Feedback from Moral Philosophy to Cognitive Science

Regina A. Rini University of Oxford

regina.rini@philosophy.ox.ac.uk

Forthcoming in *Philosophical Psychology*. Published online 4 December 2013.
This is an uncorrected manuscript. For the final, edited version, please visit
http://www.tandfonline.com/doi/full/10.1080/09515089.2013.864454

Abstract A popular argument form uses general theories of cognitive architecture to motivate conclusions about the nature of moral cognition. This paper highlights the possibility for modus tollens reversal of this argument form. If theories of cognitive architecture generate predictions for moral cognition, then tests of moral thinking provide feedback to cognitive science. In certain circumstances, philosophers' introspective attention to their own moral deliberations can provide unique data for these tests. Recognizing the possibility for this sort of feedback helps to illuminate a deep continuity between the disciplines.

0. Introduction

Quite a bit of recent work traces the implications of cognitive science for morality.[1] Cognitive science can tell us quite a bit about the mind: how it represents information, how it transforms this information in response to new experiences, how it brings the expression of this information to bear on judgment and decision-making. Moral cognition is one of the domains in which the mind operates, so it seems clear that domain-general theories about the organization of the mind should inform the study of morality. But does this influence go only one way? Can the particular study of moral thought feed back to domain-general debates in cognitive science? And can moral philosophers make a distinctive contribution to this feedback?

In this paper I start from the following popular argument form:

1. Theory of mental architecture T implies claims {C} about moral cognition.
2. Theory T is correct.

[1] See, for example, Gazzaniga (2006), Zimbardo (2007), Appiah (2008), Prinz (2009), Churchland (2011), and numerous essays in May, Friedman, and Clark (1996) and Sinnott-Armstrong (2008a).

3. Therefore {C} are true.

Here a ‘theory of mental architecture’, refers to a general theory about how the human mind is organized, including how information is represented and transformed. Such a theory involves empirical hypotheses and conceptual arguments. ‘Moral cognition’ refers to the particular domain of human thought concerned with moral evaluation and decision-making.

My goal is not to evaluate this argument form or any particular instance. Rather, I want to consider a logical consequence of accepting premise (1). Since the argument form employs modus ponens, we can generate the following modus tollens argument, starting from transposition of (1):

4. If claims {C} about moral cognition are *not* true, then theory of mental architecture T is *not* correct.
5. Claims {C} are not true.
6. Therefore, theory T is not correct.

In other words, starting again from (1): suppose that our favorite theory of mental architecture implies some set of claims about moral cognition. Given that the theory is correct, then the claims must also be true. But (now from (4)) if we come to find, independently, that the claims are *not* true, then this suggests that the theory may not be correct. The danger of a modus ponens argument is that it can be reversed.

My aim in this paper is to explore such reversals. I’ll consider instances of the familiar modus ponens, and suggest that their proponents should keep in mind that they’ve left themselves open to the corresponding tollens. So far this is an unsurprising (though little-noted) possibility. What is more interesting is *how* such a reversal might be triggered. Alongside the ordinary findings of cognitive science, we ought also pay attention to the practice of moral philosophers. After all, if our scientific theories have implications for how moral cognition works, we should expect to find moral cognition working in moral philosophers. Hence careful study of what moral philosophers do might provide feedback to psychological theory.

The broad aim of this paper is to show that the path from cognitive science to moral philosophy runs both ways. My arguments are suggestive rather than decisive – this is deliberate. I am trying to open a discussion about how to understand continuity between the disciplines, not to settle the matter. If I am right, the lesson is one of openness: cognitive scientists have much to teach moral philosophers – and perhaps much to learn as well.

1. From Mental Architecture to Moral Cognition

1.1 Descriptive (not prescriptive) implications

First, an important distinction. There are two different ways we might try to draw moral-domain implications from psychological theories. First, there are purely *descriptive* implications, which from a general psychological theory derive some particular claim about how people *actually do* engage in moral cognition. Second, in contrast, there are *prescriptive* implications, which make claims about how people *should* engage in moral cognition. In this paper, my interest is squarely on the first – descriptive – sort. Although some of the authors I'll discuss have indeed made prescriptive claims[2] , this paper is not interested in how psychology

[2] For instance, Paul Churchland (1996b, 105) and William Casebeer and Patricia Churchland (2003) have argued that the truth of association-based moral cognition (discussed below) supports Aristotelian normative ethics.

may impact prescriptive moral theory. Instead, the focus is on the purely descriptive implications.

I will discuss descriptive implications drawn from highly general psychological theories. The idea is clearest through examples. I will consider claims that differing accounts of basic mental structure – such as the dispute between rule-based and exemplar-based accounts – generate divergent implications about moral cognition. Once this example has helped to clarify the idea, in the rest of the paper I will explain how these inferences might be reversed, and show how moral philosophy can affect cognitive science.[3]

We can start from fundamental questions about mental architecture. At its most basic level, how is the mind organized? Are mental faculties innate or acquired? Is mental processing computational or associational? Do we reason by following abstract rules or by making similarity comparisons to learned exemplars? Answering any one of these questions means settling significant debates at the heart of cognitive science.

1.2 An example: computationalism and connectionism

Since at least the 1980s, a central debate in cognitive science has revolved around the relationship between computationalism and connectionism. Computationalism is, roughly, the view that the mind functions similarly to a computer, processing symbolic representations through syntactic manipulation (Putnam 1961, Fodor 1975). Connectionism claims that the mind is essentially composed of associational patterns instantiated in neural networks (Rumelhart, McClelland and the PDP Research Group 1986, Churchland 1996a). The two approaches do

[3] Great thanks to Michael Strevens for helpful suggestions on how to organize this section of the paper.

overlap in many ways, but there has been a lengthy and vigorous debate about the status of differences between them (Smolenksy 1988, Fodor and Pylyshyn 1988).

This debate remains unresolved and in many ways confusing (Piccinni 2009, Horst 2009). Among the disputed matters are the meanings of key terms, and even whether or not computationalism and connectionism are rivals in any important sense. For present purposes, I will set aside most of these disputes, along with conceptual niceties and distinctions that are not necessary to the present point. What I will focus on here, in order to provide an example of the argument form introduced in the last section, is the question of whether the mind processes information in a *rule-based* or an *association-based* manner. Many (though not all) computationalists think that the mind processes information in a rule-based way, where rules are machine-executable state-transitions relying on syntactic manipulation. Many (though not all) connectionists think that the mind processes information in an association-based way, where associations are comparisons of novel stimuli to learned category prototypes or exemplars. To simplify (with some loss of accuracy), the question boils down to how the mind applies concepts to new items: by subjecting them to algorithmic rules or by matching them to standard examples.[4]

1.3 Implications for moral cognition

[4] Henceforth I'll treat computationalism and connectionism as rivals which together exhaust the available theoretical options, though many people do not accept this. Since the point here is only to offer an example of an argument – and not to advance a conclusion - this stipulation should be acceptable. Among the options here ignored is the view that moral judgment arises from a 'kludge' – a set of functionally and phylogenetically disparate mental systems whose output we mistakenly map onto a single domain. (See Stich 2006.) If this is true, then it is a mistake to think we are investigating a single cognitive structure.

The claim so far is that computationalism and connectionism imply, respectively, rule-based and association-based accounts of cognitive processing. As we'll now see, these implications can be traced through to empirical research programs in moral cognition.

If general mental architecture is rule-based (as many computationalists would have it) then we should expect moral cognition to centrally feature the use of moral rules, or principles. And this is indeed what many researchers in the area believe. A prominent example is the theory known as 'moral grammar' (Dwyer 2006, Hauser et al. 2008, Mikhail 2011). Modeled on Chomsky's (1965) rule-based generative grammar approach to linguistics, the moral grammar program aims to use psychological evidence to recover the 'deep principles' behind moral judgments. John Mikhail, one of the most prominent proponents of this approach, locates it firmly in the rule-based computationalist tradition. Citing Chomsky and Marr (1982), Mikhail writes, "Cognitive scientists who take these ideas seriously and who seek to understand human moral cognition must devote more attention to developing computational theories of moral competence (Mikhail 2007, 151)."

According to the moral grammar account, moral intuition (like language) executes a range of biologically fixed rules, with some pruning and amplification by particular cultural circumstances. These implicit rules are recoverable through experiments linking controlled alterations of the characteristics of stories to subjects' reactions. Extending this approach, Cushman and Young (2011) argue that patterns of moral judgment can be explained through functional decomposition, wherein moral judgments are (partly) the output of cognitive operations performed on non-moral representations of the causal and intentional features of agents' actions. A central methodological assumption of this sort of research is that moral cognition is best understood in light of rule-based cognitive architecture.

By contrast, an association-based account (as in most connectionist views) will strongly de-emphasize the role of rules, focusing instead on similarity relations. Here is how Paul Churchland puts the point: "What is the alternative to a rule-based account of our moral capacity? The alternative is a hierarchy of learned prototypes, for both moral perception and moral behavior, prototypes embodied in the well-tuned configuration of a neural network's synaptic weights (Churchland 1996b, 101)." According to this view, if we want to understand how moral cognition works, we should not be looking to construct implicit principles from research subjects' reactions. Rather, we should be looking to establish the features of the prototypes to which they compare new stimuli, or the weight assigned to various features in the process of similarity-matching.[5]

Researchers sympathetic to this approach tend to stress the roles of social learning and the biological details of neuroanatomy. Patricia Churchland writes, "Early moral learning is organized around prototypes of behavior, and relies on the reward system to make us feel emotional pain in the face of some events (e.g. stealing), and emotional joy in the face of others

[5] As in other domains, connectionists disagree among themselves on the strength of their claims. Paul Churchland is quite insistent, as when he writes, "There is no hope… that we can capture the true substance of any human's moral knowledge by citing some family of 'rules' that he or she is supposed to 'follow', nor any hope of evaluating that person's character by evaluating the specific rules within any such internalized family. At the level of individual human cognition, it simply doesn't work that way (Churchland 2000, 298)." By contrast, Andy Clark, while broadly endorsing a connectionist approach, cautions against excluding a meaningful role for rules in social learning – see Clark (2000).

(e.g., rescuing). Through example, the child comes to recognize the prototypes of fairness, rudeness, bullying, sharing, and helping (Churchland 2011, 131)."

By "reward system", Churchland is referring specifically to the role of oxytocin and other chemicals in promoting the development of certain brain structures. Contrast this biological approach to remarks by computationalist Mikhail: "an adequate scientific theory of moral cognition will often depend more on the computational problems that have to be solved than on the neurophysiological mechanisms in which those solutions are implemented (Mikhail 2007, 143)."[6]

So we can see two contrasting examples of the modus ponens argument form I introduced at the start of this paper. One argument starts from a computationalist cognitive architecture and concludes that moral judgment should be understood in a rule-based way. The other argument starts from connectionist assumptions and concludes with an association-based approach to moral cognition.

In effect, these arguments generate *predictions* about the behavior of people engaged in moral deliberation; they predict that moral deliberation will be either rule-based or association-based. Now it should be clear what modus tollens reversal of these arguments would look like. If we can empirically test moral cognition and show that it does *not* conform to

[6] There is also significant disagreement between computationalists and connectionists on the extent to which moral judgments are fixed by innate cognitive structures. The moral grammar approach explicitly assumes that at least some basic principles are innate (though subject to cultural shaping), while connectionists emphasize the early equipotentiality of neural circuits and the consequent importance of cultural learning.

the prediction of a given theory of cognitive architecture, then we have some evidence against that theory.

So the scientific study of moral cognition can have implications for general theories of cognitive architecture. This is not a terribly surprising result (though it has gone largely unnoticed). After all, any general theory of cognitive architecture can be tested by examining its predictions for a particular domain.

We should, however, note certain problems with empirically testing general theories' implications for moral cognition. One is that, because of the complexity of these theories and the logic of their application, proponents can find an interpretation of almost *any* individual research finding that supports (or at least does not damage) their favored theory.[7] More importantly, the general theories give differing advice about the appropriate *methodology* for investigating moral cognition. As noted above, computationalists tend to focus on functional organization, while connectionists often make claims essentially concerning neuroanatomy. The functional and physical are different "levels" of psychological organization (see Marr 1982) and imply correspondingly different methodologies. It may then be difficult to devise experiments that both sides will agree are useful tests of their differences. [8]

[7] See, for instance, Patricia Churchland's treatment of the evidence for moral grammar (Churchland 2011, 103-111).

[8] Though this difficulty should not be overstated. In other domains, such as visual object recognition, clever experimental design has facilitated testing of deeply divergent theoretical claims (e.g. Biederman and Bar 1999). As the study of moral cognition continues to mature, it might benefit experimenters to look for methodological parallels to these domains. (Thanks to a referee for this journal for suggesting the parallel.)

Of course, it is certainly worth pursuing these questions through empirical techniques. But in the rest of this paper I will look in a somewhat different direction. I will propose that *philosophers* may have a unique role to play in these debates - though not their usual role.

2. The Philosophical Mind as Research Target

According to the arguments described above, general theories of cognitive architecture have implications for moral cognition. Among the species of moral cognition is moral cognition *by moral philosophers*. Moral philosophers are humans and have human psychologies; whatever is said about human moral psychology must have implications for the psychology of those engaged in moral philosophy.[9]

This points to an intriguing possibility. There may be value to carefully examining *how* moral philosophers think – and it may be especially valuable to have philosophers do this themselves. In this section I will defend a proposed research methodology which, conjoined to the modus tollens argument form discussed in the next section, allows feedback from moral philosophy to cognitive science. The proposal is to treat philosophers' introspective assessment of their own moral deliberations as data points against which the predictions of psychological

[9] It might be claimed that the moral cognition of professional philosophers differs from that of ordinary people in certain truth-conducive ways. This is a (partly) empirical assertion, and the subject of much current debate (Ludwig 2007, Weinberg et al. 2010, Williamson 2011, Schwitzgebel and Cushman 2012). I won't take any position on such claims here; they are orthogonal to the points about basic cognitive structure discussed in my examples.

theories can be tested. If this is right, then philosophers have an important and active role to play in empirical investigation.

First, some words about introspection in psychology. At first glance, it might seem that introspection is precisely what empirical investigation aims to move away from. But there is an old tradition in psychological research, dating to its founding days, of highlighting the introspective observations of highly trained research subjects. E.B. Titchener, student of Wilhelm Wundt and introducer of the psychological laboratory to America, employed such techniques in early studies of perception.

According to Titchener, certain mental phenomena are best investigated using subjects who are trained introspectors – that is, psychologists themselves. Titchener worried that naïve subjects would commit what he called the ‘stimulus error’, failing to distinguish between properties of a sensation and properties of an object *perceived in* that sensation. In particular, Titchener suggested that our perceptual experiences, seemingly simple and immediate, were actually highly processed, and only a very experienced introspector could effectively extract genuinely simple and intrinsic properties of perceptual states from this confusing complex.[10] Titchener’s solution was to have numerous highly trained introspectors (graduate students in psychology) carefully contemplate their subjective experience of the same perceptual situation, such as light admitted through a suddenly-opened curtain, and to look for convergence in their descriptions of this experience. Only such highly trained experts would have the powers of

[10] Titchener rejects “an account which purports to take mental phenomena at their face value, which records them as they are ‘given’ in everyday experience; the account furnished by a naïve, common-sense, non-scientific observer. … It is more than doubtful whether, in strictness, such an account can be obtained (1912, 489).”

discrimination and fine-grained vocabulary necessary to generate reports suitable for this intersubjective comparison.[11]

Of course, there are good reasons to be skeptical of introspective reports, even by highly trained introspectors. Modern research psychology has almost entirely moved away from introspection, in part because the collection of mass data from naïve subjects allows researchers to control for individual differences and to avoid the effects of hypothesis-awareness. All of this is quite reasonable, but it is consistent with allowing that there remains *some* role for introspective report in psychological research.

We can see one example in visual object recognition. Two well-development types of theory – structured description theory and view-based theory – dominated accounts of visual object recognition for a number of years (e.g. Hummel and Biederman 1992, Tarr 1995). Many forms of experimental evidence played a role in this debate, but one particularly influential argument relied in part on an introspectively-assessable feature of visual experience. Hummel and Stankiewicz (1996) devised a series of crossed-line drawings and invited participants to decide (introspectively) which ones looked more similar to one another. These drawings were designed so that one pair would clearly be regarded as most similar by view-based algorithms - yet participants gave similarity judgments incompatible with this prediction. Hence, the

[11] My reading of Titchener is heavily indebted to Schwitzgebel (2011, chapter 5). It should be noted that Schwitzgebel is skeptical of Titchener's method; he thinks that the 'training' of Titchener's experts actually resulted in a *change* in their subjective experience, thus rendering suspicious their apparent convergence. I'm also grateful to Daniel Robinson for the suggestion that I look at Titchener in this connection.

argument runs, if view-based theory predicts one pattern of similarity matching, yet introspective comparison disagrees, this is a big problem for view-based theories.[12]

So introspective evidence can matter to psychological theory. Now, why should we be particularly interested in the introspection of moral philosophers? Because philosophers –unlike ordinary research subjects – may be especially well-equipped to attend to the details of their own thinking.[13] Like Titchener's psychologist subjects, moral philosophers have a good sense of what is *important* in their domain: they have a better understanding than ordinary people of which are the controversial claims in moral theory and of how these are linked to particular mental states of theirs (i.e. intuitions). Further, moral phenomenology is complex and not easily suited to ordinary language (Chappell 2013). Moral philosophers possess a developed

[12] To be clear, Hummel and Stankiewicz (1996) did use standard naïve subject research techniques; the individuals doing the introspection were ordinary research subjects, not Titchener-style trained introspectors. However, it is informative that when Hummel (2000, 169-171) discusses this study, he notes that *everyone*, including his professional psychologist opponents (and presumably the reader) will agree about the similarity of these figures. So the argumentative force of this finding can be motivated simply be appealing to the reader's own visual experience. (Thanks to a referee for this journal for suggesting a parallel to the visual object recognition debate.)

[13] Note that this is not the same thing as claiming that philosophers are *better* at moral thinking, or are in some way immune to psychological distortions (see note 10). The idea is only that philosophers have a special skill at making introspective sense of their moral experience, through training and conceptual vocabulary – in the same way that trained jurists are skilled at introspecting upon their legal reasoning. It is a further question (not addressed here) whether this introspective ability leads to any improvement in reasoning outcomes.

theoretical vocabulary for reporting their experiences to others and making intersubjective comparisons. These are good starting points for a Titchenerian introspective moral psychology.

To be clear, I am not claiming that philosophers have any ability to introspectively determine the truth or falsity of psychological hypotheses. The claim is *not* that moral philosophers can, from the armchair, simply declare that computationalism or connectionism is false because it does not *feel* to them as if they think in a computationalist or connectionist way. The proposal is more subtle than this. What is needed is, first, for psychologists (or certain very psychologically-sophisticated philosophers) to isolate *narrow and specific* hypotheses about moral cognition, and, second, for moral philosophers to help translate these empirical hypotheses into the vocabulary of moral theory. Once this is done, moral philosophers may be asked to introspect on *narrow and specific* features of their own moral experience, to see how it comports with the hypotheses in question.

I'll now illustrate the idea with several examples. The first is short and relatively trivial: it is not an example where philosophical training is particularly necessary, nor one that has unique importance to psychological theory. The purpose of this example is only to highlight the logic of the method. More theoretically significant examples are coming in the next section.

Suppose we wish to know whether human moral psychology is intuitively consequentialist: that is, whether we are 'wired' to include only representations of the consequences of actions in our intuitive moral assessments, or whether other non-consequentialist factors play a role. To apply the introspective method, we do not simply ask moral philosophers if they feel like consequentialists. Rather, we ask them to reflect on *specific and concrete instances* of their own moral deliberation. For instance, we might ask them to think about their intuitive reactions to hypothetical scenarios that are typically taken to test the limits of consequentialist theory. One famous such case is that of the innocent accused: there have been vile crimes in a small town, and the populace will run murderously riot unless *someone* is held responsible (McCloskey 1957). To avoid these terrible consequences, may the authorities frame an innocent person and execute him?

For present psychological purposes, we do not care what is the morally correct answer to this question. Rather, we want to know whether people are psychologically disposed to accept killing the innocent accused, as the consequentialist psychology hypothesis predicts.[14] Moral philosophers can test this introspectively by examining their own reactions to the case. It may be that they are committed by their theory (if consequentialists) to endorsing killing the innocent accused. But does this feel *natural* to them? Is it intuitive? Or is it an imposition of abstract theoretical commitments over an inclination to say otherwise? According to the stalwart consequentialist J.J.C. Smart, that is exactly what the experience of considering this case is like:

> Even in my most utilitarian moods I am not *happy* about this consequence of utilitarianism. Nevertheless, however unhappy about it he may be, the utilitarian must admit that he draws the consequence that he might find himself in circumstances where he ought to be unjust [by killing the innocent]. Let us hope that this is a logical possibility and not a factual one. (Smart 1973, 71)

[14] There are complications here about what consequentialist theory is committed to saying about the case of the innocent accused. For some consequentialists (e.g. rule-consequentialists) it might be that killing the innocent is wrong after all, since the negative long-term consequences of subverting justice overwhelm short-term benefits. I leave such complications to the side here – they could be easily brought back in by changing the example in the text to a hypothesis that human moral cognition is simplistically *act*-consequentialist, or something of that sort.

Smart admits that he is logically committed (by his consequentialist theory) to endorsing killing the innocent accused. But he is "not *happy*" about doing so. Evidently, it makes him uncomfortable – it goes against his intuitions. In philosophical parlance, endorsing this conclusion requires Smart to "bite a bullet", to accept a counter-intuitive judgment about a particular case in order to protect a normative theory.

Because Smart can introspectively determine that this is a counter-intuitive conclusion, he now has grounds to infer that his own intuitive psychology is not disposed toward the consequentialist conclusion. Intuitions are, of course, the expressions or causal consequences of certain cognitive processes. So Smart could now articulate this introspective finding as a challenge to (descriptive) consequentialist moral psychology.[15]

As I warned, this is a fairly trivial example. It is not of much use to the actual practice of empirical moral psychology, because there are better ways to establish that humans are not intuitive consequentialists (i.e. through controlled naïve subject research – see e.g. Cushman et al. 2006). But the example gives us the template for a strategy that will have more utility in other instances, such as when we do not have other empirical means available, or when the particular virtues of philosophically trained introspection can be uniquely revealing. Features of

[15] Now is a good opportunity to stress, by way of reminder, that the ambitions of this paper are purely descriptive, not prescriptive. Concluding (if we do) that intuitive moral psychology is not consequentialist would *not* entail that consequentialism fails as a normative theory. Drawing that inference would require a number of supporting premises which are not to be discussed here. Interestingly, Peter Singer has drawn precisely the opposite conclusion: he claims that consequentialism is a *superior* moral theory *because* it is psychologically counter-intuitive! (Singer argues that intuitive psychology cannot track moral truth. See Singer 2005 and de Lazari-Radek and Singer 2012.) I will not discuss that claim either.

the *experience of* engaging in philosophical moral thinking – such as whether or not thinking in a certain way requires biting bullets – can be relevant to certain psychological hypotheses, and moral philosophers have the introspective training to reflect upon them productively.

Before returning to our central example of computationalism and connectionism, let me flag something interesting and rather exciting about this proposal. It appears to mark a distinctive and novel role for philosophers in advancing our understanding of moral cognition. It is already recognized that philosophers have a role to play as *theorists*, by contributing to the philosophy of science or by using their domain-expertise to help psychologists develop stimuli that are more responsive to conceptual distinctions (e.g. Kahane and Shackel 2010). But this proposal is different. It suggests that philosophers have an additional role to play as *participants* in research – as Titchener-style trained introspectors with a particular competence for understanding the experience of a certain mode of thought. Certainly this will not displace the need to conduct more pedestrian forms of psychological investigation, just as Titchener did not think that introspection was the *only* tool needed for perceptual psychology. But it does open an intriguing new avenue, one that should be investigated further whether or not much comes of the particular debate discussed in the next section.

3. From Moral Philosophy to Mental Architecture

3.1 Examples

Let's return to the central example of section 1, the debate between computationalists and connectionists. As noted, many participants in this debate would agree that these general theories of cognitive architecture generate divergent predictions about moral cognition. Computationalism, we saw, appears to imply that moral cognition will be largely rule-based.

Connectionism, in contrast, appears to imply an association-based account of moral cognition. What I will now argue is that certain instances of the experience of engaging in moral philosophy, facilitated by philosophers' training in introspection, can provide helpful tests of these rule- and association-based accounts of moral cognition – and through these, evidence for or against general theories of mental architecture.

I will discuss two areas of controversy in moral theory, showing how they might provide input to our central debate between rule-based (computationalist) and association-based (connectionist) theories of moral cognition. Importantly, the aim here is not to *complete* any instance of this argument. That is, I will not be trying to reach any particular conclusion about the plausibility of rule- or association-based moral cognition. A full defense of any particular conclusion would require quite a bit of supplementary argument, beyond this scope of this paper. The goal here is simply to make clear what *types* of arguments these are, and how they might work. Showing that there exist colorable arguments of this sort then gives us reason to claim that moral philosophers have distinctive contributions to make to cognitive science.

The two areas of controversy are about moral saints and moral principles. I have chosen these simply because their relation to the computationalism/connectionism debate is particularly clear; certainly other topics might have been chosen.

3.2 Moral Saints and Connectionism

First, consider moral saints. According to Susan Wolf (1982), a moral saint is "a person whose every action is as morally good as possible, a person, that is, who is as morally worthy as can be (419)." Wolf argues that we do not and should not want to be moral saints; moral saintliness "does not constitute a model of personal well-being toward which it would be particularly rational or good or desirable for a human being to strive (ibid).” This is because a moral saint’s single-minded devotion to maximizing moral behavior would make her incapable of enjoying other valuable things in life, and also something of a boring person.

Wolf’s paper has been very influential in moral theory (though see Carbonell 2009 for a recent dissent). Note how Wolf reaches her conclusion. She starts from a vague sense that something is amiss in the life of a moral saint. Then she applies philosophical techniques of introspection to examine her own reaction and try to work out what about moral sainthood is causing the problem.[16] So far this is just the business of moral philosophy. What I suggest is that we can connect these reflections to psychological claims. In particular, we might use them to motivate a case against association-based (connectionist) moral cognition.

Recall the main claim of associated-based moral cognition: moral cognition works by checking for similarity relations between new stimuli and learned category prototypes. But what else is a moral saint, other than a particular prototype – in fact, *the* moral prototype? As Wolf describes it, a moral saint is “as morally worthy as can be”. This generates a problem for association-based moral cognition. If our moral cognition is essentially pegged to prototypes, we

[16] Wolf starts with a “common sense” impression that the “strangely barren” life of a moral saint is one lacking “many of the interests and personal characteristics that we generally think contribute to a healthy, well-rounded, richly developed character. (421)” She expands on this by considering her reaction to characters in fiction (422-423), ruling out an alternative explanation for her reaction (423-425), and rebutting the worry that this reaction is just excuse-making for moral laxness (426-427). At each step in this process Wolf grounds her conclusions by reflecting on her own reactions to the lives of *particular* people, from Groucho Marx to Mother Theresa.

shouldn't find *the* moral prototype repellent![17] And yet, according to Wolf, that is how the moral saint comes across. So if Wolf is right, then a prediction of association-based moral cognition fails, and this is a piece of evidence against connectionist cognitive architecture.

Of course, there are plenty of ways for the association-based theorist to resist this conclusion. One is to argue that the 'saint' is *not* a prototype for morality after all. Vanessa Carbonell (2009) points out that Wolf's saint appears to be obsessed with morality-for-its-own-sake, rather than the valuable things that morality is meant to make us attend to. She writes:

> Imagine approaching a stranger and asking him 'What's your life's passion?' and getting the answer 'Morality.' Or asking 'What are you going to do today?' and getting the answer 'Morally good things.' But is that how a real moral saint would answer? The moral saint as I conceive of him would answer with the *content* of his moral commitments, not the fact that he is so committed. To 'What's your life's passion?' he would answer 'healing the sick' or 'eradicating tuberculosis.' To 'What are you going to do today?' he would answer, in similar fashion, 'see patients' or 'raise money.' (Carbonell 2009, 391-392)

Essentially, the claim goes, Wolf's 'saint' is objectionable because he does not *do* morality in the right way – his motivations are of the wrong sort to count as a moral saint. What bothers us about this character, according to Carbonell, is not that we find prototypes of

[17] Note that this is not a problem for rule-based moral cognition, though it might seem that it would be (Doesn't Wolf's moral saint also follow all the moral rules?). It is essential to the association-based account that our moral judgments involve comparison to a prototype. It is *not* essential to the ruled-based account that our moral judgments involve comparison to a person-who-has-followed-a-rule. Hence the fact that we (according to Wolf) find such a person repellent is much more troubling for the former than the latter.

morality aversive, but that he isn't a moral prototype. He is a sort of mockery of moral motivation, one who cares about claiming the banner of morality rather than about its contents.

There are some very deep issues here, including whether moral commitments must be intrinsically motivating (see Smith 1994). We needn't get into that at the moment. I wish only to highlight the fact that the status of at least one prediction of cognitive science now appears to turn on the resolution of a debate within substantive moral theory. Association-based moral cognition predicts that we should find moral prototypes compelling, not aversive. Wolf's observations about her moral saint present a failure of this prediction *if* her saint character really is a moral prototype. Whether or not Wolf's character really is a moral prototype appears to come down to substantive ethical reflection, the result of introspective consideration of what it is that Wolf makes us think of, and why we find it troubling.[18]

There is, obviously, intrinsic philosophical value in thinking about the issue raised by Wolf's paper; even Robert Adams, who rejects nearly all Wolf's conclusions, agrees that it "brings out very sharply a fundamental problem in modern moral philosophy (Adams 1984,

[18] Association-based theorists might offer another response (I owe this objection to a reviewer for this journal). Perhaps the prototype employed in association-based moral cognition is not a *person* prototype but instead an *act* prototype. That is, perhaps our moral judgments involve implicit comparison to prototype actions, independent of who does them. If this is the case, then moral saints are no problem for association-based theory; using actions as moral prototypes is consistent with finding an individual exhaustively constituted by such actions repellent. Note, however, that many association-based theorists have been explicit in saying that we *do* use virtuous persons, not just acts, as moral prototypes, and have taken this to entail support for Aristotelian normative theory (Churchland 1996b, Casebeer and Churchland 2003).

392).” In addition to that aim, I claim that we can now add a second benefit to Wolf’s reflections. They provide us a special bit of data for the claims of cognitive science, data facilitated by the introspective skills of philosophers.

3.3 Moral Principles and Computationalism

I turn now to a second moral philosophical controversy with relevance to cognitive science. The discussion of moral saints looked at a potential challenge to association-based (connectionist) moral cognition; this one will examine a challenge to rule-based (computationalist) moral cognition.

The debate in question concerns the status of moral *principles*. In contemporary ethics, it is widely assumed that a central aim of moral reflection is to systematize particular intuitive judgments into widely-applying moral principles. Here is John Rawls’ (“provisional”, he warns) characterization of the task of moral theory:

> [W]hat is required is a formulation of a set of principles which, when conjoined to our beliefs and knowledge of the circumstances, would lead us to make [our moral] judgments with their supporting reasons were we to apply these principles consistently and intelligently… We do not understand our sense of justice until we know in some systematic way covering a wide range of cases what these principles are. (Rawls 1971, 46)

In Rawls’ telling, the difficulty in moral theory is in working out *which* principles to favor, and how to prioritize principles relative to one another. This approach has been immensely influential in contemporary moral theory (e.g. Parfit 1984, Kamm 2006). But it is not without its critics. Here Bernard Williams expresses grounds for hesitation:

> In the ethical case, inasmuch as the problem is seen as the explanatory problem of representing people's ability to make judgments about new cases, we do not need to suppose that there is some clear discursive rule underlying that capacity. Aristotle supposed that there was no such rule and that a kind of inexplicit judgment was essentially involved, an ability that a group of people similarly brought up would share of seeing certain cases as like certain others. (Williams 1985, 97)

Annette Baier makes the same claim more forcefully: "It is a mere Kantian dogma that behind every moral intuition lies a universal rule, behind every set of rules a single stateable principle or systems of principles (Baier 1985, 208)." Views of this sort feature in several distinct criticisms of principle-based methodology, including 'anti-theory' (Clark and Simpson 1989) and moral particularism (Dancy 2004, Hooker and Little 2000).[19] There are important differences among these views, but they share a commitment to the claim that moral knowledge is immensely rich and complex, too rich to be captured in discrete principles, and therefore best understood as a sort of practical skill.

The status of moral principles clearly matters to the contrasting claims of rule-based and association-based moral cognition. Moral rules, in fact, may be nothing more than principles – and this relation to principle-based moral methodology has been noted by a number of computationalists. In *A Theory of Justice*, only one page after the quotation featured above,

[19] Dancy motivates his moral particularism through introspective reflection on *how* we apply moral reasons in consideration of test cases. Dancy claims that the situational invariance of certain moral reasons does not show them to be principles: it is "not a matter of the logic of such reasons, but more the rather peculiar fact that some reasons happen to contribute in ways that are not affected by other features" (Dancy 2004, 78)." This distinction, between the *logic* of reasons and how they happen to affect one another, depends on an introspective analysis of the process of moral reflection.

Rawls makes an off-hand suggestion that there might be a useful analogy between moral knowledge and Chomskyan linguistics; this remark is regularly cited by proponents of the "moral grammar" research program.[20] The rule-based (computationalist) approach strongly predicts that there are implicit principles operating behind our moral intuitions, and that these should be recoverable through careful examination of our moral practice.

The claims about moral principles quoted above are based on philosophers' introspective assessment of their own moral experience. Rawls expects to find principles readily present when one is careful and systematic in comparing individual moral judgments. Baier regards this expectation as a "mere Kantian dogma". These are not empirical claims of the sort directly relevant to psychological theory. But they do express the sincere introspective observations of moral philosophers, who like Titchener's subjects are the best equipped to evaluate thinking within their domain.

The dispute over moral principles is not resolved, or likely to be resolved anytime soon. So whatever lessons its resolution might have for moral cognition are not imminent. But we can still trace the logical chain in order to see the implications of views within this debate. If you are convinced that moral principles do a poor job of characterizing your introspective sense of what is at stake in moral theory, then you ought to be suspicious of rule-based moral cognition, and

[20] Rawls would likely have recognized the usefulness of philosophical introspection as input to psychological theory. His early writing, especially 'Outline of a Decision Procedure for Ethics' (Rawls 1951) emphasizes the need to recover "principles [that] are implicit in the considered judgments of competent judges 183". See Mikhail (2011) and Rini (2011, chapter 3) for an interpretation of Rawls that makes clear the centrality of his psychological commitments.

therefore somewhat suspicious of computationalist architecture. Inversely, if you favor computationalism and rule-based moral cognition, then you have a stake in the success of principle-driven methodology in moral theory. These disparate subjects - cognitive architecture and normative ethics - are strung together in a web of connective inferences, and the lessons go both ways.

4. Cognitive Science in a Humanistic Key

In this paper I have considered a popular form of modus ponens argument, one inferring from cognitive scientific premises to conclusions about moral thinking, and pointed out that this argument allows for a modus tollens reversal, inferring from findings about moral thinking back to general theories of cognitive architecture. I then argued for an unappreciated way in which we might learn about moral thinking itself – through careful introspective examination of the experiences of philosophers engaged in the construction of moral theory. I have illustrated this pattern with the contrast between computationalist and connectionist theories, as applied to debates in moral theory about the intuitive status of prototypes (saints) and principles.

I'll conclude with some brief remarks on the limitations and the possibilities of the foregoing argument. First, it must be admitted that the modus tollens reversal I have championed can be no more plausible than the modus ponens it reverses. That is, if you doubt (for instance) that connectionism leads to an association-based account of moral cognition, then you obviously aren't going to see a failure of association-based moral cognition as any trouble for connectionism. Of course, this is just a feature of the logic of modus tollens.

More importantly, it must also be accepted that that the tollens reversals I have described are systematically *less powerful* than the modus ponens arguments they reverse. This is because a tollens reversal must begin from making an interpretive assertion about a *particular* a posteriori outcome – e.g. that a prediction of rule-based moral cognition has failed in a specific instance. For the committed proponent of a theory of cognitive architecture, it will always be possible to resist such arguments in a range of ways: by challenging the empirical validity of the test, or the interpretation of its constructs, etc. But this is just a feature of the nature of science; theory-testing is slow and details-intensive, and there are rarely indisputable outcomes.

So it is not a problem to agree that the tollens reversals I describe have no more than an abductive, piecemeal import, especially when it comes to philosophers' introspective observations. Certainly, one would not abandon confidence in a general theory of cognitive architecture simply because a single experimental finding came back appearing to challenge its predictions. Similarly, I do not claim that introspective reflection on a single aspect of moral theory – moral saints or moral principles, for instance – would lead to anything like a decisive conclusion for general theories of cognitive architecture. The claim is only that introspective reflection on philosophical practice can be *included among the data* to be explained by a theory of cognitive architecture.

There will always be plenty of room to dispute the interpretation of a given introspective claim, particularly regarding translation between cognitive scientific terms and those of moral theory. For instance, in the discussion of rule-based cognition and principle-based method above, I did not defend the claim that moral principles are relevantly connected to 'rules' in the cognitive scientific sense (and especially in the specific sense of syntax-driven symbol-manipulation employed in computationalism) . To *complete* the argument, quite a bit would need to be said about the relationship between these terms. This would require careful conceptual work in both cognitive science and moral theory, and would be open to challenge from both directions.

But it is possible to see this multi-directional vulnerability as a virtue, not a vice.[21] The need to coordinate these concepts across disciplinary boundaries should encourage the continuation of fruitful dialogue and collaboration among psychologists and philosophers. Any philosopher browsing the 'Discussion' sections of recent empirical research on moral cognition surely appreciates how it might benefit from engagement with the careful conceptual work in normative theory – and surely analogous things might be said for the empirically-infused presuppositions concerning motivation and reasoning that philosophy has more traditionally called 'moral psychology'.[22]

Finally, although I've written here about cognitive science and the study of morality, the possibility of reversing modus ponens arguments from general cognitive architecture likely

[21] As Peter Railton writes, in a different context, "It is an attraction for me of naturalism in ethics… that it thus is constrained in several significant dimensions at once. One has such ample opportunities to be shown wrong or found unconvincing if one's account must be responsive to empirical demands as well as normative intuitions (Railton 1986, 205)."

[22] For an encouragingly rich and careful example of the former, see discussion of Greene (2008). Greene's views concerning moral judgment depend on experimental studies classifying particular stimulus scenario responses as typically consequentialist or deontic. Berker (2009), Kamm (2009) and Kahane and Shackel (2010), arguing partly by more traditional moral philosophical means, suggest that this classification is mistaken. For an example of the latter, see recent empirical investigation of psychopaths and its potential relevance to disputes concerning moral motivation (e.g. Blair 1995, McGreer 2008).

generalizes to many other areas of inquiry. Cognitive science currently enjoys a (mostly well-deserved) high place in academic and popular discourse, driving new discoveries and reinterpretations in a range of disciplines, from literature (Zunshine 2006), to cultural studies (McConachie 2010), to the visual arts (Dutton 2010). The idea seems to be that developed cognitive science gives us a sort of skeleton key to the human.

But if what I've said here is correct, then inferences from cognitive science into other disciplines also generate feedback from the target domain; failure of the purported implications is a challenge to the recommending cognitive theory. Fundamental cognitive science has lessons to teach many sorts of inquiry, but it may have a corresponding amount to learn from them. What this suggests is a much less hierarchical, much more holistic, relationship among cognitive science and humanistic pursuits. All of us, cognitive scientists, ethicists, and others, are together engaged in a deeply continuous project. We ought to welcome a little feedback.[23]

Citations

[23] **Acknowledgments:** I owe particular thanks to Ned Block, Linus Huang, Joshua Knobe, Sharon Street, Michael Strevens, Aaron Wilson, and two anonymous reviewers for thoughtful comments on earlier versions of this paper, and to participants in the University of Miami Graduate Conference on Mind and Ethics, and the NYU Thesis Prep Seminar. The argument of this paper has also benefited from the suggestions of Otávio Bueno, Justin Clarke-Doane, Jonny Cottrell, Daniela Dover, Dana Evan, Thomas Nagel, Daniel Robinson, Jeff Sebo, Jon Simon, Michael Slote, Seunghyun Song, and Amie Thomasson. Development of this research benefited from the support of the VolkswagenStiftung's European Platform for Life Sciences, Mind Sciences, and the Humanities (grant II/85 063).

Adams, R.M. 1984. Saints. *The Journal of Philosophy*. 81(7): 392-401.

Appiah, K.A. 2008: *Experiments in Ethics*. Cambridge: Harvard Univ Press.

Baier, A. 1985: Theory and reflective practices. In her *Postures of the Mind*. Minneapolis: University of Minnesota Press. 207-227.

Berker, S. 2009: The normative insignificance of neuroscience. *Philosophy and Public Affairs* 37:293-329.

Biederman, I. and M. Bar. 1999: One-shot viewpoint invariance in matching novel objects. *Vision Research* 39: 2885-2899.

Blair, R.J.R. 1995: A cognitive developmental approach to morality: Investigating the psychopath. *Cognition* 57(1): 1-29.

Casebeer, W. and Churchland, P. 2003: The neural mechanisms of moral cognition: A multi-aspect approach to moral judgment and decision making. *Biology and Philosophy* 18:169-194.

Chappell, T. 2013. Why ethics is hard. *Journal of Moral Philosophy*. Online First. doi:10.1163/17455243-4681028.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Churchland, P. 2011: *Braintrust: What Neuroscience Tells Us About Morality*. Princeton University Press.

Churchland, P. 1996a: *The Engine of Reason, The Seat of the Soul: A Philosophical Journey Into the Brain*. Cambridge: MIT Press.

Churchland, P. 1996b: The neural representation of the social world. In May, Friedman and Clark (1996): 91-108.

Churchland, P. 2000: Rules, know-how, and the future of moral cognition. *Moral Epistemology Naturalized: Canadian Journal of Philosophy Supplementary Edition* 26:291-306.

Clark, A. 2000: Word and Action: Reconciling Rules and Know-How in Moral Cognition. *Moral Epistemology Naturalized: Canadian Journal of Philosophy Supplementary Edition* 26:267-290.

Clark, S.G. and Simpson, E. (eds) 1989: *Anti-Theory in Ethics and Moral Conservatism*. Albany: SUNY Press.

Cushman, F., Young, L. and Hauser, M. 2006. The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science* 17(12): 1082-1089.

Cushman, F. and Young, L. 2011: Patterns of Moral Judgment Derive from Nonmoral Psychological Representations. *Cognitive Science* 35: 1052-1075.

Dancy, J. 2004: *Ethics Without Principles.* Oxford: Clarendon Press.

Dutton, D. 2010: *The Art Instinct: Beauty, Pleasure, and Human Evolution*. New York: Bloomsbury Press.

Dwyer, S. 2006: How Good is the Linguistic Analogy? In P. Carruthers, S. Laurence ad S. Stich, *The Innate Mind: Culture and Cognition*. New York: Oxford University Press. 237-256.

Fodor, J. 1971. *The Language of Thought*. New York: Thomas Crowell.

Fodor, J. and Z. Pylyshyn. 1988. Connectionism and Cognitive Architecture: A Critical Analysis. In S. Pinker and J. Mehler (eds.) *Connections and Symbols*. Cambridge: MIT Press.

Gazzaniga, M.S. 2006: *The Ethical Brain: The Science of Our Moral Dilemmas*. New York: Harper Perennial.

Greene, J.D. 2008: The secret joke of Kant's soul. in Sinnott-Armstrong (2008, v.3): 35-80.

Hauser, M.D., Young, L. and Cushman, F. 2008. Reviving Rawls's linguistic analogy: Operative principles and the causal structure of moral actions. In Sinnott-Armstrong (ed.) (2008, v.2) 107-144.

Hooker, B. and Little, M. 2000: *Moral Particularism.* Oxford: Oxford University Press.

Horst, S. 2009. The Computational Theory of Mind. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2011 edition). http://plato.stanford.edu/archives/spr2011/entries/computational-mind/

Hummel, J.E. 2000. Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich and A. Markman (eds.) *Cognitive Dynamics: Conceptual and Representational Change in Humans and Machines*. Hillsdale, NJ: Psychology Press. 157-185.

Hummel, J.E and Biederman, I. 1992. Dynamic binding in a neural network for shape recognition. *Psychological Review* 99: 480-517.

Hummel, J.E. and Stankiewicz, B.J. 1996. Categorical relations in shape perception. *Spatial Vision* 10: 201-236.

Kahane, G. and Shackel, N. 2010: Methodological issues in the neuroscience of moral judgment. *Mind and Language*. 25(5): 561-582.

Kamm, F. 2006: *Intricate Ethics: Rights, Responsibilites and Permissible Harm*. Oxford University Press.

Kamm, F. 2009: Neuroscience and moral reasoning: a note on recent research. *Philosophy and Public Affairs* 37(4): 330-345.

de Lazari-Radek, K. and Singer, P. 2012: The Objectivity of Ethics and the Unity of Practical Reason. *Ethics* 123(1): 9-31.

Ludwig, K. 2007: The Epistemology of Thought Experiments: First Person Versus Third Person Approaches. *Midwest Studies In Philosophy* 31 (1): 128–159.

Marr, D. 1982: *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* New York: Freeman.

May, L., Friedman, M. and Clark, A. (eds) 1996: *Minds and Morals: Essays on Ethics and Cognitive Science*. Cambridge: MIT Press.

McCloskey, H.J. 1957: An examination of restricted utilitarianism. *Philosophical Review* 66(4): 466-485.

McConachie, B. 2010: Toward a cognitive cultural hegemony. In L. Zunshine (ed.) *Introduction to Cognitive Cultural Studies*. Johns Hopkins University Press. 134-150.

McGreer, V. 2008: Varieties of moral agency: Lessons from autism (and psychopathy). In Sinnott-Armstrong (2008, v.3) 227-257.

Mikhail, J. 2007: Universal moral grammar: theory, evidence and the future. *TRENDS in Cognitive Science* 11(4): 143-151.

Mikhail, J. 2011: *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. New York: Cambridge University Press.

Parfit, D. 1984: *Reasons and Persons*. Oxford University Press.

Piccinni, G. 2009. Computationalism in the Philosophy of Mind. *Philosophy Compass* 4(3): 515-532.

Prinz, J. 2009: *The Emotional Construction of Morals*. Oxford University Press.

Putnam, H. 1961. Brains and Behavior. Reprinted in Block (ed.) 1980. *Readings in Philosophy of Psychology* Cambridge MA: Harvard University Press.

Railton, P. 1986: Moral Realism. *The Philosophical Review* 95(2): 163-207.

Rawls, J. 1951. An Outline of a Decision Procedure for Ethics. *Philosophical Review* 60(2): 177-197.

Rini, R.A. 2011: *Within is the Fountain of Good: Moral Philosophy and the Science of the Nonconscious Mind*. PhD Thesis. NYU Department of Philosophy.

Rumelhart, D.E., J. McClelland and the PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge MA: MIT Press.

Schwitzgebel, E. 2011: *Perplexities of Consciousness*. MIT Press.

Schwitzgebel, E., and F. Cushman. 2012: Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers. *Mind and Language* 27 (2): 135–153.

Singer, P. 2005: Ethics and Intuitions. *The Journal of Ethics* 9(3/4): 331-352.

Sinnott-Armstrong, W. (ed.) 2008: *Moral Psychology*. (three volumes) Cambridge: MIT Press.

Smart, J.J.C. and Williams, B. 1973: *Utilitarianism: For and Against*. Cambridge University Press.

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

Smolensky, P. 1998. The Proper Treatment of Connectionism. *Behavioral and Brain Sciences*. 11(1): 1-74.

Stevenson, C.L. 1944: *Ethics and Language*. Yale University Press.

Stich, S. 2006: Is morality an elegant machine or a kludge? *Journal of Cognition and Culture* 6(1-2):181-189.

Tarr, M.J. 1995. Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonometric Bulletin & Review* 2: 55-82.

Titchener, E.B. 1912: The schema of introspection. *American Journal of Psychology* 23: 485-508.

Weinberg, J.M., C. Gonnerman, C. Buckner, and J. Alexander. 2010: Are Philosophers Expert Intuiters? *Philosophical Psychology* 23 (3): 331–355.

Williams, B. 1985: *Ethics and the Limits of Philosophy*. Harvard University Press.

Williamson, T. 2011: Philosophical Expertise and the Burden of Proof. *Metaphilosophy* 42 (3): 215–229.

Wolf, S. 1982: Moral saints. *The Journal of Philosophy* 79(8): 419-439.

Zimbardo, P. 2007: *The Lucifer Effect: Understanding how Good People Turn Bad*. New York: Random House.

Zunshine, L. 2006: *Why We Read Fiction: Theory of Mind and the Novel*. Ohio State University Press.