questions about why they think a particular act is right or wrong, or okay or not okay. For instance, do they think an act is wrong because it causes someone harm, or because, say, it disrupts the social order?

In some surveys, children were asked to consider a situation in which their teacher tells them that they no longer need to raise their hands before talking in class. When asked if it would be okay to talk without raising their hands in such a scenario, almost all children said yes. But when told to imagine their teachers saying that it would be okay for them to hit their fellow schoolmates if they wanted to, children almost unanimously said that hitting would still not be okay. These studies have also presented questions such as, "Is it okay for children in other countries to hit their classmates?" Again, most participants answer that hitting is not okay. In one study, Nucci and Turiel asked Amish children to imagine a world in which God had made no rule against working on Sunday. Nearly all of the children said that, in such a situation, working on Sunday would be okay. However, when asked to consider a new scenario in which God did not forbid hitting

5. For an overview of this research, see Charles Helwig, Marie Tisak, and Elliot Turiel (1990).

6. I am unsure whether "social" is the right word to use, but Turiel seems to use it in order to make it clear that the moral/conventional distinction does not apply to certain other kinds of judgments (e.g., epistemic judgments or judgments about matters of "personal choice"). Judgments are social if they are about how we are to relate to other human beings.

others, 80 percent of the children said that hitting would still be wrong (Nucci and Turiel 1993).

These studies have now been performed on a wide range of participants, varying in age (toddlers to adults) as well as cultural and religious backgrounds. Even neurologically atypical participants, such as autistic individuals, have been given the moral/conventional task. Interestingly, the results have been much the same for all of these groups.[7] Participants' judgments seem to fall into two basic patterns, though as will soon become clear, articulating just what these two patterns of judgment are has proven quite difficult.

B. Domains of Judgment

Turiel has proposed that each pattern corresponds to a distinct domain, which he calls the "moral" and "conventional" domains. On this view, moral judgments are characterized as: (1) unconditionally obligatory; (2) generalizable; (3) generally regarded as more serious than conventional judgments; and, (4) relating to concepts of harm, welfare, fairness, and rights. Justifications for these judgments are given in terms of preventing harm, promoting welfare, fairness, and rights (Turiel 2000, 905). Conventional judgments, on the other hand, are thought to be: (1) contingent on authorities, rules, and existing social and cultural practices; (2) not generalizable but applicable only within the existing social arrangements to which they are connected; (3) generally regarded as less serious than moral judgments; and, (4) unrelated to concepts of harm, welfare, fairness, and rights. Justifications for conventional judgments are not given in terms of preventing harm or promoting welfare, fairness, and rights but rather are based on "understandings of social organization, including the role of authority, custom, and social coordination" (905). On Turiel's view, these features of moral and conventional judgments are pan-cultural.

To avoid caricaturing this view of the moral/conventional distinction, I should note that Turiel has been clear that he thinks conventional judgments can be regarded as very serious. He writes in "The Development of Morality" that convention "is not simply those residual regulations, to which there is little emotional intensity attached. They are uniformities and regularities of importance for social coordination" (905). But even the most important and deeply held conventions are still judged, on Turiel's view, by the criteria and justifications of the conventional domain. Moreover, Turiel has maintained that his research shows that participants generally exhibit stronger emotional responses to moral violations than to conventional violations.

7. One exception would be psychopaths, who seem to regard moral judgments in the way that most individuals regard conventional judgments. See Blair (1995).

C. Models of the Cognitive Architecture Underlying Judgments

If it is indeed the case that the moral/conventional distinction emerges early in the normal course of development for all children in every culture, what conclusions, if any, can we draw about the cognitive architecture underpinning these judgments? In this section, I will present one model designed to answer this question and another model which denies that the moral/conventional distinction is, in fact, a real distinction. I will briefly discuss their relevance to the nativism debate; and, in section V, I will consider the plausibility of each model.

Turiel and his followers have not been terribly clear on what their characterization of the moral/conventional distinction is supposed to tell us about the cognitive structures involved in moral judgments. However, these researchers have been clear on the point that the moral/conventional distinction is a distinction between two *conceptual domains*. For example, Nucci writes that morality and convention "are both part of the social order. Conceptually, however, they are not reducible to one another and are understood within distinct conceptual frameworks or domains" (Nucci 2001, 7). Furthermore, the differences between the two conceptual domains are thought to provide part of an explanation for participants' criterion judgments on the moral/conventional task.

In their critique of the moral/conventional distinction, Kelly and Stitch (2008) offer one model, referred to as the Moral/Conventional (M/C) model, designed to reflect these claims about the moral and conventional domains. On this model, each domain stores sets of *rules*,[8] as well as information about those rules. For example, rules stored in the moral domain will be authority independent[9] and generalizable, and justified on the basis of harm, welfare, fairness, and rights. Additionally, given Turiel's claim that all judgments in the moral domain pertain to harm, welfare, fairness, and rights, only rules relating to these issues will be contained in the moral domain. The conventional domain, on the other hand, will contain information indicating that the rules are contingent on authority, not generalizable, etc., and that these rules pertain to matters of social coordination.

8. Presumably, Kelly and Stitch claim that rules are stored in each domain in order to explain why moral and conventional judgments exhibit regularities. However, the fact that certain regularities can be explained by a rule does not show that there are in fact rules inside the cognitive system producing the regularities. Another possibility is that the regularities are the result of the structure of the system.

9. Kelly and Stitch replace "unconditionally obligatory" with "authority independence." Later in the paper, I will argue that the differences between these two terms are not terribly significant, though I will give some reasons for preferring the former term.

The authors argue that the only way to make sense of the idea that the two domains play a role in explaining the patterns of participants' responses on the moral/conventional task is to say that each domain is a functionally distinct component of the mind (357). In other words, Kelly and Stitch believe that if the distinction is to play an explanatory role, it must be more than just conceptually real but also *psychologically* real, i.e., it must be a distinction that reflects a division in cognitive structures. To be clear, this means that moral judgments and conventional judgments are the result of different psychological processes, organized and structured in different ways to respond to different kinds of inputs to produce different kinds of outputs. The distinction comes about, on this view, simply because conventional violations trigger the conventional judgment system, and moral violations trigger the moral judgment system.

The claim that the distinction must be psychological in order to explain participants' behavior is, I think, vulnerable to serious objections. A conceptual distinction does not require psychologically separate structures or processes to play a role in explaining behavior. The same psychological processes and organization can result in different judgments. For example, our visual systems are not likely divided into systems that respond to tables and different systems that respond to chairs. Rather, using the same system, we are able to make conceptually distinct judgments that the things we sit in are chairs and the things we sit at are tables. So, it seems to me that the moral/conventional distinction need only be conceptually real in order to explain participants' responses on the moral/conventional task.

Perhaps Kelly and Stitch think that if, as Turiel and his followers have claimed, the moral/conventional distinction appears cross culturally and early in development, it is more likely that the distinction is psychological. However, I do not see why the same psychological process, or organized set of psychological processes, could not be structured in such a way to reliably produce a conceptual distinction early in childhood development. If Chomsky and his followers are correct, many of the grammatical rules which children employ early in the course of normal development are specifiable in terms of abstract rules, and according to these theorists, the rules are the product of a set of psychological process dedicated to the linguistic domain. It seems then that even if the moral/conventional distinction appears in all cultures quite early in development, this fact does not provide any reason for preferring an explanation which posits a psychological distinction over one which posits only a conceptual distinction. This is not, of course, to say that we should prefer the view that the domains are only conceptually distinct, but only that Kelly and Stitch have offered no reason for thinking the domains are psychologically distinct. For the remainder of this paper, I will withhold judgment on

this issue, though I will consider the significance of each position for the nativism debate.

The alternative model proposed by Kelly and Stitch, and the one the authors actually endorse, is referred to as the Sripada and Stitch (S&S) model.[10] Instead of two cognitive mechanisms, one for each domain, their model posits an "Acquisition Mechanism" and an "Execution Mechanism." These mechanisms are thought to play an important role in the acquisition and implementation of "norms," which the authors define as "a theoretically important class of behavior-regulating social rules" (Kelly and Stitch 2008, 349). The authors claim that norms specify which behaviors are required or forbidden independently of any legal or social institution, that violations of norms result in a variety of punitive attitudes including anger and blame, that norms are present in every society, that they develop in all normal children between the ages of three and five, that the norms children ultimately "internalize" are the norms of their respective cultures, and that there is substantial diversity in the norms different cultures adopt.

In the S&S model, the Acquisition Mechanism somehow locates and internalizes the prevailing norms in the child's social environment. Once norms are acquired, the Execution Mechanism produces intrinsic motivation to comply with the norms and to punish violators (350). These features of the model are intended to explain the putative fact that norms are authority independent. Since participants are intrinsically motivated to comply with norms and to punish violators, norms do not need to be supported by authority in order for participants to feel bound by them, though Kelly and Stitch recognize that norms are often backed by authorities.

It is important not to view the S&S model as merely another way of characterizing the moral/conventional distinction because for Kelly and Stitch, there is no such distinction. The norm acquisition and execution mechanisms should not be understood as analogs to the moral domain and likewise, the cognitive structures that store all other kinds of rules should not be seen as corresponding to the conventional domain. The relevant distinction on this model is between rules with which humans are intrinsically motivated to comply (norms), and rules with which we are not. Although describing norms as "authority independent" might lead us to think of norms as similar to moral judgments, Kelly and Stitch are clear that norms need not be regarded as generalizable to other times and places, nor will they necessarily pertain to harm, welfare, fairness or rights. A person's norms will pertain to whichever norms are adopted by the culture in which he or she is raised, and Kelly and Stitch think that cultures can and do adopt norms to regulate a wide range of behaviors. It also appears that Kelly and Stitch may not think that norm violations will

10. For the original development of this model, see Sripada and Stitch (2006).

consistently be regarded as more serious than non-norm rule violations, since they write that some non-norm rules "might evoke an authority independent response" and also claim in the next sentence that non-norm rules "may evoke any pattern of answers on the seriousness…questions" (359).

Both the M/C model and the S&S model posit domain-specific cognitive structures and therefore reject the empiricist view that moral judgments are entirely the products of general purpose learning mechanisms. On the S&S model, though, the distinction between the moral and conventional domains does not exist, psychologically or conceptually, ruling out the possibility of that distinction providing evidence on which to ground a PoS argument. However, the bare claim that the moral/conventional distinction is a distinction between two conceptual domains is neutral with respect to the nativism debate. As stated earlier, one can accept that participants in the moral/conventional task rely on a conceptual distinction in making their judgments without also accepting that the distinction is psychologically real.

If the distinction is conceptual, but not psychological, we then need to ask why moral and conventional judgments get put in different conceptual domains. One explanation could be that the two domains develop through the workings of a single cognitive structure or set of structures dedicated to producing these kinds of judgments. An empiricist explanation worth considering is that these two conceptual domains are the product of general purpose learning systems.[11]

These considerations are important for the nativism debate because they help us see that the strength of the moral/conventional distinction as a line of evidence for nativism does not rest on its reflecting a psychological reality. In fact, the issue of whether it is psychological or just conceptual is an issue that should be decided after the nativist question is answered. If we find that the early emergence of the distinction is best explained by positing some kind of innate, domain specific structure(s), we then need to ask whether the distinction reflects a division in cognitive structures designed for different tasks, or a single structure designed, at least in part, to enable children to develop a conceptual distinction between the moral and conventional domains.

4. Problems with the Moral/Conventional Distinction

A. Conceptual Problems

Before we get into empirical data, there are conceptual problems with the moral/

11. I will have more to say about the prospects of an empiricist explanation in Section VI.

conventional distinction that deserve attention. First, although this point is not necessarily a criticism, it should be noted that the terms "unconditionally obligatory" and "generalizable" are not neutral terms in ethical theory. They are the product of a Kantian tradition and are absent from, for instance, virtue theories of ethics.

Furthermore, it is not easy to spell out just what these terms mean. The first, "unconditionally obligatory," involves two rather complicated concepts. The term "obligation" is a *theoretical* notion in that it is the product of theories constructed to provide accounts of right action (e.g., Kantianism or Utilitarianism). While most people are probably familiar with words such as "obligatory," "permissible," and "impermissible," it is unclear whether these words are ordinarily understood to mean anything more than just "okay" and "not okay."[12]

It is unclear from Turiel's writings whether he takes "obligatory" to simply mean okay or something stronger. However, the main idea he seems to be trying to convey in describing moral judgments as "unconditionally obligatory" is that they apply under all circumstances. So, on this formulation, the judgment that, say, lying is morally wrong would mean that it is not okay to lie under any circumstances. This idea is potentially problematic. Although we may judge that lying is wrong, it is not hard to think of conditions in which lying seems okay. It might be okay to lie to save someone's life or even just to save someone from embarrassment. For example, if a friend asks me if his or her new pair of pants makes him or her look fat, there does not seem to be anything wrong with me answering "no," even if I really believe the pants do make him or her look fat. Context seems important in determining the applicability of moral judgments.

I believe we can understand "unconditional" in a way that allows for moral judgments to be sensitive to context. We could, for instance, say that though people might make judgments such as "lying is wrong," certain implicit assumptions are built into these judgments. For example, one assumption might be that it is not wrong to lie if there are more compelling moral reasons to lie. If that is right, then one could say that under no conditions is it okay to lie if there are not at least equally strong moral reasons for lying. This kind of judgment can plausibly be thought of as unconditionally obligatory, but it is clear that context does matter for the applicability of such a judgment. So the idea that moral judgments are unconditionally obligatory, plausibly understood, allows for judgments to be context-sensitive.

The idea that moral judgments are generalizable also presents conceptual challenges. If this claim is to have any plausibility, it should not be taken to mean that judgments

12. This point is borrowed from Dwyer, unpublished manuscript.

generalize to *all* situations. For instance, it seems odd to say the judgment that it is wrong to steal generalizes to a situation in which a poor man steals a loaf of bread because it is the only way he can feed his starving family. However, judgments about the wrongness of stealing do seem appropriate in a situation in which a wealthy man steals a loaf of bread just to see if he can get away with it.

A way of explaining these intuitions is to say that moral judgments generalize only to cases similar in all *morally relevant* respects. To see if a judgment made in one context applies in another, we have to consider whether the two contexts are similar in ways that are morally relevant before determining whether the judgment can be applied in the latter case. For example, in the case of the two bread thieves, the wealthy man's action is judged as wrong, while the poor man's action is not, because their reasons for stealing the bread were different, and this difference is morally relevant to an assessment of the wrongness of the act. I do not think there is anything particularly controversial about this notion of generalizability, but it will become important later on in the paper when I consider some objections to the moral/conventional distinction.

I think one broader reason for why it is difficult to articulate just what is meant when terms such as "unconditionally obligatory" and "generalizable" are used to describe moral judgments is that it is far from clear what exactly moral judgments apply to. Some theorists maintain that moral judgments are fundamentally about *rules*. There are many variations of this view, but they generally hold that we have rules that specify which classes of actions are okay and which are not. To judge a particular action as "not okay" is to judge that the action falls within a more general class of actions that are prohibited by a given rule. Other theorists maintain that moral judgments are most fundamentally about *particular* actions, character qualities, states of affairs, etc. They may still think it is helpful to formulate rules as a way of systematizing our particular judgments, but the particular judgments are more fundamental. The rule-based account might seem more plausible given the regularities that we observe in moral judgments. However, as I discussed earlier, even if one accepts a rule-based account, one must still allow judgments to have some degree of context-sensitivity in order to capture ordinary intuitions. This requirement can be met by claiming either that exceptions are built into the rules (e.g., it is wrong to lie *unless* there are more morally compelling reasons to lie) or that our moral rules form some sort of hierarchy such that higher-level rules sometimes requires us to break lower-level rules. I will not try to settle whether moral judgments are fundamentally about particulars or about rules in this paper, but the issue will have some relevance to the discussion in section V.

A final challenge for Turiel's view, one which I think is a more serious problem, is that

it seems unduly restrictive to claim that moral judgments relate only to harm, welfare, fairness, and rights, and that justifications of moral judgments are only given in terms of these concepts. It is true that on the moral/conventional task, the only judgments that have been found consistently to exhibit the moral response pattern have related to these issues, but these studies presented participants with a limited range of moral issues. Much more evidence would be needed to infer that all moral judgments do, in fact, relate only to welfare, harm, fairness, and rights. As I will explain below, studies that have examined participants' responses to a wider range of issues have found that moral judgments are often about things having nothing to do with any of these concepts.

B. Empirical Problems

There are now many studies that have yielded results that seem to undermine various elements of Turiel's characterization of the moral/conventional distinction. One early study by Haidt, Koller, and Dias (1993) found that participants in low socio-economic groups in both the United States and Brazil judged certain acts, such as masturbating into a chicken carcass and cleaning one's toilet bowl with the national flag, to be seriously wrong. They also judged that these actions would be wrong even if they were performed in other times and places or in cultures where such practices were supported by authority. This result is important because neither act seems to involve harm, welfare, fairness, or rights, yet they are judged by criteria that, on Turiel's view, belong exclusively to the moral domain. It is also important because participants in higher socioeconomic groups were far less likely to judge these acts as moral issues. This result challenges Turiel's claim that across the human species, only certain kinds of acts are judged by moral criteria. It suggests rather that socioeconomic status can have a significant impact on the kinds of acts that are regarded as warranting moral or conventional judgments.

The philosopher Shaun Nichols has also conducted studies in which participants judged violations of certain etiquette rules by moral criteria. In one study of college students (Nichols 2002), Nichols found that many of the students regarded the act of spitting in one's cup before drinking as more seriously wrong than drinking tomato soup straight out of the bowl at a party. Many also said that the former action would still not be okay even if it was approved by custom or authority. Furthermore, over 60% of participants said the reason spitting in one's cup before drinking is wrong is that it is gross. Here, again, participants' criterion judgments would suggest that they judge the action to be morally wrong, but the act in question does not involve welfare, harm, fairness, or rights, and participants did not appeal to these concepts in their justifications.

Most of the studies that challenge Turiel's characterization of the moral/conventional

distinction suggest that actions that do not involve harm, welfare, fairness, or rights are often judged by criteria thought to belong exclusively to the moral domain. If these studies tell us anything, it is that a broader range of actions can evoke the moral response than Turiel and many of his followers have assumed. The studies do not provide any reason to think there is no distinction. In fact, to say that a broader range of actions can evoke the moral response presupposes a distinction between morality and conventions. Moreover, these studies do not challenge the claim that issues relating to harm, welfare, fairness and rights consistently evoke the moral response.

Nevertheless, some researchers have questioned whether the distinction is even conceptually real. Although few studies have directly investigated this claim, Kelly et al. (2007a) produced results which the authors argue provide preliminary grounds for skepticism about the moral/conventional distinction.[13]

Kelly et al. investigated whether *rules* pertaining to harm are always judged by the criteria of the moral domain. Two things deserve mentioning before I describe their survey and the conclusions the authors draw from it. First, Kelly et al. characterize the moral/conventional distinction as a distinction between rules, and not fundamentally as a distinction between judgments. Second, the authors' list of moral domain criteria differs in one respect from the list I have been using. Whereas I have taken the Turiel characterization to be committed to the claim that moral judgments are regarded as unconditionally obligatory, Kelly et al. describe moral judgments as being "authority independent" on the Turiel view. Although I will argue that these points are not terribly significant in evaluating Kelly et al.'s conclusions, it is worth noting these distinctive features of the authors' account of the moral domain. Now, let me describe their survey and the conclusions the authors' draw from the results.

Kelly et al. created an online survey designed to elicit participants' reactions to a number of scenarios involving one character or set of characters harming others. In

13. To be clear, Kelly et al. state that they are only challenging the claim that the moral/conventional distinction reflects a psychological distinction. For example, they seem to take Turiel and his followers to be claiming that the moral/conventional distinction is psychologically real (117), and in their conclusion, they write that their results suggest that "the moral/conventional task is not a good assay for the existence of a psychologically important distinction" (130). However, their position on this matter seems to reflect the claim made in Kelly and Stitch (2008) that only a psychological distinction could explain how participants consistently exhibit behavior which allegedly involves patterns associated with either the moral or conventional domain. If I am right in thinking that such behavior may only reflect a conceptual distinction, then Kelly et al. are effectively challenging the idea that the moral/conventional distinction reflects a genuine conceptual distinction since they hold that the behavior of participants does not, in fact, exhibit either the moral or conventional responses.

discussing their motivations for conducting the study, the authors argue for a need to study individuals' reactions to a broader range of moral violations than the ones presented in research on the moral/conventional task. They note that in those studies, the moral violations tend to be of the "schoolyard" variety (e.g., one child hitting another child). In their online survey, they asked participants about scenarios that involved other kinds of harms, varying the time and location of the harms to test whether moral rules (specifically harm rules) are generalizable, and varying whether or not the transgression had been sanctioned by an authority in order to see if moral rules are authority independent.

In each scenario, participants were asked whether it was okay for the character in the hypothetical scenario to engage in the act described, and then asked to rate the character's action on a scale from 0 to 9, with 0 being "not bad at all" and 9 being "very bad."

The first scenario set, referred to as Whipping/Temporal, was designed to see if moral rules relating to harm generalize. It goes as follows:

> Scenario 1
> Three hundred years ago, whipping was a common practice in most navies and on cargo ships. There were no laws against it, and almost everyone thought that whipping was an appropriate way to discipline sailors who disobeyed orders or were drunk on duty. Mr. Williams was an officer on a cargo ship 300 years ago. One night, while at sea, he found a sailor drunk at a time when the sailor should have been on watch. After the sailor sobered up, Williams punished the sailor by giving him 5 lashes with a whip.
>
> Scenario 2
> Mr. Adams is an officer on a large modern American cargo ship in 2004. One night, while at sea, he finds a sailor drunk at a time when the sailor should have been monitoring the radar screen. After the sailor sobers up, Adams punishes the sailor by giving him 5 lashes with a whip. (123)

A second version of this scenario set, Whipping/Authority, was presented to participants to test whether the sanctioning of whipping by an authority figure affected participants' judgments about whether whipping is okay. This version involves Mr. Adams (the officer on the modern American ship) again, and again in one of the scenarios he gives a drunken sailor five lashes. However, in the second scenario, Mr. Adams is told by

the captain of the cargo ship that "on this ship it is OK for officers to whip sailors."[14]

In total, Kelly et al. presented participants with eight different sets of "harm" scenarios.[15] Two of the sets were designed to test whether judgments about harm transgressions generalize to other times and places. These included Whipping/Temporal and Slavery/Spatio-temporal, which asked participants whether it was okay for ancient Greeks and Romans to own slaves and whether it was okay for people in the American South to own slaves 200 years ago. Six of the sets were designed to investigate whether harm norms are judged to be authority independent. These included Whipping/Authority, Military training/Authority, Prisoner abuse/Authority, Spanking/Authority, Hair pull/Authority, and Hitting/Authority. All of these scenarios involved a character inflicting some sort of harm on another character, and each included a version in which the harmful activity is endorsed by an authority figure and a scenario in which the authority figure has either said nothing regarding the harmful activity or said specifically that it is not okay in the particular context.

Results showed that participants were, in most cases, much more likely to judge characters' behaviors as okay and to give them a low ranking on the "how bad" scale in the versions of the scenarios in which either the behavior took place in a different time and social context or some authority figure had sanctioned the behavior. For instance, in Whipping/Temporal, 51 percent of participants said it was okay for Mr. Williams (the officer on the cargo ship 300 years ago) to whip one of his sailors as punishment for drunkenness, whereas only ten percent said it was okay for Mr. Adams (the officer on the modern American cargo ship) to whip a sailor for the same reason. In Spanking/Authority, there was also a significant difference in participants' answers. 44 percent of participants said that it was okay for the teacher to spank her student when spanking was not against the law and she had the principal's permission, but only five percent said that spanking was okay in the scenario where the teacher did not have permission and there was a law against spanking. Only in Slavery/Spatio-temporal was the difference not so stark. Seven percent of participants reported that slavery was okay 200 years ago in the American South (a surprisingly high number), and just 11 percent of participants said that slavery was okay in Greco-Roman times (126).

14. For all of the scenarios and results, Kelly et al. (2007b).

15. They also presented a ninth scenario designed to test Nichols' hypothesis that transgressions of rules which evoke negative affect are more likely to evoke the moral response pattern. Since it is not my intention to consider the viability of Nichols' hypothesis in this paper, I have excluded the results of that scenario from this discussion.

Kelly et al. admit that more research is needed before we can reach any definitive conclusions about the viability of Turiel's moral/conventional distinction as a hypothesis, but they do think that, along with other studies mentioned earlier regarding participants' judgments about issues unrelated to harm, welfare, fairness, or rights, there is good evidence for completely rejecting Turiel's characterization of the moral/conventional distinction, and even for rejecting the claim that the distinction is conceptually real altogether (130). Kelly et al. argue that if humans really do recognize a moral/conventional distinction that is anything like Turiel's characterization, we would expect participants in the authors' survey to judge harmful acts such as whipping and spanking as either okay in each scenario or not okay in each scenario. We should expect this result, they think, because acts involving harm clearly pertain to the moral domain (assuming for the sake of argument that there really is such a domain), so judgments about these acts should be regarded by participants as authority independent and generalizable. Since their findings suggest that people's judgments about the same harmful act (e.g., whipping) are often different in different contexts, they conclude that it is unlikely that the moral/conventional distinction reflects a real psychological or conceptual distinction. They suggest that further studies should reexamine the previous research on schoolyard transgressions to see why participants reported that these kinds of harm norms apply generally and hold independently of authority.

5. Whither Moral/Conventional Distinction?

I agree with Kelly et al. that there are problems, both conceptually and empirically, with some of the ways in which the Turiel tradition has characterized the moral/conventional distinction. However, I do not think the evidence put forth to challenge the moral/conventional distinction as it has been characterized gives any reason to think that the distinction is not even conceptually real, nor do I think that the evidence is sufficient for a wholesale rejection of Turiel's characterization. In this section, I will explain why I think criticisms of the moral/conventional distinction fall short and suggest some ways in which the Turiel characterization can be improved.

A. Getting Straight What is Being Distinguished

It seems to me that one of the biggest challenges facing both supporters and critics of Turiel's view is that it is not clear what exactly is being distinguished. This is hardly surprising, given the subject. It seems that any attempt to define morality or conventions is sure to raise further questions. I do not propose to even attempt to give a full analysis of either concept here, but these difficulties are worth raising because I suspect they have

much to do with the confusion. Turiel's writings are not always easy to interpret, but he seems to think that the moral/conventional distinction is a distinction between moral and conventional *judgments*. However, the distinction is often described in the literature as a distinction between two kinds of *rules*.

As I mentioned earlier, the M/C model which Kelly and Stitch develop and criticize offers a rule-based account; that is, to make a judgment about a particular act is to say that the act either complies with or transgresses a given rule. This view does imply a distinction between moral and conventional judgments, but it is more of a derivative distinction. A judgment is moral if it is a judgment about whether or not an act complies with a rule stored in the moral domain, and it is a conventional judgment if it is about whether an act complies with a rule stored in the conventional domain (Kelly and Stitch 2008, 357).

This articulation of the M/C model is reflected in Kelly et al. (2007a), in which the authors frequently talk about moral and conventional norms and rules. I think there are problems with the idea that judging that *x* is wrong means judging that *x* transgresses some rule. The idea has some appeal, since moral judgments do follow certain obvious regularities. For instance, if someone regularly judges instances of promise-breaking to be wrong, one way to explain this regularity is to say that the person is employing a rule that says something like "persons ought to keep their promises." However, even if judgments are just the applications of rules, it seems implausible to think that many people have rules such as "persons ought to keep promises" stored in their minds/brains because rules such as this one are often ignored or transgressed. More importantly, in some situations we might judge that it is okay, or even obligatory, to break a promise. For instance, if a person promises his friend that he will return the friend's rifle to him in a week, it seems wrong for the person to return the rifle a week later if he finds the friend in a drunken rage.

Someone could respond that we do have rules such as "persons ought to keep their promises" that we make use of in our moral judgments, but that we also have other rules about, say, not needlessly putting others' lives in danger, and that, furthermore, these rules can and do conflict with one another. In these cases, we may have higher-order rules that guide us in deciding which lower-order rules to break. It may also be that though we make judgments such as "persons ought to keep their promises," these rules have exceptions built into them (e.g., persons ought to keep their promises unless there are more morally compelling reasons to break them). There may be some truth to one or even both of these accounts, but it seems clear that if one is going to take the position that moral judgments are just the applications of rules to particular actions, situations,

etc., one is going to have to allow that the rules we use in our judgments are fairly sophisticated and may involve higher-order rules.

Second, I want to suggest from this discussion that, while the rule-based account of judgments is appealing because it gives a somewhat clear view of what judgments are, it is not obvious that our judgments (particularly our moral judgments) exhibit enough regularity to infer that they are just the applications of rules. Whatever clarity the rule-based account offers, it is somewhat lost once we try to account for the fact that almost all moral rules are sometimes suspended.

Finally, even if our moral judgments exhibit enough regularity to be specifiable in terms of rules (e.g., persons ought to keep their promises), that by itself does not tell us that the regularities are actually governed by rules in our cognitive structures, as both the M/C model and the S&S model suppose. To say that there are rules *in* our cognitive structures is to say that these structures literally contain nonconscious rules. However, it may be rather that the rules are in the structure of the cognitive systems, i.e., the arrangement of psychological processes produces a set of information transformation that is specifiable in terms of rules, but the system does not contain any such rules. To see the difference, it is helpful to think of the motherboard of a computer. The motherboard is arranged so that certain information transformations occur in a very specific way—a way that is specifiable in terms of rules. However, it is not the case that the motherboard contains such rules—what it is is a certain structure and arrangement for information transformation.

B. Why the Data does not Really Pose a Problem for the Distinction

As stated earlier, Kelly et al. argue that the results of their survey strongly suggest that a particular subset of moral rules, those relating to harm, do not consistently evoke the moral response pattern as Turiel and his followers have characterized it. That is, these rules are not consistently regarded as generalizable or authority independent. I believe the authors' conclusion is incorrect, but to see just where their analysis goes wrong, we will need to examine more carefully the scenario sets designed to test the generalizable hypothesis and the authority independent hypothesis.

One methodological problem with Kelly et al.'s study is that the survey does not make it clear in the instructions what is meant by "OK." If participants took "OK" to simply mean "OK relative to the standards of the time and place in which the action was performed," then it would make sense that many participants judged the actions by conventional domain criteria and the study's results would not pose any problem

for the moral/conventional distinction.[16] The descriptions of all eight scenarios provide information regarding whether the acts are considered okay relative to the specific conventions in place. For example, in Hair pull/Authority, the teacher says "at this school there is no rule against pulling hair. Anybody can pull someone's hair if they want to." The other descriptions contain similar phrases. This information might have led some participants to think that the questions were simply asking whether the acts were okay relative to the conventions of the contexts in which they occurred. If a significant number of participants interpreted the questions in this way, the fact that participants were more likely to judge acts as not okay when there were authorities forbidding them provides no evidence that harm violations do not evoke the moral response.

This explanation is particularly helpful, I think, in making sense of the results of Hair pull/Authority and Hitting/Authority. These scenario sets are quite similar to the schoolyard transgressions typical of moral/conventional task studies, yet participants' responses differed dramatically from the moral response pattern observed in those studies. Hair pull/Authority involved one eight-year-old girl pulling another eight-year-old girl's hair on the playground. Four percent of participants said it was okay without the teacher's permission, but 14 percent responded that pulling hair was okay when the teacher said it was okay and there was no school rule against it. Hitting/Authority involved one grade school boy hitting another. 14 percent said it was okay without the teacher's permission, whereas 53 percent responded that hitting was okay with the teacher's permission and in the absence of a school rule against it.

These scenario sets involve similar kinds of harms to the ones typically presented on moral/conventional task surveys, but Kelly et al.'s scenarios differ in that the descriptions do not make it clear how "OK" is to be understood. This omission contrasts with many of the studies conducted by Turiel and his followers, such as Nucci (1985) and Nucci and Turiel (1993). These studies examined whether Catholic, Protestant fundamentalist, and Jewish youths regard moral and conventional judgments differently, even conventional judgments that are supported by religious authorities such as the Catholic Church and/or Christian and Hebrew scriptures. In these studies, the question descriptions make it clear that participants are being asked whether religious authorities can change the status of moral judgments about typical moral issues such as harm and unfairness, not just in local contexts, but for everyone. By making the scope of the judgments clear, participants are required to consider whether the authorities really have jurisdiction over these issues. Not surprisingly, the results of these studies were in line with what the moral/conventional

16. For a similar criticism, see Rosas (2012, 6).

distinction predicts. Over 90 percent of Catholic adolescents said it would be wrong for religious authorities to suspend judgments against transgressions such as stealing and harming others, and over 80 percent of Protestant fundamentalist and Jewish adolescents said it would be wrong for religious authorities to suspend these moral judgments.

It might be asked whether distinguishing between senses of "OK" presupposes that people do in fact distinguish between moral and conventional judgments. However, one can accept that OK is sometimes regarded as having a wide scope and sometimes regarded as having a narrow scope without accepting that this difference reflects a genuine distinction between two conceptual (or perhaps psychological) domains. For example, one can believe the judgment "it is wrong to kill people just for fun" has a wide scope while the judgment "it is wrong to place one's fork on the right side of one's plate" has a narrow scope without any prior commitment to the moral/conventional distinction.

Even assuming most participants took OK to mean "morally OK" in the majority of Kelly et al.'s cases, the results of five of the scenario sets (Whipping/Temporal, Whipping/Authority, Military training/Authority, Prisoner abuse/Authority, and Spanking/Authority) do not seem problematic for the moral/conventional distinction if we consider some important differences between these scenarios and the ones typically presented to participants in moral/conventional task surveys. Although Kelly et al. are right to point out that these five scenarios are unique in that they do not involve schoolyard harm violations, they are also unique in that the harmful acts are not performed on innocent victims, as is typical of moral/conventional task studies (e.g., one student hitting another student just for the fun of it). In Whipping/Temporal, Whipping/Authority, and Spanking/Authority, the harmful acts are performed as punishment for wrongdoing. Additionally, in Whipping/Temporal, the description makes it clear that, in the Mr. Williams case, whipping was a common practice and "almost everyone thought whipping was an appropriate way to discipline sailors who…were drunk on duty" (Kelly et al. 2007a, 123). From this information, participants could have reasonably concluded that, unlike in the modern whipping case, the drunken sailor on Mr. Williams' cargo ship would have known that he would be whipped as punishment for being drunk on duty when he agreed to join the crew and that, given the views of the time, he would have regarded whipping as an appropriate punishment for such behavior.

The Prisoner abuse/Authority cases involve torture, not punishment, though the victim is described as a suspected terrorist and he is tortured in order to obtain information about future terrorist attacks. The Military training/Authority cases involve individuals training to become elite American military commandos. Participants likely assumed that the individuals freely chose to enroll in the training program and they

may have assumed further that the trainees were aware of the physical abuse training would involve. Moreover, the description states that "most people in the military believe that these simulated interrogations were helpful in preparing trainees for situations they might face later in their military careers" (Kelly et al. 2007b, 5). It is clear, then, that the abuse was performed for what was perceived to be the good of the trainees.

Since none of these cases involved the harming of innocent victims and many of the descriptions provided what participants might have considered to be justifiable reasons for committing the acts, it is likely that many participants regarded these acts as morally okay, at least when they were not forbidden by authorities. Although the Turiel characterization includes "concepts of harm" within the moral domain and claims that moral judgments are often justified in terms of preventing harm, it does not claim that all instances of harm will evoke the moral response pattern, so this conclusion is entirely compatible with the moral/conventional distinction. However, we still need to explain why many participants judged the acts as not okay when they were forbidden by authorities.

If my point about generalizabilty is on track, the fact that an action is judged as morally okay in one context does not guarantee that it will be judged as morally okay in another context, since there may be morally relevant differences between the contexts. Acts involving harming the innocent, stealing, and other typical moral transgressions are widely regarded as wrong in other times and places, and there may be few circumstances in which participants would regard these kinds of acts as okay. However, there are plausible explanations for why participants regarded the acts Kelly et al. presented as okay in certain contexts and not okay in other contexts.

Despite Kelly et al.'s claim that in the Whipping/Temporal case, "Clearly, many subjects think that whipping was OK 300 years ago though they do not think it is OK now" (Kelly et al. 2007a, 126), there are a number of differences between the two cases which participants may have regarded as morally relevant. As I said, participants would have likely assumed from the description that the sailor who was whipped 300 years ago would have known that whipping was the standard punishment for being drunk on duty before he decided to get drunk or to even serve on the ship. In the modern case, however, it could reasonably be assumed that the man who was whipped was unaware that he would be punished in such a manner when he decided to get drunk or when he decided to join the ship's crew.

In Whipping/Authority, Military training/Authority, Prisoner abuse/Authority, and Spanking/Authority, participants were more likely to say the acts were not okay when they were forbidden by authorities. These scenarios were designed to test whether

harm judgments are regarded as "authority independent" rather than "unconditionally obligatory." I attempted to argue earlier that we should not take "unconditionally obligatory" to mean that moral judgments hold under all circumstances. There may be important contextual assumptions built in, or, if the rule-based account is correct, the rules may be structured in some sort of hierarchy that makes them context-sensitive. If we consider these cases carefully, I do not think the results suggest that moral judgments are not regarded as unconditionally obligatory, nor do they suggest that moral judgments are not regarded as authority independent for that matter.

One explanation for the results is that, though participants think the acts are generally okay, they believe it is morally wrong to disobey authorities[17] in the absence of a compelling moral reason to do so. Many philosophers have held views along these lines, and regardless of the merits of this position, something like it may be held by many people outside of academic philosophy. Such a judgment, it should be noted, fits Turiel's other criterion judgments. It can be generalized to other times and places and may, with certain contextual assumptions, apply unconditionally. Furthermore, a judgment about our obligations to authorities could also be authority independent. Some people might think that, regardless of whether authorities say so, it is morally wrong to disobey authorities unless one has a compelling justification.

Participants might have believed the harmful acts had sufficient justification to make them generally okay, but thought these considerations were insufficient to justify disobedience. The moral/conventional distinction does not, of course, include violations of authorities' commands in the moral domain, but, as I've already argued, Turiel's characterization needs to be modified somewhat to allow that some people will moralize things having nothing to do with harm, welfare, fairness, or rights.

Even if participants do not hold this view, they still might have thought that disobedience would be wrong in Kelly et al.'s cases. Perhaps they thought disobedience would undermine the authority of the individuals who issued the orders or undermine the effectiveness of the organizations that adopted the rules. For example, a school in which teachers ignore school policies and the principal's explicit instructions might encounter serious difficulties in educating its students, and a military in which sergeants disregard orders from the Pentagon may be more likely to fail to uphold U.S. values in its efforts to defend the nation.

These points help us see that, even if we prefer to characterize moral judgments as "authority independent" rather than "unconditionally obligatory," Kelly et al.'s results are

17. There may, of course, be contextual assumptions—e.g., the authorities/laws are reasonably just.

not problematic for the moral/conventional distinction. If my interpretation of the data is broadly correct, participants do not judge acts as morally wrong simply because there is some general consensus that they are wrong or because an authority says so. Rather, these judgments are the consequences of beliefs about one's obligations to authorities and to the law, or beliefs about the negative consequences of disobedience. It seems unlikely, then, that participants think that for these harmful acts, whether or not they are morally okay is merely a matter of what authorities/laws say on the matter.

One additional possibility is that the moral ambiguities of the scenarios led participants to interpret "OK" in the narrow sense, in which case the results of these scenarios reflect the conventional response pattern. The scenarios' descriptions provided enough information to make it fairly easy to determine whether the acts were okay relative to the conventions of their specific contexts, so participants may have opted for the narrow interpretation in order to avoid making difficult moral judgments.

The results of the slavery cases are more difficult to explain away. Since slavery was considered okay by the conventions of both the American South and Ancient Greece and Rome, we would not expect more participants to regard slavery as okay in the latter context if they understood "OK" in a narrow sense. Moreover, both cases involve acts of harm against innocent victims.

Nevertheless, I do not think the slavery cases provide much evidence against the moral/conventional distinction. First, 11 percent of participants regarded slavery as okay in Ancient Greece and Rome, which is only four percent more than in the American South case. Second, 83.9 percent of participants identified as living in the United States (122), and the association of Southern slavery with the United States' history of racism and marginalization of African Americans may have influenced participants' judgments.

Although Kelly et al.'s study does not seem to pose a problem for the moral/conventional distinction, I think we can draw two lessons from Haidt and Nichol's work that will help us re-characterize the moral/conventional distinction in a way that gets us closer to a real distinction. First, however, it is worth emphasizing again that this work does not provide any evidence that certain issues, such as those involving the harming of an innocent victim, do not universally evoke the moral response pattern. Keeping this point in mind, the authors' work does nevertheless provide compelling evidence that a much wider range of issues than ones relating to harm, welfare, fairness, and rights can evoke the moral response pattern. Haidt has done considerable work on trying to identify all of the areas human cultures include as part of the moral domain. For example, Haidt and Bjorklund (2008) offer five sets of what they call "moral intuitions" which they claim are pan-cultural: harm/care, fairness/reciprocity, authority/respect, purity/sanctity,

and concerns about boundaries between in-group and out-group. Haidt and Bjorklund's view may or may not be correct; the point is that we need to broaden the range of acts, character qualities, etc. that can be included in the moral domain.

Second, we should recognize that an individual or group may moralize something that is not moralized by another individual or group. To give an example from Haidt, Koller, and Dias' work, low socioeconomic groups in Brazil and the U.S. may view the act of masturbating with a chicken carcass as an act that warrants a moral judgment, while high socioeconomic groups in Brazil and the U.S. may see such an act as wrong only from a conventional standpoint. It may very well be that, on the final analysis, we cannot give a comprehensive list of the kinds of things that can and cannot be included in the moral domain. Accepting this point does not undermine the moral/conventional distinction. In fact, to say that one individual views judgments about a particular act as moral judgments, while another views them as conventional judgments, is to accept that there is a distinction, at least conceptually.

C. Problems with the S&S Model

In order to adequately evaluate the plausibility of the moral/conventional distinction, we need to compare its strengths and weaknesses with the strengths and weaknesses of competing views. The S&S model is the only psychological model I am aware of that does not treat the moral/conventional distinction as real, so I will use it as the comparison model.[18]

One of the motivations behind the S&S model is that it is supposed to do a better job explaining why such a wide range of judgments exhibit something like Turiel's characterization of the moral response. On the S&S model, any norm prevailing in a given community can be internalized through the Acquisition Device. For this reason, Kelly and Stitch think the S&S model can explain why, for example, many people in low socioeconomic groups in the U.S. and Brazil judged washing the toilet bowl with the national flag to be wrong independent of what authorities say while people in high socioeconomic groups generally did not. However, this feature of the model does not seem to give it an advantage over the moral/conventional distinction because, as I argued, we can modify Turiel's characterization to allow for a divergence of judgments in borderline cases.

The other important component of the model, the Execution Device, ensures that

18. Since I have already discussed some of the problems with rule-based accounts of moral judgments, I set those problems for the S&S model (and the M/C model) aside here.

individuals will be intrinsically motivated to comply with norms and to punish violators. Kelly and Stitch emphasize that the intrinsic motivation attached to norms is what distinguishes them as a special class of social rules. They write that the intrinsic motivation to comply with norms and to punish violators "sharply distinguishes norms from other rules or information that may be mentally represented elsewhere in an agent's cognitive system" (Kelly and Stitch 2008, 350).

However, the model says little about non-norm rules and, at times, the authors seem to suggest that the distinction between norms and other rules is not so sharp. In fact, Kelly and Stitch suggest that rules outside of the norm database may exhibit many of the same features as rules in the database. The authors even write that non-norm rules "may evoke an authority independent response" (359). This concession seems to leave the model with little predictive power. If it is not at all clear on the S&S model what distinguishes norms from other rules, it is difficult to see how we might use it to make predictions about participants' judgments on moral/conventional task surveys.

Supporters of the S&S model would likely respond that, at the very least, the S&S model predicts that studies will not reveal a strong tendency for actions which are viewed as authority independent to also be viewed as generalizable,[19] nor will there be a great deal of uniformity in which kinds of actions are regarded as authority independent in different cultures. However, the evidence suggests neither claim is correct. Paradigm moral transgressions (e.g., hitting someone just for the fun of it) are consistently regarded by participants of different ages and cultural and religious backgrounds as wrong in other times and places and as unconditionally obligatory. Paradigm conventional transgressions (e.g., talking without raising one's hand in class) are consistently regarded as neither generalizable nor authority independent (or unconditionally obligatory). Such consistency is inconsistent with the S&S model. Even though the consistency breaks down somewhat in borderline cases, other transgressions that are sometimes regarded as both authority independent (or unconditionally obligatory) are also *typically* regarded as generalizable. Kelly et al. mention only one study, Nichols (2002), in which violations of certain etiquette norms were regarded by American college students as authority independent but not generalizable. But the most we can conclude from this result is that people do not always get the distinction right in borderline cases.

It seems then that the most serious problem with the S&S model is that it denies

19. Kelly and Stitch are clear that while norms are regarded on the S&S model as authority independent since participants allegedly have intrinsic motivation to comply with them, norms are not necessarily regarded as generalizable, even beyond one's immediate social context (359).

even the conceptual reality of a distinction which participants in moral/conventional task surveys reliably make in paradigm cases. That is not to say, though, that the M/C model Kelly and Stitch offer fully captures the distinction. As I have said, I do not think the fact that the distinction helps explain participants' behavior is evidence that it reflects a psychological distinction. Moreover, I think the M/C model wrongly assumes that simply because we can characterize moral and conventional judgments as the applications of rules means that our cognitive structures literally contain propositional rules.[20] There is also a good deal of evidence to suggest that, contrary to the M/C model, violations having nothing to do with harm, welfare, fairness or rights can be regarded as generalizable and unconditionally obligatory (or authority independent). Given these concerns with the M/C model and the concerns I raised earlier with the Turiel characterization, I will now try to sketch more of a positive account of how I think the moral/conventional distinction breaks down, one which I believe largely preserves the spirit of Turiel's characterization.

D. Why the Moral/Conventional Distinction is an Important Conceptual Distinction

If my criticisms of the Turiel characterization and the M/C model are on track, a good place to start in articulating a positive account of the moral/conventional distinction is to address why characterizing the distinction has proven so difficult. This problem could be seen as evidence that the distinction is rather abstract and sophisticated, or that it is vague and confused. I actually think both explanations are correct to an extent.

I have said that there is a considerable degree of consistency in participants' judgments (including children's) in "paradigm" moral and conventional cases, though much more variability in "borderline" cases. The paradigm moral cases, it seems, have to do with issues such as harming innocent victims, welfare, fairness, and rights. Paradigm conventional cases, according to Turiel, are based on "understandings of social organization, including the role of authority, custom, and social coordination." This description may not be as clear as one would like, but issues such as whether it is appropriate to speak in class without raising one's hand and whether church should be held on Sunday or Saturday seem to fit this category and there is ample evidence that these issues do consistently evoke the conventional response.

We have already examined a number of cases which may be called borderline cases since participants often differ on whether they regard them as moral or conventional issues. These include issues relating to disgust (e.g., masturbating with a chicken carcass, violations of etiquette norms), in-group out-group relations (washing one's toilet bowl

20. Indeed, as I have explained, even characterizing our judgments as the applications of rules is no simple task.

with one's national flag seems to belong in this category), perhaps certain harmful acts if the victims are not innocent and/or there are possible justifications for the acts, and perhaps even issues related to paradigm moral cases (e.g. Kelly et al.'s Hair pull/Authority and Hitting/Authority) if the questions are worded in a way that suggests to participants that they are being asked whether the acts comply with the conventions of the specific context.

The divergence of responses in these borderline cases suggests, I think, that the moral/conventional distinction may be somewhat vague. However, the near uniformity with which it is applied in paradigm cases suggests that it is not a hopelessly confused concept and, moreover, that it is an important conceptual distinction and an important line of evidence in the nativism debate. Regardless of how consistently the distinction is applied in borderline cases, the consistency of its application in paradigm cases, even among young children from a range of backgrounds, is something that requires an explanation. The extent to which empiricist and nativist accounts are successful on this front is significant in evaluating the debate between moral nativists and empiricists.

To say the distinction is applied consistently in paradigm cases requires a sufficiently clear characterization of the distinction so that we can determine whether it is in fact applied consistently. We have found that, for instance, the judgment that it is not okay to hit other students just for the fun of it is thought by most people to apply in other times and places, even hypothetical situations in which the teacher has said hitting is okay. The judgment that it is wrong to talk without raising one's hand does not apply if the teacher says so, nor does it apply in cultures with different customs regarding talking in class. It is, of course, difficult to say just what these differences are telling us about how the two judgments are distinguished, though it seems clear that they are regarded differently by participants.

Recall that for Turiel, criterion judgments for the moral domain are: (1) generalizable; (2) unconditionally obligatory; (3) generally regarded as more serious than conventional judgments; and, (4) relating to concepts of welfare, fairness, and rights. It will be helpful to consider each of these in developing a modified version of the distinction.

Although it is incorrect to say that only issues involving harming innocent victims, welfare, fairness, and rights evoke the moral response, these cases do reliably evoke that response, so the Turiel characterization is not completely wide off the mark on this point. To amend the original Turiel characterization, I think it would be best to say that issues relating to harming innocent victims, welfare, fairness, and rights consistently evoke the moral response, though other issues can and do evoke the moral response as well. The evidence amassed so far suggests that, except in certain cases, issues unrelated to these

concepts consistently evoke the conventional response. The exceptions I have considered involved things such as disgust, purity/sanctity, and in-group out-group relations. Nichols (2002 and 2004) and Haidt and Bjorklund (2008) offer some suggestions for why other issues sometimes evoke the moral response. A critical evaluation of these works is beyond the scope of this paper, though I am somewhat skeptical whether the authors' theories can account for the consistency of the distinction's application in paradigm cases and the inconsistency of its application in borderline cases.

I am less optimistic that we can preserve the idea that moral judgments are generally regarded as more serious. This claim is quite vague and it is unclear whether it tells us anything important about the moral and conventional domains. Moreover, it is easy to think of examples that do not fit the pattern. To borrow an example from Prinz (2008, p.384), one might judge public displays of nudity as more serious than eating the last available cookie without offering to share, but still regard the former case as merely a matter of convention and the latter as a moral issue.

The core of the distinction, in my view, has to do with which judgments are regarded as "generalizable" and "unconditionally obligatory." But as I argued in section IV A, to apply these terms to moral judgments is not to say that they are entirely context-independent.[21]

If my argument in section IV A is on track, we must take "generalizable" to mean that moral judgments generalize to cases similar in all morally relevant respects. One might argue that this definition blurs the distinction between the moral and conventional domains, because it seems one could likewise say that conventional judgments generalize to cases similar in all conventionally relevant respects. Answering this objection requires elaborating on the conditions under which moral and conventional judgments generalize.

This problem can best be approached by considering what it means to say moral judgments are unconditionally obligatory, because I think it is this feature of moral judgments that causes them to generalize in a way that is distinct from the way conventional judgments generalize. The important difference between moral and conventional judgments is that the conditions under which the latter judgments apply is determined *entirely* by what the authorities, rules, customs, etc., determine to be okay or not okay for a specific context. For example, the judgment that it is wrong to place one's

21. If what I have said about rules so far is on track, it should be clear that everything I say here is compatible with a rule-based account of moral judgments. As I have said, there may be contextual assumptions built into the rules or they may be arranged in some sort of hierarchy that determines when a given lower order rule is applicable.

fork on the right side of one's plate generalizes to all cultures in which authorities, rules, customs, etc., say that forks should be placed on the left.

We need to be careful not to think that what authorities, rules, customs, etc., determine to be appropriate for a specific context is never morally relevant. As I argued in my discussion of Kelly et al.'s results, people may think certain acts are generally okay but wrong when they are forbidden by political authorities. The key point, however, is that the acts are regarded as wrong not simply because authorities say they are wrong, but because of moral judgments that are authority independent. For example, the judgment that it is generally wrong to disobey reasonably just political authorities holds irrespective of what authorities say about the matter. To be clear, then, although some of the contextual considerations that determine when moral judgments apply are such that what authorities, rules, customs, etc., say may impact what is okay or not okay in a specific context, these contextual considerations are authority independent, i.e., they apply whether or not authorities, rules, customs, etc., say they apply.

If am correct, then moral judgments are, properly speaking, authority independent, but given the complicated relationship between moral judgments and authority, I believe the term "unconditionally obligatory" is preferable. Of course, this term can also be misleading if taken to mean that moral judgments such as "it is wrong to lie" apply in every possible circumstance. However, this confusion can be avoided if we keep in mind that moral judgments contain some contextual assumptions, if only implicitly. Furthermore, given that some authors have claimed that judgments may be authority independent without being generalizable, the term unconditionally obligatory has the virtue of suggesting a tight relationship between the two criterion judgments, a relationship which is supported by a broad range of data.

If these points are correct and moral judgments have a more complex relationship to authority and context than previously supposed, we have prima facie evidence that the distinction is abstract and sophisticated. With this point in mind, I will turn now to the question of whether the moral/conventional distinction could plausibly be acquired by all children in the normal course of development without any domain-specific cognitive structures.

6. The Moral/Conventional Distinction and PoS Arguments

Some empiricists accept that the moral/conventional distinction is conceptually real but argue that the moral "data" available to children is sufficient for them to develop this distinction without any domain specific faculty. I believe it is highly implausible to think

that children are capable of the kind of meta-cognizing required to form a distinction between the moral and conventional domains and that, consequently, the fact that children as young as 3 ½-years-old employ this distinction is the right kind of evidence for a PoS argument.

A. Empiricist Explanations

Jesse Prinz (2008) argues for an empiricist view of the moral domain and attempts to undermine a variety of nativist arguments. Although Prinz has expressed skepticism about whether the moral/conventional distinction is real, he has claimed that even if it is real, there is a plausible empiricist explanation for how children acquire the distinction.

According to Prinz, children come to acquire the distinction because, while parents do not explicitly teach it to their children, parents do treat violations of moral and conventional rules differently. Along with Kelly and Stitch, Prinz treats the distinction as a distinction between rules. He assumes along with researchers in the Turiel tradition that these two domains of rules are associated with different patterns of reasoning, patterns exhibited by children and adults. Prinz argues that if parents exhibit different reasoning patterns for violations of moral and conventional rules, children will be exposed to these different reasoning patterns and be able to "imitate" and "internalize" them. He cites a number of studies which suggest that parents adapt their disciplinary measures to the kinds of rules their children violate. Moral rule violations, he claims, "are likely to be enforced using power assertion and appeals to rights, and conventional rules are likely to be enforced by reasoning and appeals to social order" (392).

Prinz's argument seeks to establish two key claims. First, that the moral data available to children is not impoverished because parents consistently treat violations of moral and conventional rules differently. Second, that the moral/conventional distinction children employ is entirely the result of their using general purpose learning systems to internalize this data and imitate the reasoning patterns exhibited by their parents. I think both claims are problematic.

B. Problems for Empiricist Explanations

The first problem with Prinz's account is that it seems to assume that parents discipline their children in a highly systematic and consistent manner. He is, of course, right to point out that if the moral/conventional distinction is a real conceptual distinction, it will be exhibited in the reasoning patterns of adults. However, as Dwyer (2008, 416) explains, some parents treat violations of conventional rules very seriously. Furthermore, even if most parents think that, say, hitting is worse than failing to clean up some food spilled on the floor, parents are often tired and stressed and can become angry over anything

their children do. It seems highly unlikely that parents' emotional attitudes with respect to violations of moral and conventional rules differ significantly and consistently enough to provide children with a robust distinction that they can imitate and internalize. Prinz's claim looks even more dubious in light of the fact that in the normal course of development, *all* children employ the distinction consistently in paradigm cases at an early age. For his account to have any plausibility, it must be the case that *all* parents discipline their children consistently along the lines Prinz suggests.

Another serious challenge for empiricist accounts such as the one Prinz gives relates to the fact that we employ the moral/conventional distinction to make judgments about other individuals' use of the distinction. For instance, we might call someone who treats rules about how to properly set a table as moral rules an "over-moralizer." In calling someone an over-moralizer, we seem to be making some kind of normative judgment to the effect that they are moralizing something which they *ought not* moralize. The fact that individuals disagree about which judgments ought to be regarded as moral and which ought to be regarded as conventional helps drive home the point that there are not clear labels for which things belong in which category. For empiricists, this presents an acute challenge because they must explain how young children grasp the distinction sufficiently well to apply it consistently in paradigm cases when the categories "moral" and "conventional" are not clearly laid out in children's experience.

These difficulties for empiricist accounts of the development of the moral/conventional distinction suggest that it is the right kind of evidence for a PoS argument. If, at an early age, children across cultures employ a distinction that they could not plausibly have learned from the evidence available to them, we have good reason to believe that domain-specific cognitive structures are involved in the development and use of this distinction.

7. Conclusion

In order for the moral/conventional distinction to provide nativists with a compelling line of evidence for a PoS Argument, it must meet two conditions. First, unsurprisingly, it must at least be a real conceptual distinction. Second, it must be the kind of distinction that could not plausibly be acquired simply through experience and general purpose learning systems.

I have argued that the moral/conventional distinction meets both conditions. The large body of work generated by Turiel and his followers provides strong evidence that the distinction is recognized early by children across cultures. The work of Haidt, Nichols, and

Kelly et al. suggests that certain elements of the distinction need to be refined. We need to make clear that characterizing moral judgments as generalizable and unconditionally obligatory/authority independent does not mean they are insensitive to context. We also need to recognize that participants' judgments exhibit a considerable degree of uniformity in paradigm moral and conventional cases and that there are a number of borderline cases in which judgments do not exhibit such uniformity. The upshot, in my view, is that the moral/conventional distinction is both vague and sophisticated. It is vague because the consistency with which it is applied breaks down in borderline cases, and sophisticated because the distinction seems to be specified in terms of abstract and subtle principles regarding the conditions under which judgments generalize and the relationship between judgments and authorities.

I believe the second condition is also met. Here, the relevant facts are that the distinction appears early, in children as young as 3½ years old, and across cultures, where children are exposed to a wide range of moral upbringings. If I am right that the distinction is abstract and sophisticated, we have further evidence for the nativist position since, all other things being equal, the more complex the moral "output" the less likely it is that children are able to learn the distinction using only general purposing learning systems and the moral data plausibly available to them.

Further investigations will hopefully shed more light on the role this distinction plays in our judgments. I want to be clear that I have only offered a preliminary sketch of how the moral/conventional distinction might break down. If I am right in thinking that the moral/conventional distinction is highly abstract and sophisticated, future studies should be careful not to draw overly simplistic conclusions about what is going on in participants' minds/brains when they seem to regard certain judgments differently than others. It would be helpful to examine more closely participants' beliefs about the ways in which moral and conventional judgments generalize and their relationship to context and authorities. Future studies should also try to get a clearer idea of what the paradigmatic cases are and what it is about the "borderline cases" that causes the distinction to break down when applied to them.

References

Blair, James. 1995. "A Cognitive Developmental Approach to Morality: Investigating the Psychopath." *Cognition* 57: 1–29.

Chomsky, Noam. 2005. *Rules and Representations*. New York: Columbia University Press.

Dwyer, Susan. 1999. "Moral Competence." In *Philosophy and Linguistics*, edited by Kumiko Murasugi and Robert Stainton, 169–90. Boulder: Westview Press.

——— . 2008. "How Not to Argue that Morality Isn't Innate: Comments on Prinz." In *Moral Psychology: Vol. 1: The Evolution of Morality: Adaptations and Innateness*, edited by Walter Sinnott-Armstrong, 407–18. Cambridge: MIT Press.

——— . 2008. "Moral Psychology as Cognitive Science: Explananda and Acquisition." Unpublished manuscript.

Haidt, Jonathan, Silvia Koller, and Maria Dias. 1993. "Affect, Culture, and Morality, or is it Wrong to Eat Your Dog." *Journal of Personality and Social Psychology* 65 (4): 613–628.

Haidt, Jonathan, and Fredrik Bjorklund. 2008. "Social Intuitionists Answer Six Questions About Moral Psychology." In *Moral Psychology: Vol. 2: The Cognitive Science of Morality: Intuition and Diversity*, edited by Walter Sinnott-Armstrong, 181–217. Cambridge: MIT Press.

Harman, Gilbert. 2000. *Explaining Value and Other Essays in Moral Philosophy*. New York: Oxford University Press.

Helwig, Charles, Marie Tisak, and Elliot Turiel. 1990. "Children's Social Reasoning in Context." *Child Development* 61 (6): 2068–2078.

Kelly, Daniel, and Stephen Stitch. 2008. "Two Theories About the Cognitive Architecture Underlying Moral Judgment." In *The Innate Mind: Vol. 3: Foundations and the Future*, edited by Peter Caruthers, Stephen Laurence, and Stephen Stitch, 348–366. New York: Oxford University Press.

Kelly, Daniel, Stephen Stitch, Kevin J. Haley, Serena J. Eng, and Daniel M.T. Fessler. 2007a. "Harm, Affect, and the Moral/Conventional Distinction." *Mind and Language* 22 (2): 117–131.

——— . 2007b. "Scenarios and Results," *RCI Rutgers*." http://www.rci.rutgers.edu/~stich/Data/Data.htm.

Mikhail, John. 2000. "Rawls' Linguistic Analogy: A Study of the 'Generative Grammar' Model of Moral Theory Described by John Rawls in *A Theory of Justice*." Doctoral Dissertation. Cornell University.

Nichols, Shaun. 2002. "On the Genealogy of Norms: A Case for the Role of Emotion in Cultural Evolution." *Philosophy of Science* 69: 221–236.

Nucci, Larry. 1985. "Children's Conceptions of Morality, Societal Convention, and Religious Prescription." In *Moral Dilemmas: Philosophical and Psychological Issues in the Development of Moral Reasoning*, edited by C. Harding, 115–36. Cambridge: Cambridge University Press.

Nucci, Larry. 2001. *Education in the Moral Domain*. Cambridge: Cambridge University Press.

Nucci, Larry, and Elliot Turiel. 1993. "God's Word, Religious Rules and Their Relation to Christian and Jewish Children's Concepts of Morality." *Child Development* 64: 1485–1491.

Prinz, Jesse. 2008. "Is Morality Innate?" In *Moral Psychology: Vol. 1: The Evolution of Morality: Evolution and Innateness*, edited by Walter Sinnott-Armstrong, 367–406. Cambridge: MIT Press.

Rosas, Alejandro. 2012. "Mistakes to Avoid in Attacking the Moral/Conventional Distinction." *The Baltic International Yearbook of Cognition, Logic, and Communication* 7: 1–10.

Sripada, Chandra, and Stephen Stitch. 2006. "A Framework for the Psychology of Norms." In *The Innate Mind: Vol. 2: Culture and Cognition*, edited by Peter Caruthers, Stephen Laurence, and Stephen Stitch, 280–301. New York: Oxford University Press.

Turiel, Elliot. 2000. "The Development of Morality." In *Handbook of Child Psychology: Vol. 3: Social, Emotional, and Personality Development, 5th Edition*, edited by William Damon, Richard M. Lerner, and Nancy Eisenberg, 863–932. Hoboken: Wiley.

Naturalized Rationality, Evolutionary Psychology and Economic Theory

Yakir Levin
Ben-Gurion University of the Negev

Arnon Cahen
Ben-Gurion University of the Negev

Izhak Aharon
The Interdisciplinary Center, Herzliya
The Hebrew University of Jerusalem

Biographies

Itzhak Aharon (Gingi) is a lecturer at the Interdisciplinary Center in Herzliya, Israel, and a visiting fellow of the Center for Rationality and Interactive Decision Making in Jerusalem. After receiving his PhD in computational neuroscience from the Hebrew University in Jerusalem, he worked for several years in the high-tech industry, before returning to academia as a member of the Martinos Center for Biomedical Imaging in Boston. His research centers on the neurobiology of motivation and decision making (neuroeconomics).

Arnon Cahen is a postdoctoral research fellow at the philosophy department of Haifa University in Israel. He received his PhD in Philosophy-Neuroscience-Psychology from Washington University in St. Louis. Prior to coming to Haifa University, he was a Kreitman Postdoctoral Research Fellow at the philosophy department of Ben-Gurion University of the Negev, Israel. His research concerns the nature of mental content, perception, action theory, and rationality.

Yakir Levin is a senior lecturer in philosophy at Ben-Gurion University of the Negev, Israel. After receiving his PhD in philosophy from the Hebrew University, he was a Postdoctoral Chevening Scholar at Oxford University, Golda Meir Postdoctoral Fellow at the Hebrew University, and a visiting scholar at Harvard University. His research interests are early-modern philosophy, analytic metaphysics and philosophy of mind.

Acknowledgements

Earlier versions of the paper were presented at the philosophical colloquium at Ben-Gurion University, and at the "Re-Thinking Rationality" workshop co-organized by the Center for Rationality and Interactive Decision Making at the Hebrew University and the philosophy department at Ben-Gurion University. We would like to thank the participants in both these events for their helpful comments and suggestions. Thanks are also due to the referee of the *Journal of Cognition and Neuroethics* for a very helpful constructive suggestion.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). June, 2013. Volume 1, Issue 1.

Citation

Levin, Yakir, Arnon Cahen, and Izhak Aharon. 2013. "Naturalized Rationality, Evolutionary Psychology and Economic Theory." *Journal of Cognition and Neuroethics* 1 (1): 39–72.

Abstract

The rationality assumed by mainstream Rationalistic Economics (RE) and the irrationality purportedly revealed by Behavioral Economics (BE) are like a Hegelian "thesis" and "anti-thesis." Well aware of this conflict, proponents of RE have pursued several strategies to reconcile RE's rationalistic thesis with the ample evidence supporting BE's irrationalistic anti-thesis. Yet, none of these attempts appear satisfactory, which casts serious doubt on the possibility of reconciling RE with BE. Recently, Robert Aumann, 2005 Economics Nobel Laureate, has suggested a novel approach to this conflict, indeed, a new paradigm (Rule-Rationalism, or RR), which is supposed to synthesize RE and BE in terms of an innovative evolution-grounded notion of rationality. One aim of this paper is to cast further doubt on the possibility of reconciling RE with BE by showing that Aumann's suggested non-standard reconciliation fails. Another, related, aim is to show that RR does not genuinely present an alternative to RE and standard BE, but, rather, falls well within the bounds of the latter. Yet another aim is to address two fundamental issues in philosophy and psychology that RR involves—namely, the possibility of naturalizing practical rationality in evolutionary terms, and the scope of evolutionary explanations in psychology and economics. These aims are interwoven, as is reflected in the structure of the paper's central argument. We begin by arguing that practical rationality cannot be naturalized in evolutionary terms. Based on this we then argue that RR fails to synthesize RE and standard BE, yet that it may still be a new and better paradigm than either one of them. Next, we argue that evolutionary explanations in psychology and economics are of a limited scope. Based on this, we then conclude the paper by arguing that RR does not form a new paradigm in economics after all, and is, instead, tantamount to a call for evolutionary explanations in BE (when possible).

Keywords: Aumann, Rationalistic Economics, Behavioral Economics, Neuroeconomics, practical rationality, act rationality, rule rationality, naturalized rationality, intentional agency, Evolutionary Psychology, Human Behavioral Ecology

1. Introduction

The assumption of rationality—that people act in their own best interests, or seek to maximize utility—underlies most of economic theory.[1] Indeed, utility maximization is the central paradigm of mainstream economic theory, or Rationalistic Economics (RE) as we shall call it (Simon 1999; Tomer 2007, 466-467; Aumann 2008, 2; Etzioni 2011). Yet, ample evidence has accumulated suggesting that subjects deviate systematically from utility maximization. This evidence has led to the development of Behavioral Economics (BE), a core claim of which is that rationality is bounded, and that subjects are irrational to the extent that they do not always seek to maximize utility.[2] Thus, a core assumption of RE and a core claim of BE are like a Hegelian thesis and anti-thesis.[3]

In an attempt to reconcile their core rationalistic thesis with the evidence supporting BE's irationalistic anti-thesis, proponents of RE have pursued four main strategies (Etzioni 2011, 1101-1108). First, marginalize the recalcitrant findings—i.e., acknowledge their validity but insist that their implications are rather limited. Proponents of this strategy generally insist that the recalcitrant findings merely illuminate relatively minor foibles, or that the required departures from RE are not very radical, but merely relax simplifying assumptions of RE that are not central to this economic approach. Second, reframe the irrational as rational—i.e., reinterpret the recalcitrant findings to make them fit RE's core thesis: no matter how strange or irrational a particular economic decision might seem,

1. The rationality of *actions* or practical rationality of the sort referred to here differs from the rationality of *beliefs* or theoretical rationality that involves evaluation of the reasons for and against beliefs. Unless otherwise indicated, hence forward we shall mean by 'rationality' practical rationality. There are of course close ties as well as important analogies between theoretical and practical reason (Audi 1990 and 2004). But we shall almost completely ignore these ties-cum-analogies.

2. For a nuanced account of the development of BE, see Sent 2004. Actually, BE consists of quite a few strands as well as individual practitioners whose work does not fit neatly into any one of these strands. However, since there is enough commonality in these strands, they do form a whole (Tomer 2007).

3. RE also assumes unbounded willpower as well as unbounded selfishness, which BE also rejects (Mullainathan and Thaler 2000). Notice that the tension between RE and BE with respect to rationality, willpower, and selfishness does not entail that RE and BE completely exclude each other. As Camerer and Loewenstein (2004, 3) have aptly put it: "At the core of behavioral economics is the conviction that increasing the realism of the psychological underpinnings of economic analysis will improve the field of economics *on its own terms*—generating theoretical insights, making better predictions of field phenomena, and suggesting better policy. This conviction does not imply a wholesale rejection of the neoclassical approach to economics based on utility maximization, equilibrium, and efficiency. The neoclassical approach is useful because it provides economists with a theoretical framework that can be applied to almost any form of economic (and even noneconomic) behavior, and it makes refutable predictions."

construct a rational explanation for it. Third, cut some, rather limited slack in RE's core thesis to accommodate the recalcitrant findings—e.g., relax the conditions on rationality, thereby admitting recalcitrant behavior within the bounds of rationality.[4] Fourth, delegate the recalcitrant factors to exogenous realms—i.e., divide the social world between realms of behavior that adhere to the laws of RE and those that are governed by different rules, and then confine economics to the former realm. However, none of these strategies appear satisfactory (ibid.), which casts serious doubt on the possibility of reconciling RE with BE.

Building upon ideas that have long been "in the air," Robert Aumann, 2005 Economics Nobel Laureate, has recently suggested a novel approach to the conflict between RE and BE, indeed a new paradigm (Rule-Rationalism, or RR), which is supposed to synthesize RE's rationalistic "thesis" and BE's irrationalistic "anti-thesis" (Aumann 2008; see also Aumann 1997, § 2.3, and van Damme 1998). One aim of this paper is to cast further doubt on the possibility of reconciling RE and BE by showing that Aumann's suggested non-standard reconciliation in terms of RR fails. A second, related aim is to show that RR does not really form a *new* paradigm in economics that might supplant RE and standard BE. A third aim is to address two fundamental issues in philosophy and psychology that RR involves—namely, the possibility of naturalizing practical rationality in evolutionary terms, and the scope of evolutionary explanations in psychology and economics. Indeed, at the core of the paper is an extended argument that weaves all these themes together.

In the first part of the paper (§ 2), we outline RR and explain how it is supposed to synthesize RE and BE by way of a novel notion of rationality ("rule-rationality"), which Aumann suggests as an alternative to the standard notion of rationality used in economics ("act-rationality"). Unlike act-rationality, rule-rationality relies on the notion of evolutionary fitness (to be defined below). Our critical examination of RR centers on this crucial aspect of rule-rationality. By parity of reasoning, it applies to other accounts of rationality in evolutionary terms to which we shall return in due course.

In the second part of the paper (§ 3), we argue that practical rationality cannot be naturalized by way of evolutionary fitness (§§ 3.1-3.4). On this basis we then argue that due to the reliance of RR on the notion of evolutionary fitness, rather than synthesizing RE and BE, RR is to be viewed as a version of BE (§ 3.5).

As such, RR may still be better off than RE and other versions of BE, and if so should

4. The main difference between the second and third strategies may be characterized as a difference in the direction of fit: while the second strategy seeks to fit the recalcitrant findings to a preconceived notion of rationality, the third strategy seeks to fit the notion of rationality to the recalcitrant findings.

supplant them. In the third and final part of the paper (§ 4), we argue that the scope of evolutionary explanations in psychology and economics is rather limited (§§ 4.1-4.5). On this basis, we then argue that due to the central role that evolutionary fitness plays in grounding the notion of rule-rationality, RR cannot really supplant standard versions of BE but can at most supplement them (§ 4.5); indeed, it is tantamount to a call for evolutionary explanations in BE (when possible).

2. Rule-Rationalism

2.1 Act-Rationality vs. Rule-Rationality

At the core of RR lies a distinction between two notions of rationality: "act-rationality" and "rule-rationality." "[W]hen making [an act-rational] decision, economic agents choose an act that yields maximum utility among all acts available in that situation" (Aumann 2008, 2). In contrast, "under rule-rationality people do not maximize over acts. Rather, they adopt rules, or modes of behavior, that maximize some measure of total or average or expected utility, taken over all decision situations to which that rule applies; then, when making a decision, they choose an act that accords with the rule they have adopted. Often this is the act that maximizes their utility in that situation, but not necessarily always; the maximization is over rules rather than acts" (ibid., 2). In the following two sections we illustrate these rather abstract definitions drawing on Aumann's discussion of the Ultimatum Game (ibid., 6-7).[5]

2.2 Ultimatum Game Scenarios I—Behavioral vs. Rationalistic Economics

In the Ultimatum Game, two players are offered the opportunity to win a certain amount of money. One of the players (the proposer) suggests how to split the money, while the other player (the responder) may accept or reject the offer. If the responder accepts, both get the share agreed to, otherwise none of them receives any money. According to standard, act-rationality-based utility theory, responders should behave on the principle that any monetary amount is preferable to none. However, responders tend to reject unfair offers although this means missing out altogether. As a result, their behavior seems to manifest a clear violation of act-rationality.

Behavioral economists consider this apparent violation of act-rationality as evidence for the involvement of irrational elements—e.g., wounded pride, self-respect, a desire for revenge, etc.—in economic decision making. In their view, the *apparent* violations of act-rationality are *genuine* violations, which should be accounted for in terms other than

5. Originally introduced by Güth et al., (1982).

utility maximization. When responders in the Ultimatum Game behave as they do, they are driven not by reason alone but by various psychological factors that have nothing to do with utility maximization. And, they argue, these psychological factors should find their due place in accounts of economic decision making.[6]

Alternatively, the apparent violation of act-rationality at issue may be *explained away* by considering the psychological factors in the explanation of economic decision making as simply broadening what counts as utility for rational agents: the satisfaction of anger or of the desire for revenge in the Ultimatum Game is a kind of utility that enters into the agent's rational choice of an appropriate response. Pride, insult, self-respect, and revenge are legitimate sources of utility and disutility, so responders are behaving entirely according to norms of act-rationality when rejecting an unfair offer; they actually get positive utility from taking revenge, and would get negative utility from accepting an insulting offer.[7] On this view, then, act-rationality is not really violated in cases such as those illustrated by Ultimatum Game scenarios.[8]

2.3 Ultimatum Game Scenarios II—Rule-Rationalistic Economics

6. Among those arguing for the *utility-independent* role of emotions in economic choice behavior (in Ultimatum Game scenarios and others) are Elster (1998), and Rick and Lowenstein (2008). Some appeal in particular to the agent's sense of fairness, for example, Thaler (1988), Rabin (1993), and in some respect also Boudon (1998). The latter appeals to the agent's commitments or principles (among which might be fairness) that enter into decision making, yet have nothing to do with enhancing utility (indeed they might go strongly against utility). Still others appeal to framing effects in decision making, most famously, Tversky and Kahneman (1981 and 1986., and see also the various articles in Kahneman and Tversky 2000).

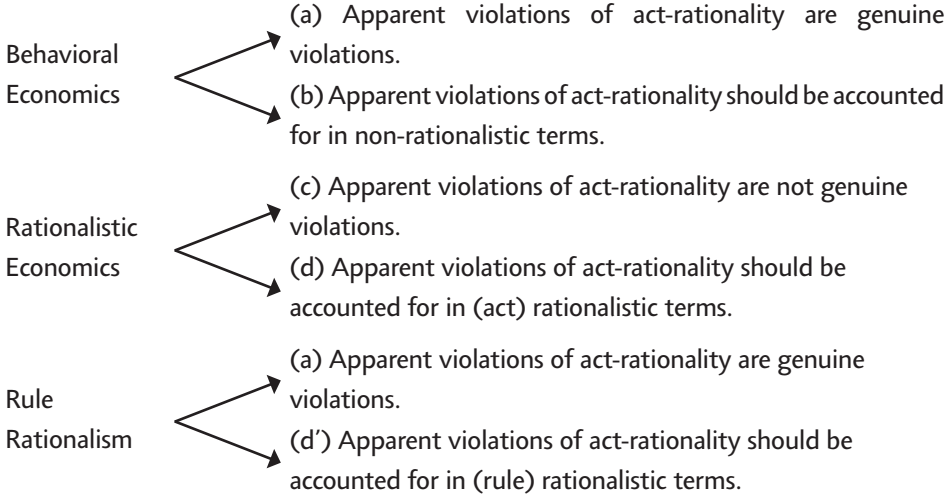
7. See, for example, Fehr and Schmidt (1999) who incorporate a variable of fairness in their utility function, and Bolton and Ockenfels (2000) who replace a utility function with a motivation function that the agent aims to maximize, and in which various psychological factors are represented as variables. Roth 1995 provides a good survey of various approaches to emotions that aim to explain their operation in terms consistent with the maximization of utility.

8. Aumann focuses in his discussion on RE's (1) "reframing the irrational as rational" approach to apparent violations of act-rationality, and ignores completely the three other RE's approaches to these apparent violations: (2) marginalizing them, (3) cutting some slack in RE's assumptions, (4) delegating the apparent violations to exogenous realms. What may explain Aumann's sole focus on (1) are the following assumptions that seem to underlie his discussion: (a) that violations of act-rationality in the Ultimatum Game say, if genuine, are not really marginal; (b) that one cannot really cut a slack in RE's assumptions by defining down what is entailed by being act-rational; (c) that apparent violations of act-rationality as in the Ultimatum Game should be treated within economic theory, and not be delegated to some other field. Since in this part of the paper we outline Aumann's approach, we will follow him in assuming (a)-(c), and therefore also in focusing solely on (1). Notice, however, that nothing in our critical discussion of Aumann's paradigm in the next two parts of the paper (Parts 3 and 4) will hang on these assumptions.

One major response, then, to the apparent violation of act-rationality in the Ultimatum Game is the BE approach of considering it a genuine violation. Another major response is the RE approach of explaining it away as a mere apparent violation. RR suggests a path between the two, synthesizing them by combining major features of both. Like BE it takes act-rationality to be genuinely violated in the Ultimatum Game. However, like RE it aims to account for the decisions of agents in rationalistic terms—viz., in terms of a certain kind of utility maximization. On this view, while it is not act-rational for the responder to reject an unfair offer, it is rule-rational to do so. As a rule, one should reject lop-sided offers, for the reputational reason of being treated more even-handedly in the future—an outcome of high overall utility. Given that the Ultimatum Game involves an anonymous, one-shot, interaction, it is clear that no reputational gain can be had by rejecting whatever offer is proposed. Nevertheless, people follow this rule even in this situation because doing so usually is act-rational, specifically, in almost all—or perhaps even all—natural, "real-world" negotiations, which are not anonymous. In contrast, following alternative rules that *accept* such lop-sided offers might result in the agent's being act-rational in Ultimatum Game scenarios but would reduce overall utility across negotiation situations, and thus would not be a favorable rule overall.

2.4 Diagrammatic Presentation

The differences between the approaches of BE and RE to apparent violations of act-rationality, as well as the way RR synthesizes these approaches, can be presented diagrammatically as follows:



RR synthesizes BE and RE by combining element (a) of the former economic paradigm

with a modified version (d') of element (d) of the latter paradigm which preserves the gist of (d)—viz., that apparent violations of act-rationality should be accounted for in *rationalistic* terms.

2.5 Behavioral, Rationalistic or Rule-Rationalistic Economics?

In Aumann's view, the RR approach to apparent violations of rationality by economic agents is to be preferred to both the BE approach and the RE approach. It is to be preferred to the former because it purportedly shares with the prevailing paradigm in economics the view of economic agents as rational. And it is to be preferred to the latter approach because it provides a deeper and more satisfactory account of decision making.

According to the RE approach, which denies that we have before us *genuine* violations of act-rationality, the emotions—viz., self-respect, wounded pride, a desire for revenge, and so on—are accorded a utility value. Yet the utility of these psychological factors is left unanalyzed. No account of their utility is provided in more fundamental terms, such as their evolutionary or other purpose or function. No answer is provided to questions such as: What purpose does it serve to feel insulted, or to take revenge? What is the function of self-respect? Without providing some answer to these questions, the RE attribution of utility to the operation of such psychological factors seems ad hoc and question begging, thus threatening to trivialize the very explanation that RE's appeal to utility maximization aims to provide.[9]

The RR approach offers a potentially richer and deeper explanation, both methodologically and conceptually, of the *rational* operation of these psychological factors in the decision making process. It does so by rather plausibly suggesting that such factors may be taken to have evolved genetically or memetically as mechanisms for the implementation of rules that usually, though not always, maximize utility. A combination of these factors, for example, may be considered the mechanism for implementing a rule such as "Reject lop-sided offers" in situations such as the Ultimatum Game. By following such a rule the responder (normally, but not necessarily always) establishes a reputation

9. Fehr and Schmidt (1999), for example, construct a utility function that includes a variable for fairness. Clearly the same can be done with respect to any psychological factor presumed to play a de facto role in decision making. However, without having some substantive answer as to why such a factor, a sense of fairness for example, ought to play any role whatsoever in determining the utility function it amounts to little more than fitting the utility function to the data. Zamir (2001) also points out a similar difficulty with such accounts. As he notes, such accounts merely provide a description of the phenomena in need of explanation. Such descriptions do not constitute an *explanation*. Thus, he argues that "[i]t may be useful to have a utility function, incorporating 'motivations' or 'fairness', that captures observed behavior fairly well, but it is more challenging to explain these attributes from the very basic assumptions of rationality" (ibid., 3).

within her community that promises progressively greater benefits in future negotiation situations, benefits that serve to enhance the responder's fitness, defined as "the expected number of offspring" (Aumann 2008, 3).[10]

Although, from the responder's perspective, she is merely acting on the basis of her emotions, she is in fact acting on the basis of a mechanism that has survived evolutionary pressures precisely because it instantiates a rule that maximizes fitness. Thus, in acting on their emotions, responders in the Ultimatum Game are de facto maximizing some measure of total or average or expected utility—viz., fitness. In so doing, they optimize a rule so as to do well "in general," but not necessarily always.[11]

2.6 Rule-Rationality and Evolutionary Fitness

Fitness, then, plays a key role in Aumann's RR approach to apparent violations of rationality by economic agents. It serves as the link between the operation of seemingly irrational elements in the subject's decision making process and a notion of rationality understood in the traditional RE terms of utility maximization.

According to Aumann, "rationality is a product of evolution. Rationality has evolved, alongside of physical features like eyes, stomachs, limbs, and breasts, because it maximizes fitness. A person shopping around for lower prices is maximizing fitness, because the money saved can be used to purchase food, theater tickets for a date, shelter, attractive clothing, education for the children, and so on – all of which increase fitness" (Aumann 2008, 4). Thus, by viewing rationality from the prism of evolutionary theory, we can explain the operation of *seemingly* irrational psychological factors in decision making as

10. Fitness, in this sense, provides the basis for a formal definition of natural selection which is the agent of adaptation in Darwinian evolution (Travis and Reznick 2009, 106). Aumann's use of this notion reflects his commitment to evolutionary psychology. In the literature on evolutionary psychology, however, the notion usually used is that of adaptation rather than fitness. So adaptation is the notion that will take center stage in the third part of the paper, where we discuss the role of evolutionary psychology in Aumann's paradigm as well as in economic theory in general.

11. Notice that an analogous account cannot be deployed within the framework of the RE approach to the emotions. Viewing the emotions from the prism of act-rationality, this approach takes the satisfaction of an economic agent's emotion in a given situation as a kind of utility that enters into the agent's rational choice of an appropriate response in that *specific* situation. So, it can account for the utility of a satisfied emotion in a given situation in terms of the contribution of this satisfaction to fitness only if the satisfaction at issue does contribute to fitness enhancement. However, the satisfaction of an emotion and fitness enhancement may part ways in specific situations: I may be getting positive utility from taking revenge; indeed, it may become the goal of my life. Yet this may cost me my life. The big advantage of rule-rationality over act-rationality, insofar as evolutionary explanations of choices in terms of fitness are concerned, is that the former but not the latter may be indifferent to disparities between the positive utility of satisfied emotions and fitness enhancement.

having survived evolutionary pressures by their instantiating rules that increase fitness. And, because the operation of these psychological factors increases fitness *by virtue* of increasing overall utility, we can see their operation as contributing to the rationality of the agent, contra BE.

At the same time, we also evade the difficulty of ad hockery raised above against RE's attempt to broaden what counts as utility to resolve the threat of irrationality. Without some substantive account that justifies such broadening, the strategy is no better off than defining utility by revealed preferences. Yet, defining utility in this way trivializes the notion, making any choice by an agent utility maximizing, and thus rational, by definition (Sen 1977);[12] it would be in danger of turning the notion of utility into a target that automatically moves to wherever the arrow is aimed (Bermúdez 2009, 2).

RR resolves this problem by providing a substantive account of the utility of the purportedly problematic intervening psychological factors in terms of fitness maximization. Of course, crucial for the appeal to fitness to play this explanatory mediating role is that we understand utility in terms of fitness enhancing goods. And indeed Aumann makes this "substantive relation" between fitness and utility maximization explicit. To avoid a trivialization of the notion of utility "one must define [it] more substantively—in terms of basic desiderata like time, money, family welfare, life, health, food, and so on—which are indeed closely related to fitness. With such a definition, an act that maximizes utility is then indeed act-rational, and a rule that usually maximizes it, rule-rational" (Aumann 2008, 17).[13]

As we shall argue in the next part of the paper, Aumann's suggested synthesis of BE and RE is hindered in large part because of this way of thinking about the "substantive relation" between rationality and fitness.

12. As Sen points out, however, despite this feature of the revealed preference approach it does not lead to a mute theory. For, it imposes a requirement of internal consistency of observed choices, and this might well be refuted by actual observations (Sen 1977, §§ II and III; see also Kaclnick 2006, § 2.3).

13. Compare Sen's claim that "once we give up the assumption that observing choices is the only source of data on welfare, a whole new world opens up, liberating us from the informational shackles of the traditional approach" (Sen 1977, 339-340). Indeed, in Sen's view, a person whose revealed preferences fulfill the requirement of internal consistency and is thus rational according to the standard economic wisdom, yet behaves unrelatedly to happiness or other substantive concerns must be a bit of a fool. "The purely economic man is […] close to being a social moron [… or a] rational fool," as he puts it (ibid., 336). Accordingly he proposes a view of rationality that does take into account the substantive interests of mankind. For an illuminating discussion of revealed preference (or operational) accounts of utility and substantive accounts, see Bermúdez 2009, Chap. 2.

3. Aumann's Paradigm and the Evolutionary Approach to Practical Rationality

3.1 The Basic Argument

Our basic argument for the failure of Aumann's proposed synthesis of BE and RE is based on a more general argument against the possibility of naturalizing practical rationality by way of evolutionary fitness. Schematically put, our basic argument runs as follows:

(a) Rationality is in the normative "logical space of reasons."

(b) Fitness maximization is in the factual "logical space of nature."[14]

(c) Thus, an account of rationality in terms of fitness maximization must be incorrect. Rationality cannot be naturalized by way of evolutionary fitness.

(d) But Aumann's synthesis takes utility maximization to be in the last analysis some sort of fitness maximization. It rests on an account of rationality in terms of evolutionary fitness maximization.

(e) Thus, Aumann's synthesis fails.

Step (d) follows from our discussion in § 2, and it does not require further elaboration (though we will return to an important presupposition of this premise in the last part of the paper). So, in the rest of § 3 we shall elaborate on the other steps of the argument.

3.2 Step (a): Rationality is in the "Logical Space of Reasons"

3.2.1 The Deontological Dimension of Rationality

Act-rationality is *perscriptive*: "it tells people how they should behave to advance their self-interest" (Aumann 2008, 13-14). Relatedly, act-rationality is also *justificatory*—i.e., its dictates may *justify* our behavior: in their lights our behavior may look reasonable, or proper, or right. Indeed—and this is the other side of the same coin—in case we violate these dictates we may be subject to blame or reproach for acting improperly, or unreasonably rationally speaking. To be act-rationally justified in having done something, one's action must not violate the dictates of act-rationality. The action is, then, permissible and one cannot be rightly blamed for doing it. Contrariwise, if one acts in violation of the dictates of act-rationality one is act-rationally unjustified and blameworthy. Thus, act-rationality has a *deontological dimension*: it has to do with obligation, permission, requirement, blame, and the like (cf. Alston 1988).

As we shall argue in the next section (§ 3.2.2), any notion of practical rationality worthy of the name must share these aspects of act-rationality, or be *normative* in all of

14. We borrow the notions "logical space of reasons" and "logical space of nature" from McDowell (1996) and Sellars (1997).

these respects. Any such notion must be, to paraphrase Wilfrid Sellars, in the logical space of reasons, of prescribing and justifying (Sellars 1997, 76). This needs argumentation because, as we shall see in § 3.2.3, *non-normative* notions of practical rationality might be proposed on the model of available non-normative notions of theoretical rationality. Indeed, Aumann himself takes rule-rationality to be a genuine notion of practical rationality, yet appears, at times, to consider it a non-normative, or "positive concept [… that] describes how people *do* behave [rather than how they *should* behave]" (Aumann 2008, 14).

3.2.2 Rationality and Intentional Agency

Some of our behavior, and much of the behavior of other animals is a function of invariant responses to detected stimuli, which can be characterized in terms of what cognitive ethologists call "innate releasing mechanisms"—namely, fixed patterns of behavior that are more complex than reflexes, often involving a chained sequence of movements rather than a simple reaction, and yet seem to be instinctive. Such mechanisms are triggered by specific stimuli; they always take the same form; they occur in all members of the relevant species; their occurrence is largely independent of the individual creature's history; once launched they cannot be varied; and they have only one function (Bermúdez 2003, 7-8). Behavior of this type, then, is automatic, mechanical, inflexible, context-bound, and not under the control of the animal. When behaving in this way an animal is no more than a stimulus-response system.[15]

It is also the case, however, that much of our behavior and some of the behavior of other animals consist of flexible and plastic responses to the environment, which are the result of complex interactions between internal states: organisms respond flexibly and plastically to their environment in virtue of the fact that their representational states respond flexibly and plastically to each other, usually through the influence of stored representations on present representations (Bermúdez 2003, 9). Typically, behavior of this type is goal-directed: it is best explained either in terms of the purposes and the desires that it is intended to satisfy or, more minimally, simply in terms of those that it does satisfy. Relatedly, it may involve intentional agency in which case ends and information

15. A classic example of this type of behavior is the egg rolling behavior of Greylag Geese. A goose of this species will roll a displaced egg near its nest back to the others with its beak. The sight of the displaced egg triggers this mechanism. If the egg is taken away, the Goose continues with the behavior, pulling its head back as if an imaginary egg is still being maneuvered by the underside of its beak. Moreover, it will also attempt to move other egg shaped objects, such as a golf ball, door knob, or even an egg too large to have possibly been laid by the goose itself.

about ends/means contingencies interact-rationally to produce intentions to act.[16]

Corresponding to varying degrees of flexibility and plasticity intentional agency forms a spectrum onto which different types or levels of rationality can be mapped. Indeed, a recurring theme of recent work on animal cognition has been the importance of escaping from a crude dichotomy between an inflexible, rigidly context-bound stimulus-response system on the one hand, and full-fledged conceptual, inferential, and mindreading abilities on the other. It is becoming continuously clearer that various finer distinctions between locations on this spectrum should be drawn (Hurley 2006). How exactly to do this—most especially, how to incorporate into this scheme animals that are devoid of language and are of limited conceptual capacities—is a matter of controversy (Bermúdez 2003 and 2006; Fodor 2003; Hurley 2006; Milikan 2006).[17] Fortunately, for our purposes we need not get into this vexed issue but rather elaborate a bit on another, though related, aspect of intentional agency.

To be an intentional agent or to act intentionally is to act for a reason (Audi 2006, Chaps. 4-5). This means, firstly, that for every intentional action there is a corresponding practical argument whose premises represent the structure of the causal and explanatory basis of the action, namely, the set of wants and beliefs—the motivational and cognitive elements—that explain why it is performed (ibid., 103).[18]

In addition, it means that the practical argument corresponding to the action is realized either inferentially, by an actual process of practical reasoning, or merely behaviorally (ibid., 114-115). The case of inferential realization, in which an agent acts on the basis of practical reasoning, is the paradigm of action for a reason, and does not really need an example. An example of a behavioral realization would be a spontaneous action that expresses the agent's motivating want and is guided by her belief where neither the want nor the belief is entertained or otherwise occurrent in the agent's consciousness.[19]

16. An example of a relatively plastic, though non-intentional behavior is the navigational behavior of wild animals such as the Tunisian desert ant and Clark's nutcracker (Hauser 2001, Chap. 4). An example of intentional behavior by wild animals is tool use (ibid., Chap. 2), and perhaps also social play and predatory and antipredatory behavior (Allen and Bekoff 1997, Chaps. 6-7).

17. A closely related controversy is of whether propositional attitudes such as belief can be ascribed to animals, and if so how related they are to propositional attitudes that may be ascribed to humans (Stitch 1979; Davidson 1982; Allen and Bekoff 1997, Chap. 5).

18. This most important thesis goes back to Aristotle, and seems also to be held by Hume and Kant (Audi 2006, 103).

19. A few philosophers—e.g., Davidson (1970), and Harman (1976)—maintain the thesis that for an action to

Finally, the equivalence between intentional action and action for a reason also means that whether the practical argument corresponding to the action is realized inferentially or behaviorally, the explanatory basis of the action is the same: in both cases the action is explained by the motivational and cognitive elements of its corresponding practical argument which are invoked for this purpose either directly (in case of inferential realization) or reconstructively (in case of behavioral realization) by the agent, or indeed someone else (ibid., 114).[20] In other words, the system of answers to the question what to do of practical reason is one and the same system of answers to the question "why?" of action explanation (Rödl 2007, 44-55; see also Anscombe 1963, §§ 37-40).[21]

From the normative and explanatory nexus of practical reason, then, springs intentional agency.[22] But the close ties between practical reason and intentional agency via this nexus are also the source of the centrality of the normative dimension of rationality: if practical rationality is characterized by the role it plays in making an action intentional—i.e., the space of reasons is coextensive with the space of intentional action (Hurley 2006, 167), or better still, acting intentionally *is* being of a rational mind (Rödl 2007, 49)—and if normativity is a central feature of practical rationality in this role, then normativity is a constitutive feature of practical rationality. Any notion of practical rationality worthy of

---

be intentional its corresponding practical argument must be inferentially realized (cf. Audi 2006, 233n1). But as Audi (ibid., Chap. 5) convincingly argues, this requirement is too strong, and for an action to be intentional it is sufficient that its corresponding practical argument be realized behaviorally. See also McDowell 2006, §§ 2, 3 and 8. McDowell's example of behavioral realization is of a person who, following a marked trail, at a crossing of paths goes to the right in response to a signpost pointing that way (ibid., 129).

20. Animals, who unlike us enjoy only limited, domain specific intentional agency—they occupy "islands of intentional agency" to paraphrase Hurley (2006)—and who lack a public language, cannot explain their behavior in these terms. This does not mean, however, that animals cannot have second order thoughts within their islands of intentional agency (unless it is assumed with e.g., Bermúdez (2003) that such thoughts require a public language, a rather implausible assumption as Fodor (2003) argues). Relatedly, it does not mean that animals are incapable of the self-determination—or the capacity to step back from their natural inclinations to act in a specific way and rationally assess this act—that we are capable of qua intentional agents (McDowell 2006, §§ 2, 3 and 8), though unlike our rather general self-determination, theirs would be limited to the islands of intentional agency that they occupy.

21. The view of intentional agency outlined here implies the guise-of-the-good thesis according to which doing something intentionally is thinking it good. This thesis has been attacked recently by philosophers with an empiricist bent—e.g., Velleman (1992). However, in light of Boyle's and Lavine's (2010) effective defense of the guise-of-the-good thesis, we consider this thesis as sound.

22. Intentional agency may require more than this nexus (Korsgaard 1997, 221; cf. Rödl 2007, 49-51). But even if the nexus at issue is only necessary for intentional agency, that would suffice for our purposes.

the name must be normative.[23]

3.2.3 Internalism vs. Externalism

In grounding the normativity of practical reason in what is constitutive of intentional agency[24] we have also provided support for an internalist approach to practical reason according to which the reasons that rationalize intentional actions must be cognitively accessible to the agents performing these actions: in our account, what an agent is doing is an action if and only if her reasons for doing it explain why she is doing it. This is no mere coincidence, since non-normative accounts of rationality go hand in hand with externalism, or the view that justifying reasons need not be cognitively accessible to agents. Thus, an argument for the normativity of rationality must be ipso facto also an argument for an internalist conception of reason, a point on which we would like to elaborate a bit.

A prominent example of an externalist view is the reliabilist account of theoretical reason (Bonjour and Sosa 2003, 24-26). According to reliabilism, a belief is justified if and only if it is produced by a reliable cognitive process, or a process that makes it objectively likely that the belief is true.[25] And since a reliable process in this sense need not be cognitively accessible to the holder of the belief, reliabilism does not require cognitive accessibility of the justifying grounds of a belief.

An example of an externalist account of practical reason may be the view that an agent is behaving rationally if and only if her behavior is produced by processes that are likely to be conducive to utility maximization and which need not be cognitively accessible to her. As we shall see in § 3.4, rule-rationality may be considered externalist in

23. The notion of rationality emerging from our discussion here is a reasoned assessment notion that concerns itself in the first place with the mental processes that lead to behavior. In contrast, both the act-rationalistic and the rule-rationalistic notions are consequentialist act evaluation notions that concern themselves first and foremost with the patterns of behavior resulting from inner mental processes. (For the difference between these types of notions of rationality, see Kacelnik 2006.) Implying that the most basic or primary notion of rationality is of the reasoned assessment type—that "the notions of rationality and reasoning are correlative notions" (Bermúdez 2006, 136)—our discussion in this section also implies that consequentialist notions of rationality must be derivative or secondary notions. This point has an important consequence to which we shall turn in footnote 30.

24. Notice that this grounding is rather different from recent attempts to extract moral or other norms from what is constitutive for action, or to resolve major metanormative problems on this basis. Thus, our grounding is immune to criticisms of these attempts. For a thorough critical examination of these attempts, see Enoch 2006.

25. There are actually quite a few versions of reliabilism, but the differences between them are unimportant for our purposes. For the appeal of reliabilism to philosophers with a naturalistic bent, see Fodor 1995.

this sense, and so may two more related conceptions of rationality.[26]

In grounding rationality in processes that may be beyond the cognitive ken of agents, and thus also beyond their voluntary control, externalist conceptions of (theoretical or practical) rationality are not prescriptive; they are not concerned with what rational subjects should or ought to believe or do: ought, as Kant reminds us, implies can (Kant 1929, A 548/B 576).[27] For the same reason they are also not deontological: they have nothing to do with obligation, permission, requirement, blame, and the like. From the externalist perspective, for a belief or action to be rational is not for the belief or action to be permissible or unblameworthy. And to be irrational is not to be blameworthy. Thus, the externalist conception of rationality is non-normative.

In contrast, internalist conceptions of rationality are normative. Thus, in order to show that rationality is of a normative nature, we had to argue that it must be an internalist conception, which is what we did by grounding it in intentional agency. For this reason, it also cannot be taken for granted that the norms of practical reasoning are constitutive of practical rationality; that to behave rationally is to follow the principles of practical reasoning and to be responsive to their dictates. To argue that the normativity of practical rationality stems from its conceptual ties with the norms of practical reasoning is to get things backwards. One must first show the normativity of practical rationality by showing that it must be an internalist conception, from which would follow, as we saw, its close ties with practical reasoning, rather than the other way around.

3.3 Step (b): Evolutionary Fitness is in the "Logical Space of Nature"

Unlike rationality, evolutionary fitness is neither prescribing nor justifying. To say that maximization of fitness is an evolutionary driving force in developing traits such as a tendency to behave (act or rule) rationally is not to say that evolution justifies these traits or that we should or ought to have them or behave as they cause us to behave. To say this

26. See also Audi's outline of an externalistic conception of action according to which "an action is intentional when it *in fact* is produced or sustained in an appropriate way by a suitable set of one's wants and beliefs, say non-waywardly produced or sustained by an overriding want for something and a (rational) belief that the action is necessary for achieving it. This condition may hold even if, in principle, one could not introspectively come to know or justifiably believe that one has those grounds for action" (Audi 1990, 233). As Audi notes, such behavioral externalism "cuts us off from our actions in a way that seems to make us more like spectators of our own doings than their agents. Agents can […] know (or at least form justified beliefs concerning) what they are about, in a sense implying a capacity to know for what reason(s) they are acting; they are not in the position of observers whose only route to such knowledge (or justified belief) is observational" (ibid.). Highly relevant here is Moran 2001, Chap. 4.

27. "The action to which the *'ought'* applies must indeed be possible under natural conditions."

is merely to say how things *are*, and not how they *ought* to be. It is to express something factual rather than something normative. In Immanuel Kant's memorable words,

That our reason has causality […] is evident from the *imperatives* which in all matters of conduct we impose as rules upon our active powers. *'Ought'* expresses a kind of necessity and of connection with grounds which is found nowhere else in the whole of nature. The understanding can know in nature only what is, what has been, or what will be. We cannot say that anything in nature *ought to be* other than what in all these time-relations it actually is. When we have the course of nature alone in view, *'ought'* has no meaning whatsoever. It is just as absurd to ask what ought to happen in the natural world as to ask what properties a circle ought to have. All that we are justified in asking is: what happens in nature? What are the properties of the circle? (Kant 1929, A 547/B 575)

Nature is governed by laws, but it does not follow these laws or behave in accordance with them. The laws of nature do not prescribe but rather describe, and it is not up to nature whether its behavior falls under them. In the spectrum in between mere automatism and intentional agency nature is clearly at the far end of the automatism side. So the whole language of deontology is inapplicable to it. It would be meaningless to say that nature's behavior was permissible, and that it cannot therefore be blamed for it. And it would be just as meaningless to consider nature's behavior as improper or wrong, and to take it to be blameworthy or irresponsible. All this is particularly true of evolution, which unlike rationally governed intentional actions, just blindly happens. Unlike intentional actions which are in the normative logical space of reasons, evolutionary fitness is in the non-normative logical space of nature.

3.4 Step (c): Rationality cannot be Naturalized by way of Evolutionary Fitness

Since practical rationality is a normative concept and fitness maximization lacks this dimension, practical rationality cannot be explicated or defined in terms of fitness maximization. A notion of utility that rests on fitness maximization is not a notion of rationality at all.

This is true of rule-rationality, as this notion rests in the last analysis on fitness maximization. But it is true, more generally, of any account of rationality in evolutionary terms that might be suggested. Such is the case, for example, with the notion of B-rationality ('B' is for biological) that has been recently suggested by the Oxford zoologist Alex Kacelnik (2006). B-rationality may be a useful notion in evolutionary biology. Yet, since it is defined in terms of fitness maximization it is not really a notion of rationality at all. And the same considerations apply to the human behavioral ecology model which considers humans not as utility maximizers but as fitness maximizers (Sternly and Jeffares

2010, 380).[28]

Another, closely related way of seeing this is via the internalism/externalism distinction. Practical rationality, we saw, is an intenalist concept: it takes the reasons that rationalize intentional actions to be in the cognitive ken of intentional agents. In contrast, evolutionary processes and their relation to behavior need not be in the cognitive ken of agents. So a notion of rationality that is based on the notion of evolutionary fitness—be it of the rule-rationality, or of the B-rationality, or of the human behavior ecology model type—must be an externalistic concept that does not require that justifying reasons be cognitively accessible to agents. Thus, the concept of rationality and the concept of fitness maximization are foreign to each other.

The notion of practical rationality cannot be founded then on the notion of evolutionary fitness. It is, of course, perfectly legitimate to define the word 'practical rationality' in terms of the meaning of the word 'fitness maximization.' One is certainly fully authorized to mean by a term whatever one likes. Nevertheless, just as defining the word 'white' in terms of the meaning of the word 'black' does not turn black into white, defining 'practical rationality' in terms of fitness maximization does not turn the notion of fitness maximization into a notion of rationality.[29,30]

---

28. Aumann, of course, identifies the two (see § 2.6).

29. In response, Aumann may possibly bite the bullet and argue that while fitness maximization may not be a genuine notion of rationality, what economists, himself included, are interested in is rationality in the sense of utility maximization. Thus, insofar as fitness maximization is a notion of utility maximization, that is all the rationality he needs for his paradigm. Yet, is the notion of fitness maximization *really* a notion of utility maximization? After all, the notion of utility is a normative notion while that of fitness is not a normative one. So, the response at issue seems to be begging the question against our argument: it actually assumes what our argument rejects—viz., that a fundamentally non-normative notion may be of the same type as a fundamentally normative notion. Our rejection of this assumption presupposes that normative notions cannot be analyzed or defined in non-normative terms. We are also committed to the related view that normative properties cannot be reduced to non-normative properties. So another, closely related response to our argument that is open to Aumann is to attack these two related presuppositions of ours. This is not the place to go into the vexed issue of our two presuppositions; suffice it to say that there are very strong arguments in their support (e.g., Smith 1994, Chap.2, and Bilgrami 2005, 7-15). In any case, we hereby put them on the table.

Moreover, the very fact that economists find puzzling apparent violations of the norms of (act) rationality—i.e, that apparently people do not behave as they *should* according to these norms, or in the best of their interest—attests to their commitment to a normative notion of rationality. And if so, this is another reason why Aumann cannot really respond to our argument by biting the bullet as suggested above.

30. A consequence of our discussion in § 3.2.2 is that the most basic or primary notion of rationality is of the reasoned assessment type. Yet, both act and rule-rationality are consequentialist act evaluation notions (see

3.5 Step (e): Losing Touch with Rationalistic Economics

RR seeks to synthesize BE and RE by combining stance (a) of the former (according to which apparent violations of act-rationality should be considered as genuine violations) with stance (d) of the latter (according to which the apparent violations at issue should be accounted for in rationalistic terms).[31] A central feature of this paradigm, however, is that it bases the rationalistic terms of stance (d) on fitness maximization. And by the previous section this turns the rationalistic terms of stance (d) to non-rationalistic terms, thereby transforming stance (d) to stance (b) of BE (according to which the apparent violations of act-rationality should be accounted for in non-rational terms). Thus, due to the central place it must give fitness maximization if it is to have its advantage over RE, RR actually loses touch with the latter theory. Short of having one leg—(a)—in BE, and another leg—(d')—in RE, RR has both legs—(a) and (d')-turned-(b—in BE. Rather than combining a rationalist thesis with a behavioral anti-thesis into a rationalist and behavioral synthesis, RR boils down to the behavioral anti-thesis.

4. Aumann's Paradigm and the Evolutionary Approach to Human Psychology and Behavior

4.1 Two Evolutionary Accounts

Despite its failure as a synthesis between RE and standard BE, the evolutionary approach to human psychology and behavior that RR presupposes (§§ 2.5-2.6), may still make this paradigm a serious and interesting alternative to both RE and standard BE. This immediately raises the question of the alleged explanatory value of the evolutionary approach at issue. Does this approach really give RR any explanatory edge over its rivals? Does RR really provide a deeper and better account than do RE and standard BE, and so should supplant them?

Aumann's evolutionary approach can be read in two ways. On the first reading, it

footnote 23). This means that for act or rule-rationality to be genuine notions of rationality they must have some substantial common denominator with the reasoned assessment notion of rationality. However, while act-rationality fulfills this requirement, rule-rationality does not fulfill it: act-rationality can be analyzed in terms of the *normative* notion of utility maximization, thereby sharing a central feature of the reasoned assessment notion of rationality. Yet, since rule-rationality is non-normative, it cannot be analyzed in these terms (see footnote 29). Thus, because of the non-normativity of rule-rationality, this notion does not have any other substantial common denominator with the most basic notion of rationality—i.e., the non-normativity of rule-rationality strikes rather deep.

31. See § 2.4.

takes human behavior to be generated by psychological mechanisms that are evolutionary *adaptations*: these mechanisms evolved because they produced behavior in our ancestors that enabled them to survive and reproduce. Thus read, Aumann's approach is akin to the Evolutionary Psychology research paradigm within the evolutionary approach to psychology (Buller 2000; Dowens 2008; Staratt and Shackelford 2010): like Evolutionary Psychology, Aumann's approach focuses under this reading on the question of whether a trait is an *adaptation*, rather than on whether it is currently *adaptive*.[32]

On the second reading of Aumann's evolutionary approach, it is mainly interested in the current adaptivity of human behavior. Thus read, Aumann's approach is akin to the Human Behavioral Ecology research program within the evolutionary approach to psychology (Borgerhoff Mulder 1991; Downes 2010): like Human Behavioral Ecology, Aumann's approach focuses under this reading not so much on the question of whether the proximal psychological mechanism that triggers a specific behavior is an adaptation as on the question of whether the behavior triggered is currently adaptive.[33,34]

---

32. For the adaptation/adaptive distinction, see Gould and Vrba 1983, 4-6, and Sober 2000, 85. As illustrated by the following examples, a trait can be adaptive without being an adaptation, and an adaptation without being adaptive: a sea turtle's forelegs are adaptive insofar as they are useful for digging in the sand to bury eggs, but they are not adaptations for nest building (Sober 2000, 85). And vestigal organs such as our appendix or vestigial eyes in cave dwelling organisms are adaptations but not currently adaptive (Dowens 2008, § 4).

33. This reading of Aumann's evolutionary approach fits better than the first reading the original synthesizing goal of RR: fitness enhancing adaptations in ancestral environments need not be fitness enhancing in modern environments. So, for an evolutionary account of rule-rationality in terms of fitness enhancement to preserve the rationality of agents (which is a main aspect of Aumann's synthesizing move), it must focus on current adaptivitiy (as the second reading does) rather than on evolutionary origin in adaptation (as the first reading does). At this stage of our discussion, however, the emphasis has shifted from the synthesizing goal of RR to its explanatory value. And here both readings will do: both may provide an evolutionary, albeit different, explanation of why people behave as they do.

34. Another major approach to human evolution is Richerson and Boyd's (2005) gene-culture co-evolution approach, according to which gene evolution molds cultural evolution, while culture affect the relative fitness of different genotypes in many ways. While this approach differs in important respects from both Evolutionary Psychology and Human Behavioral Ecology, it shares with the latter the explanatory element relevant for RR—viz., the emphasis on current fitness. Thus, the aforementioned differences would be rather unimportant for our purposes.

Still another evolutionary approach to human behavior and culture is the "memes" approach, according to which cultural evolution is the outcome of competition between cultural units, or "memes," that replicate and are selected in a way analogous to but separate from genes. For reasons we cannot get into here, this approach is highly problematic (Jablonka and Lamb 2006, 206-212). Indeed, even Dawkins, who contributed perhaps more than anyone else to its popularity has backed off his meme-talk (Dawkins 1982, 111-112). So we can

In this part of the paper we shall argue that the scope of evolutionary explanations in psychology is rather limited. This we shall do by arguing that culture involves the extension of human capacities in radically novel directions, which are neither adaptations (contra Evolutionary Psychology) nor adaptive (contra Human Behavioral Ecology)—§§ 4.2-4.4. We shall then show that this is true in particular with respect to economics, and on this basis conclude that RR cannot supplant standard BE but at most supplement it (§ 4.5).

4.2 Man the "Symbolic Animal"

While very similar in many respects to other inhabitants of the animal kingdom, most especially to the great Apes, humans are unique among those inhabitants in their highly complex and sophisticated ways of using symbols in both thinking and communication. Indeed, it has been rather plausibly argued that this use of symbols is the distinctive mark of humankind; that man is the "symbolic animal" (Cassirer 1944, Chap. 2; Jablonka and Lamb 2006, Chap. 6).

Closely related to this characteristic of humankind is its unique civilization and culture. While other animals may also have cultural traditions of a sort (de Waal 2001, Chaps. 5-8; Jablonka and Lamb 2006, Chap. 5), human cultural traditions are unique in their extreme and overwhelming richness, variety, and sophistication. Social animals may have socially transferable systems of patterns of behavior, preferences and products of activity.[35] But only human beings compose music and do mathematics, send missiles into space and build cathedrals, write books of poetry and philosophize, alter at will the genetic nature of their own and other species, interpret themselves and others, and exhibit an unprecedented level of creativity and destruction, rewriting the past and molding the future. In these respects, *Homo sapiens* is totally unlike any other species. And it is undoubtedly its symbolic system that opened the way to humankind's unique civilization. Due to this system "man lives not only in a broader reality [than other animals], he lives so to speak, in a new *dimension* of reality" (Cassirer 1944, 43).

Culture has brought then a dramatic qualitative change into human life, "transforming the very core of our being" (de Waal 2001, 29). It opened new ways of adapting oneself to one's environment (Richerson and Boyd 2005). Yet, at the same time, it has enabled the development of human traits and behavior that even if adaptive, pace Evolutionary Psychology, are not adaptations. Indeed, as we shall now argue, a great many cultural

safely ignore it in our discussion.

35. Cf. de Waal's (2001, 30-32), and Jablonka's and Lamb's (2006, 160) closely related definitions of culture.

inventions and the traits and behaviors that they involve must be of this sort.

4.3 Spandrels and Exaptations

Consider first the now familiar phenomenon of *spandrels*—traits that are by-products of selective processes (Lewontin and Gould 1977; Gould 1997) which "just come along for the ride," as Fodor (2007) put it. A famous example of a spandrel is the "male-mimicking" genitalia of the female spotted hyena that arose as a by-product of the evolution of female dominance and superior size, an adaptation built by high testosterone titers, which induce masculinized genitalia as an automatic result (Gould 1997). In all likelihood at least some of the mechanisms underlying human behavior are spandrels. Indeed, it has recently been suggested that certain specific aspects of the faculty of language may be by-products of preexisting constraints rather than end products of a history of natural selection (Hauser et al. 2002, 1574).

If there is reason to think that some of the mechanisms underlying human behavior are spandrels, there is a better reason still to think that a great many of these mechanisms are *exaptations*—i.e., features that were evolved for certain usages (or for no function at all) and then "coopted" or re-appropriated for other roles (Gould and Vrba 1983).[36] A classic example of exaptation is how feathers which initially evolved for insulation were coopted for flight (ibid.).[37] Another example is how the African Black Heron uses its wings to prey on small fish (ibid.). Yet another example is the hypothesis that mirror neurons, which are widely held to underlie social cognition, develop from visual and motor neurons through associative learning processes (Heyes 2009 and 2010). Still another is the Cheney and Seyfarth (2008, 143) suggestion that skills such as the ability to learn from others, invent new behaviors, and use tools piggybacked and built upon mental computations that had their origins in social interactions. A related example is their hypothesis that many of the rules and computations found in human language first appeared as an elaboration of the rules and computations underlying social interactions (ibid., 269-270; cf. Hauser et al. 2002, 1578). One more example is Darwin's thesis, supported by recent findings that the configuration of emotional facial expressions has evolved from a functional role in

36. As the great master of evolution himself said, "throughout nature almost every part of each living being has probably served, in a lightly modified condition, for diverse purposes, and has acted in the living machinery of many ancient and distinct specific forms" (Darwin 1886).

37. The conversion of the adaptation of feathers for insulation into an exaptation for flight set the basis for subsequent adaptations and exaptations, one of which forms our next example (Gould and Vrba 1983, 7-8). Yet, the possible interplay between adaptations and exaptations that this exemplifies is of no consequence for our discussion.

regulating sensory intake. These ancestral configurations may later have proven useful as social signals, assuming a new function without needing to change their basic form (Susskind et. al., 2008).

As the last four examples illustrate, exaptations may already be found rather far back in the evolutionary history of mankind. Thus, mirror neurons are found in monkeys (Heyes 2010). Intentionally modified stone tools already appeared ca 2.6 Myr BP (Stout et al. 2008, 1939). Speech capacity emerged by ca 200,000 BP with H. Sapiens, or maybe earlier and gradually over a much longer period (Renfrew 2008, 2042). And, as already said, the configuration of emotional facial expressions is rather ancient. However, exaptations must have become particularly widespread in our rather recent history, since the end of the Pleistocene period, say, ca 10,000 years BP. The reason for this is as follows.

At the end of the Pleistocene culture and society started developing rapidly and massively in ways unprecedented before (Renfrew 2008, 2042-2043). However, the biological basis of our species, the human genome, has been established for much longer, at least since the out-of-Africa dispersals of some 60,000 years ago (ibid., 2042).[38] Thus, many cultural domains and inventions must have coopted or re-appropriated traits and capacities that were naturally selected for other purposes; they must exemplify exaptations rather than adaptations (Sperber and Hirschfeld 2004; Dehaene and Cohen 2007; Renfrew 2008).

A concrete example of a cultural innovation that illustrates this general point is the practice of writing and reading. This practice was invented ca 5400 years ago by the Babylonians. Moreover, until very recently, only a very small fraction of humanity was able to read and write. Thus, it is impossible that human brain regions evolved specifically for the purpose of reading. And indeed, it has recently been found out that word reading utilizes brain areas that were initially unrelated to reading but rather to object and scene recognition, a function significantly different from the mapping of written language onto

38. This does not imply that humans have not continued to be subject to natural selection since the out-of-Africa dispersals or even during the last 10,000 years known as the Holocene period (e.g., the domestication of animals at the beginning of the Holocene introduced a new selection pressure that produced human lactose tolerance - Richerson and Boyd 2005, 191-192). Rather, it means that the bulk of our species genetic make-up has remained relatively constant during the last 60,000 years or so—i.e., that most genetic variation amongst humans occurred before they migrated out of Africa (Downes 2010, 250-251). This is suggested e.g., by the fact that variations among populations of different continents account for only about 10% of all genetic differences (Staratt and Shackelford 2010, 232). But notice that it does not preclude the possibility of some genetic variation occurring during the Holocene, for which there is also some evidence (Hawks et al. 2007). In any case, as we shall illustrate by two major examples of cultural inventions, one in this section and the other in § 4.5, many such inventions are too recent for significant genetic adaptation to their existence to be possible.

sound and meaning (Dehaene and Cohen 2007). A brain mechanism that evolved for a certain purpose has been re-appropriated by a cultural invention, and began to fulfill a completely different purpose;[39] even if its new use is adaptive, it is not an adaptation for this use.

Many cultural domains and inventions must have involved, then, the extension of human traits and capacities in radically novel ways that were not anticipated by evolution. These extensions must have exemplified exaptations rather than adaptations. As we shall argue next, it is very likely that, pace Human Behavioral Ecology, many such extensions are non-adaptive.

4.4 Non-Adaptive Extensions of Traits and Capacities

To see what we mean by non-adaptive extensions of traits and capacities and how they may emerge, consider the simple cultural tradition of the famous Japanese macaques of Koshima Island, the discovery of which set off the cultural revolution of primatology (de Waal 2001, 194-204; Jablonka and Lamb 2006, 178-179). Wanting to study these macaques in the early 1950s, Japanese primatologists used sweet potatoes to lure them from the forest to the sandy seashore, where they were easier to observe. This worked well, but had unexpected consequences. At one point, a juvenile female macaque known as Imo started washing the potatoes in a nearby stream, thereby removing the soil from them. The new habit spread to other monkeys, at first to juveniles and then to adults, gaining in complexity: the monkeys switched from the stream to the nearby sea, and started to bite the potatoes before dipping them into the salty water, thus seasoning them as well as washing them.

In addition to potatoes, the macaques were fed by wheat, which was difficult to collect and eat because inevitably it became mixed with sand. And here again Imo's genius proved highly instrumental. Her solution to this problem was to throw the mixed sand and wheat into the water, where the heavier sand sank while the wheat floated, making it easy for her to collect it. This new habit also spread to other monkeys and after some while most of the macaque community collected wheat in this way.

These sea-related habits of the macaques gave rise in turn to further habits. Infants that were carried by their mother when she washed the food became used to the sea, and started playing and bathing in it. Swimming, jumping and diving became popular. In addition, adult males began eating fish that the fishermen had discarded, a habit which

39. Indeed, as the case of reading shows, re-appropriations of this sort may profoundly alter cortical organization (Dehaene et al. 2010).

also spread in the community.

Since the scientists first started feeding the macaques in Koshima, a new life-style, or cultural tradition of a sort, has developed. One modification in behavior produced the conditions for the generation and propagation of other modifications, and a whole set of socially transferable patterns of behavior had evolved. Indeed, and this is the important point for our purposes, at least one of these evolving behavioral patterns—viz., the water playing habits—was not adaptive: even if the capacity to feel pleasure evolved in the first place for its adaptive utility, in the context of the macaque culture it no longer serves only adaptive purposes. Water playing has no adaptive value for the macaques. So, the pleasure it involves cannot be considered a mechanism in the service of enhancing adaptation. Rather, it should be considered a goal unto itself. What explains the macaques' water playing is not some other (unconscious) interest, but the very interest in water playing. They water play for the sake of water playing and not for some other hidden purpose.

What is true of aspects of the very simple cultural tradition of the macaques of Koshima is many times truer of human culture. Even if the exaptations that underlie a great many of our cultural practices (§ 4.3) emerged from adaptations rather than spandrels, many of them are like the macaques' practice of water playing in being pursued for their own sake, and not in the (unconscious) interest of evolutionary fitness. After all, what's the adaptive value of, e.g., listening to music, doing mathematics, writing philosophical articles, reading novels, reflecting on ourselves and on human nature, decorating our homes, performing a religious ritual, etc.? It may be possible, of course, to ascribe some strained adaptive value to at least some of these practices. Thus, according to Pinker (1997, 543), "fictional narratives supply us with a mental catalogue of the fatal conundrums we might face someday and the outcomes of strategies we could deploy in them. What are the options if I were to suspect that my uncle killed my father, took his position, and married my mother?" Short of vindicating, however, the view of these practices as adaptive, such ascriptions rather strengthen our claim that these practices are pursued for their own sake, or for the sake of several non-adaptive motives. "What if," elaborates Fodor (1998), albeit sarcastically, on Pinker's suggestion, "it turns out that, having just used the ring that I got by kidnapping a dwarf to pay off the giants who built me my new castle, I should discover that it is the very ring that I need in order to continue to be immortal and rule the world? It's important to think out the options betimes, because a thing like that could happen to anyone and you can never have too much insurance."[40]

40. Thirty years after the publication of Gould and Lewontin's (1979) seminal criticism of adaptationism, says

Moreover, not only are many of our cultural practices non-adaptive, but also, as we know all too well, many of our cultural practices may be counter-adaptive. "Human cultural practices can be orthogenetic and drive towards extinction in ways that Darwinian processes, based on genetic selection, cannot" (Gould and Lewontin 1977, 584). Man is a child of nature, but also of culture: adaptationist processes play an extraordinarily important role in shaping us, but we are also shaped by cultural processes that may well be non- or even counter-adaptive. As we shall now see, this conclusion reflects on RR.

4.5 Back to Economics

What is true of human psychology and behavior in general must be true of economic psychology and behavior. Thus, based on extensive theoretical and empirical literature about money Lea and Webley (2006) argue that major aspects of the extraordinary and reinforcing power of money are best explained by what they call Drug Theory rather than by the more standard Tool Theory. Drug Theory's basic explanatory concept is that of a functionless motivator that obtains its motivational effect by a parasitic action on a functional, evolutionarily adaptive system. In contrast, Tool Theory's basic explanatory concept is that of a motivator that, although of no biological significance in itself, is used instrumentally to obtain biologically relevant incentives. Tool Theory has been the standard account of the motivational power of money. But as Lea and Webley argue, (i) there are a number of significant phenomena that cannot be accounted for by a pure Tool Theory of money motivation; (ii) supplementing Tool Theory with a Drug Theory enables these phenomena to be explained; and, (iii) the human instincts that, according to a Drug Theory, money parasitizes include trading (derived from reciprocal altruism) and object play. According to Lea and Webley, then, money use has non-adaptive aspects. Moreover, money use cannot be the result of some evolutionary process that occurred within the hominid period: money has emerged only within the last 3,000 years or so (ibid., 162), too short a time for significant genetic adaptation to its existence; besides, individuals born into cultures that have never used money quickly come to use it if they move into a money-using culture. Thus, aspects of money use provide a nice illustration from the economic sphere of our general claim that many cultural inventions are neither adaptations nor adaptive. Indeed, the conceptual framework developed by Lea and

---

Rasmus Nielsen in a recent commentary in *Evolution*, "evolutionary biologists are […] arguably, much more reluctant to invent adaptive stories without direct evidence for natural selection acting on the traits in question. We still regularly encounter very naïve adaptive stories, particularly about human behavior, but rarely in journals […] with high standards of peer review, and rarely from researchers with a background in evolutionary biology" (Nielsen 2009, 2487).

Webley to deal with these aspects of money use may be employed in a more general account of non-adaptive cultural exaptations.

Clearly, RR cannot apply to those aspects of our economic psychology and behavior that are neither adaptations nor adaptive. So the fact that there are such aspects implies that it cannot supplant non-evolutionary accounts of economic behavior such as those provided by standard BE but *at most* supplement them—i.e., it can *at most* deepen these accounts when it comes to aspects of economic behavior that *are* adaptations or adaptive. Thus, consider once again the Ultimatum Game scenarios with which we started the paper. Recent research concerning the psychological mechanisms underlying the rejection of unfair offers in such scenarios strongly suggests that they are driven by a feeling of moral disgust, which is a cooption of the rejection impulse characteristic of distaste, an impulse that can already be found in sea anemones, which evolved about 500 million years ago (Chapman et al., 2009; Rozin et al., 2009). Insofar as moral disgust plays an important role in the regulation of social behavior of humans it has a significant adaptive value. And reference to this fact does indeed deepen accounts of economic behavior that is driven by this feeling and is of explanatory value. To that extent RR improves on standard BE. However, stripped as it should be of its alleged alternative account of rationality (§3), it is nothing but a call for employing evolutionary explanations in economics (when possible).

5. Conclusion

In conclusion, then, the appeal of RR to evolutionary explanations is a double-edged sword. On the one hand, such explanations may sometimes, though certainly not always, deepen accounts of economic psychology and behavior, and in that respect be of explanatory value. On the other hand, due to this very appeal to evolutionary psychology RR can neither fulfill its stated goal of synthesizing BE and RE, nor supplant any of these approaches, but at most supplement the former. Indeed, short of a new paradigm in economics, Aumann's suggestion is tantamount to a call for evolutionary explanations in BE.

As we argued in our four-fold move: (A) Practical rationality cannot be naturalized in evolutionary terms (§§ 3.1-3.4), and for this reason (B) RR fails to synthesize RE and standard BE, yet may still be a new and better paradigm than either one of them (§3.5). However, (C) evolutionary explanations in psychology and economics are of a limited scope (§§4.1-4.5), and for this reason (D) RR is not a new paradigm in economics after all and remains well within the bounds of BE (§4.5).

References

Allen, C. and Bekoff, M. 1997. *Species of Mind: The Philosophy and Biology of Cognitive Ethology*, Cambridge, MA: The MIT Press.

Alston, W.P. 1988. "The Deontological Conception of Epistemic Justification." Reprinted in W.P. Alston, 1989, *Epistemic Justification: Essays in the Theory of Knowledge*, 115–152. Ithaca: Cornell University Press.

Anscombe, G.E.M. 1963. *Intention*, 2nd edition. Oxford: Basil Blackwell.

Audi, R. 1990. "An Internalist Conception of Rational Action." *Philosophical Perspectives* 4: 227–245.

Audi, R. 2004. "Reasons, Practical Reason, and Practical Reasoning." *Ratio* XVII: 119–149.

Audi, R. 2006. *Practical Reasoning and Ethical Decision*. London: Routledge.

Aumann, R. J. 1997. "Rationality and Bounded Rationality." *Games and Economic Behavior* 21: 2–14.

Aumann, R.J. 2008. "Rule-Rationality versus Act-Rationality." HUJ Center for the Study of Rationality, Discussion Paper #497.

Bilgrami, A. 2005. "Self-Knowledge, Intentionality, and Normativity." *Iyuun*, 54: 5–24.

Bermúdez, J.L. 2003. *Thinking without Words*. Oxford: Oxford University Press.

Bermúdez, J.L. 2006. "Animal Reasoning and Proto-Logic." In S. Hurley and M. Nudds (eds.), *Rational Animals?*, 127–137. Oxford: Oxford University Press.

Bermúdez, J.L. 2009. *Decision Theory and Rationality*. Oxford: Oxford University Press.

Bolton, G. E. and Ockenfels A. 2000. "A Theory of Equity, Reciprocity, and Competition." *The American Economic Review* 90 (1): 166–193.

Bonjour, L. and Sosa, E. 2003. *Epistemic Justification: Internalism vs. Externalism, Foundations vs. Virtues*. Oxford: Blackwell Publishing.

Borgerhoff Mulder, M. 1991. "Human Behavioral Ecology." In J. Krebs and N. Davies (eds.), *Behavioral Ecology: An Evolutionary Approach*, 69–98. Oxford: Blackwell.

Boudon, R 1998. "Limitations of Rational Choice Theory." *American Journal of Sociology*, 104 (3): 817-828.

Boyle, M. and Lavin, D. 2010. "Goodness and Desire." In S. Tenenbaum (ed.) 2011. *Desire, Practical Reason and the Good*, 202–233. New York: Oxford University Press.

Buller, D.J. 2000. "Evolutionary Psychology." In M. Nani and M. Marraffa (eds.), *A Field Guide to the Philosophy of Mind*.

Camerer, C. F. and Lowenstein, G. 2004. "Behavioral Economics: Past, Present, Future." In C. F Camerer, G. Lowenstein, and M. Rabin (eds.), *Advances in behavioral economics*, 3–51. Princeton, NJ: Princeton University Press.

Cassirer, E. 1944. *An Essay on Man: An Introduction to the Philosophy of Human Culture*. New Haven: Yale University Press.

Chapman, H.A., Kim, D.A., Susskind, J.M. and Anderson, A.K. 2009. "In Bad Taste: Evidence for the Oral Origins of Moral Disgust." *Science* 323: 1222–1226.

Davidson, D. 1970. "How is Weakness of the Will Possible?" Reprinted in D. Davidson. 1980, *Essays on Actions and Events*, 21–42. Oxford: Clarendon Press.

Davidson, D. 1982. "Rational Animals." Reprinted in D. Davidson, *Subjective, Intersubjective, Objective*, 95–105. Oxford: Clarendon Press.

Dawkins, R. 1982. *The Extended Phenotype*. Oxford: W.H. Freeman and Company.

Dehaene, S. and Cohen, L. 2007. "Cultural Recycling of Cortical Maps." *Neuron* 56: 384–398.

Dehaene, S., Pegado, P., Braga, L.W., Ventura, P., Nunes Filho, G., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J. and Cohen, L. 2010. "How Learning to Read Changes the Cortical Networks for Vision and Language." *Science* 330: 1359–1364.

Darwin, C. 1886. *On the Various Contrivances by which British and Foreign Orchids are Fertilized by Insects and on the Good Effects of Intercrossing*, London: John Murray.

Downes, S.M. 2008. "Evolutionary Psychology." *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (ed).

Downes, S.M. 2010. "The Basic Components of the Human Mind Were Not Solidified During the Pleistocene Epoch." In F.J. Ayala and R. Arp (eds.), *Contemporary Debates in Philosophy of Biology*, 243–252. Oxford: Wiely-Blackwell.

Elster, J. 1998. "Emotions and Economic Theory." *Journal of Economic Literature* 36 (1): 47–74.

Enoch, D. 2006. "Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action." *The Philosophical Review* 115: 169–198.

Etzioni, A. "Behavioral Economics: Toward a New Paradigm." *American Behavioral Scientist* 55 (8): 1099–1119.

Fehr, E. and Schmidt, K. M. 1999. "A Theory of Fairness, Competition, and Cooperation", *The Quarterly Journal of Economics* 114 (3): 817–868.

Fodor, J. 1995. "Encounters with Trees." *London Review of Books* 17: 10–11.

Fodor, J. 1998. "The Trouble with Psychological Darwinism." *London Review of Books* 20: 11–13.

Fodor, J. 2003. "More Peanuts." *London Review of Books* 25: 16–17.

Fodor, J. 2007. "Why Pigs Don't Have Wings." *London Review of Books* 29: 19–22.

Gould, S.J. 1997. "The Exaptive Excellence of Spandrels as a Term and Prototype." *Proceedings of the National Academy of Sciences* 94: 10750–10755.

Gould, S. J. and Lewontin, R. 1979. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme." *Proceedings of the Royal Society London B* 205: 581–598.

Gould, S.J. and Vrba, E.S. 1982. "Exaptation—A Missing Term in the Science of Form." *Paleobiology* 8: 4–15.

Güth W., Schmittberger, R. and Schwarze, B. 1982. ''An Experimental Analysis of Ultimatum Bargaining.'' *Journal of Economic Behavior and Organization* 3: 367–388.

Harman, G. 1976. "Practical Reasoning", *Review of Metaphysics*, 29, pp. 431-463.

Hauser, M.D. 2001. *Wild Minds: What Animals Really Think*, London: Penguin Books.

Hauser, M.D., Chomsky, N. and Fitch, W.T. 2002. "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?" *Science's Compass* 298: 1569–1579.

Hawks, J., Wang, E.T., Cochran, G.M., Harpending, H.C. and Mozis, R.K. 2007. "Recent Acceleration of Human Adaptive Evolution." Proceedings of the National Academy of Sciences 104: 20757–20762.

Heyes, C. 2009. "Where Do Mirror Neurons Come From?" *Neuroscience and Biobehavioral Reviews* 34: 575–583.

Heyes, C. 2010. "Mesemerizing Mirror Neurons." *NeuroImage* 51: 789–791.

Hurley, S. 2006. "Making Sense of Animals." In S. Hurley and M. Nudds (eds.), *Rational Animals?*, 139–171. Oxford: Oxford University Press.

Jablonka, E. and Lamb, M.J. 2006. *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge, MA: MIT Press.

Kacelnik, A. 2006. "Meanings of Rationality." In S. Hurley and M. Nudds (eds.). *Rational Animals?*, 87–106. Oxford: Oxford University Press.

Kahneman, D. and Tversky A. (eds.), 2000. *Choices, Values, and Frames*. Cambridge, UK: Cambridge University Press.

Kant, I. 1929. *Critique of Pure Reason*, translated by N. Kemp Smith. London: Macmillan.

Lea, S.E.G. and Webley, P. 2006. "Money as Tool, Money as Drug: The Biological Psychology of a Strong Incentive." *Behavioral and Brain Sciences* 29: 161–176.

Korsgaard, C. 1997. "The Normativity of Instrumental Reason." In G. Cuillity and G.N. Berys (eds.), *Ethics and Practical Reason*, 215–254. Oxford: Oxford University Press.

McDowell, J. 1996. *Mind and World*. Cambridge MA: Harvard University Press.

McDowell, J. 2006. "Conceptual Capacities in Perception." Reprinted in J. McDowell, 2009. *Having the World in View: Essays on Kant, Hegel, and Sellars*, 127–144. Cambridge MA: Harvard University Press.

Milikan, R.G. 2006. "Styles of Rationality." In S. Hurley and M. Nudds (eds.), *Rational Animals?*, 117–126. Oxford: Oxford University Press.

Moran, R. 2001. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton: Princeton University Press.

Mullainathan, S. and Thaler, R.H. 2000. "Behavioral Economics." NBER Working Paper 7948.

Nielsen, R. 2009. "Adaptationism – 30 Years after Gould and Lewontin." *Evolution* 63: 2487–2490.

Pinker, S. 1997. *How the Mind Works*. New York: W.W. Norton and Company.

Rabin, M. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83: 1281–1302.

Renfrew, C. 2008. "Neuroscience, Evolution and the Sapient Paradox: The Factuality of Value and of the Sacred." *Philosophical Transactions of the Royal Society B* 363: 2041–2047.

Richerson, P.J. and Boyd, R. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press.

Rick, S. and Loewenstein, G. 2008. "The Role of Emotion in Economic Behavior." In M. Lewis, J. M. Haviland-Jones and L. F. Barrett (eds.), *Handbook of Emotions 3rd Ed.*, 138–156. NY, NY: The Guilford Press.

Rödl, S. 2007. *Self-Consciousness*. Cambridge MA: Harvard University Press.

Roth, A. E. 1995. "Bargaining Experiments." In J. H. Kagel and A. E. Roth (eds.), *Handbook of Experimental Economics*, 253–348. Princeton, NJ: Princeton University Press.

Rozin, P., Haidt, J. and Fincher, K. 2009. "From Oral to Moral." *Science* 323: 1179–1180.

Sellars, W. 1997. *Empiricism and the Philosophy of Mind*. Cambridge MA: Harvard University Press.

Sen, A. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy and Public Affairs* 6: 317–344.

Sent, E.-M. 2004. "Behavioral Economics: How Psychology Made Its (Limited) Way Back into Economics." *History of Political Economics* 36 (4): 735–760.

Simon, H. 1999. "The Potlatch between Political Science and Economics." In J. Alt, M. Levi and E. Ostrom (eds.), *Competition and Cooperation: Conversations with Nobelists about Economics and Political Science*, 112–119. Cambridge: Cambridge University Press.

Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.

Sober, E. 2000. *Philosophy of Biology*. Boulder, CO: Westview Press.

Sperber, D. and Hirschfeld, L.A. 2004. "The Cognitive Foundations of Cultural Stability and Diversity." *Trends in Cognitive Science* 8: 40–46.

Starratt, V.G. and Shackelford, T.K. 2010. "The Basic Components of the Human Mind Were Solidified During the Pleistocene Epoch." In F.J. Ayala and R. Arp (eds.), *Contemporary Debates in Philosophy of Biology*, 231–242. Oxford: Wiely-Blackwell.

Sterenly K. and Jeffares, B. 2010. "Rational Agency in Evolutionary Perspective." In T. O'Connor and C. Sandis (eds.), 374–383. *A Companion to the Philosophy of Action*. Oxford: Wiley-Blackwell.

Stitch, S. 1979. "Do Animals Have Beliefs?" *Australasian Journal of Philosophy* 57: 15–28.

Stout, D., Toth, N., Schick, K. and Chaminade, T. 2008. "Neural Correlates of Early Stone Age Toolmaking: Technology, Language and Cognition in Human Evolution." *Philosophical Transactions of the Royal Society B* 363: 1939–1949.

Susskind, J.M., Lee, D.H., Cusi, A., Feiman, R., Grabski, W. and Anderson, A.K. 2008. "Expressing Fear Enhances Sensory Acquisition." *Nature Neuroscience* 11: 843–850.

Thaler, R. H. 1988. "The Ultimatum Game." *The Journal of Economic Perspectives* 2 (4): 195–206.

Tomer, F.J. 2007. "What is Behavioral Economics?" *The Journal of Socio-Economics* 36: 463–479.

Travis, J. and Reznick, D.N. 2009. "Adaptation." In M. Ruse and J. Travis, 2009. *Evolution: The First Four Billion Years*, 105–131. Cambridge MA: Belknap Press.

Tversky, A. and Kahneman, D. 1981. "The Framing of Decisions and Psychology of Choice." *Science* 211: 453–458.

Tversky, A. and Kahneman, D. 1986. "Rational Choice and the Framing of Decisions." *Journal of Business* 59: s251 –s278.

van Damme, E. 1998. "On the State of the Art in Game Theory: An Interview with Robert Aumann." *Games and Economic Behavior* 24: 181–210.

Velleman, D.J. 1992. "The Guise of the Good." Reprinted in D.J. Velleman, 2000. *The Possibility of Practical Reasoning*. Oxford: Clarendon Press.

de Waal, F. 2001. *The Ape and the Sushi Master: Cultural Reflections of a Primatologist*. New York: Basic Books.

Zamir S. 2001. "Rationality and Emotions in Ultimatum Bargaining." *Annals d'Economie et de Statistique* 61: 1–31.

cognethic.org