Moral Agency and Moral Learning: Transforming Metaethics from a First to a Second Philosophy Enterprise

William A. Rottschaefer
Lewis and Clark College

ABSTRACT: Arguably, one of the most exciting recent advances in moral philosophy is the ongoing scientific naturalization of normative ethics and metaethics, in particular moral psychology. A relatively neglected area in these improvements that is centrally important for developing a scientifically based naturalistic metaethics concerns the nature and acquisition of successful moral agency. In this paper I lay out two examples of how empirically based findings help us to understand and explain some cases of successful moral agency. These are research in moral internalization and aggression management. Using these examples, I sketch some lessons for investigating successful moral learning and moral action. My proposal reflects a common theme in scientifically based philosophy generally: the shift from the armchair methods of analyzing concepts and finding a priori foundations, the enterprise of first philosophy, to an effort to study the phenomena themselves, using empirical findings and theories to answer philosophical questions about these phenomena, an endeavor recently characterized as second philosophy.

Key words: aggression management, experimental philosophy, first philosophy, moral agency, moral learning, moral internalization, second philosophy

Arguably, one of the most exciting recent developments in moral philosophy is the ongoing scientific naturalization of both normative ethics and metaethics, in particular moral psychology. In a series of papers John Doris and Stephen Stich (2005, 2006) have described this work in some detail, laying out how empirical studies and empirically based theoretical advances in evolutionary theory, biology, neuroscience, cognitive science, social psychology, and anthropology have provided findings, hypotheses, and empirically supported theories relevant to the traditional subject matter of normative ethics and metaethics. For instance, they have highlighted studies demonstrating the problems with the predominant intuitional methodology of moral philosophers. They note that empirical studies of folk conceptions of moral responsibility have opened up further questions about philosophical analyses of what features of agency are necessary for moral agency.

AUTHOR’S NOTE: A version of this paper was presented at the Philosophy of Science Association 2006 meeting in Vancouver, British Columbia. Please address correspondence to William A. Rottschaefer, Department of Philosophy, Lewis and Clark College, Portland, OR 97034; Email: rott@lclark.edu. I thank two anonymous referees and the editor of this journal, Jack Marr, for their helpful comments and suggestions.
and they discuss cross-cultural studies that challenge traditional claims about the distinction of the moral realm from conventional, religious, and prudential realms. They describe studies in social psychology that pose problems for virtue ethics in its emphasis on the centrality of character as opposed to situation in moral action. They reflect on findings concerning moral agreement and disagreement that affect central discussions about what counts as belonging to the moral realm and about its consequences for claims concerning moral realism and anti-realism. In addition, multiple and large-scale studies of moral judgment have provided a wealth of information about people’s responses to moral dilemmas such as the trolley dilemma, and this information has led to much theorizing about the nature of moral judgment (Hauser, 2006). Doris and Stich (2005, 2006) also review studies in evolutionary biology and psychological theories of altruism and fairness, as well as experimental findings, which have provided a wealth of materials that extend and challenge traditional moral philosophical positions on these central ethical topics. In addition, moral philosophers Shaun Nichols (2004), Jesse Prinz (2007), and Joshua Greene (2003), among others, are developing metaethical theories that make extensive use of empirically based moral psychology and its discoveries about the role of emotions in moral judgment, challenging in various ways traditional rationalistic metaethical positions. Further, these moral philosophers and others have begun to do their own empirical work in pursuit of answers to the traditional questions of normative ethics and metaethics (Sinnott-Armstrong, 2008).

In this paper I want to point to and illustrate an area of empirical research, findings, hypotheses, and theories that I contend is centrally important for developing a scientifically based naturalistic metaethics. This is the investigation that flies under the banner of social cognitive theory in empirical psychology. It is an area that has been relatively neglected in the recent attempts to bring scientific sources to bear on the study of morality.

I shall begin by describing briefly what I understand to be a scientifically naturalized account of metaethics, contending that its central focus should be on providing an account of moral agency. Next I will concentrate on one aspect of that account: learning to be a moral agent, where I understand moral learning to be concerned with the development and activation of capacities central to successful moral agency. I shall lay out two examples of how empirically based findings help us to understand and explain some cases of successful moral agency. These are research in moral internalization and aggression management. Using these examples, I shall sketch some lessons for investigating successful moral learning and moral action. Finally, I shall assess some major objections to such a project and describe some challenges in its pursuit.

Metaethics as the Study of Moral Agency

As is well known, traditional Anglo-American analytic ethics divides itself into three sub-areas: metaethics, normative ethics, and applied ethics. Applied ethics addresses moral questions about the rightness and wrongness of various
actions and practices that occur in everyday life, business, medicine, law, etc., making use of the theories of moral rightness and wrongness that are developed in normative ethics, such as ethical egoism, virtue theory, consequentialism, Kantian theories of duty, and natural law theories. Such normative theories develop accounts of moral obligation and moral permissibility as well as theories of what counts as morally right and wrong, morally good and bad, and the requirements for acting in a morally responsible fashion. Metaethics traditionally presents itself as an armchair discipline, making use of conceptual and a priori techniques. It involves an attempt to analyze the central concepts of morality that are employed in and presupposed by normative ethics, for instance, GOOD, BAD, RIGHT, WRONG, and RESPONSIBILITY. In addition, making use of conceptual analysis, a priori reflection, and intuition, it addresses itself to the epistemic status and ontological implications of moral judgments and, in the process, makes various sorts of assumptions about human moral psychology. My focus is on some of the moral psychology that lies in the background of metaethics.

I have proposed elsewhere that metaethics be reconceived as a scientifically based naturalistic philosophical investigation of the phenomenon of moral agency (Rottschaefer, 1998). This investigation examines moral agency’s various dimensions—biological, psychological, social, and cultural—and, in so doing, addresses the ontological status of moral agents, their cognitive, motivational, and emotional capacities, as well as their actions and the contexts in which these actions are performed. On this conception of metaethics it pursues questions about the nature, acquisition, activation, and well functioning of moral agency. Questions about how we learn to be moral agents focus on how we acquire, develop, and maintain the various capacities that constitute well-functioning moral agency.

A larger contention that I shall not argue for here is that such a scientifically based naturalistic account of moral agency better solves the semantic, ontological, metaphysical, and epistemic issues addressed in metaethics than does the traditionally conceived armchair methods of first philosophical approach. My proposal reflects a common theme in scientifically based philosophy generally—what Penelope Maddy (2007) has recently felicitously called second philosophy: the shift from the armchair activities of analyzing concepts and finding a priori foundations, an enterprise of first philosophy, to an effort to study the phenomena themselves by using empirical findings and theories to answer philosophical questions about these phenomena. While Maddy addresses the significance of scientific findings for understanding metaphysics, mathematics, and logic, I turn my attention to metaethics.

Metaethical questions about what constitutes moral agency, how it is acquired, how it is activated, and what makes for successful performance tread on territory that scientists often contend they do not address and that are frequently thought to be beyond the proper of domain of scientists. In the case of moral agency, this territory includes the existence and nature of moral values and norms, how they are known, and how they are achieved. However, precisely because the naturalist postulates that moral values are to be found in human individual, social, and cultural life, if they are to be found anywhere, and because the scientific
naturalist contends that the sciences are our best way to understand human individual, social, and cultural life, the second philosophical student of metaethics argues that the questions of metaethics are best pursued by making use of best current findings and theories of the relevant sciences (Rottschaefer, 1998).

I now turn to illustrating how this might work concerning the question of how we become successful moral agents.

Social Cognitive Theory and Learning to be a Moral Agent

Although the new metaethics of second philosophy described by Doris and Stich has provided an admirably rich range of scientifically based materials to pursue various sorts of metaethical questions, it has paid little attention to the issue of learning to be a moral agent. The issue is implicitly addressed in some of the research used in discussing the relative importance of situation and character (Doris, 2002) and in determining the relevant differences in the moral judgments of psychopaths, criminals, variously impaired agents, and normal moral agents (Nichols, 2004; Sinnott-Armstrong, 2008). However, I maintain that there is a notable lack of attention to the question of what makes for successful moral agency, as well as to the extensive empirical research relevant to answering this question that stems from social cognitive theory and studies of very young children.

I note that the widespread attention paid to moral judgment by the new second philosophical metaethicians eerily parallels the attention paid to moral judgment—to the exclusion of moral action—in the classical cognitive moral developmental theories of Piaget, Kohlberg, and their successors. As is well known, one of the major failures of this tradition was its inability to account adequately for successful moral action. Ironically, it was social cognitive theorists such as Walter and Helen Mischel (1976) who pointed out this failure and helped develop a social cognitive approach that has potential for explaining moral success. Psychologists within the moral cognitive developmental tradition also recognized this problem, for instance Augusto Blasi (1980, 1983).

Social cognitive theories are rooted in the cognitive revolution in psychology that transformed behaviorism in the 1960s. They have retained behaviorism’s emphasis on the role that the environment plays in bringing about behavior, including the role of social environmental factors in successful or dysfunctional human agency. They have also preserved behaviorism’s emphasis on careful controlled studies. But, in addition, social cognitive theories have expanded behaviorism’s explanatory toolbox to include the cognitive, emotional, and motivational factors influencing the actions of both an agent and the people that make up her social environment. The result has been an impressive array of careful studies providing well-confirmed findings and hypotheses concerning the factors influencing the successful or less than successful performance of a wide range of human activities and skills. These include cognitive, motivational, and emotional functioning, skills in achieving physical and psychological wellbeing, athletic skills, career competencies, and social aptitudes (Bandura, 1986).
I maintain that much of this work is importantly relevant for second philosophers who are attempting to understand and explain successful moral agency, in particular moral learning. To make my case, I shall discuss work done on moral internalization, deriving from advances in moral developmental psychology, and on the management of aggression, anger, and conduct disorder, stemming from work using social cognitive theory.¹

**Empirical Findings About Reliable Mechanisms of Moral Learning**

We can distinguish substantive and functional issues concerning moral learning. Substantive issues concern the areas of human action that count as being in the moral realm. Functional issues, on the other hand, concern what psychological factors are required for an action concerning a moral matter to be a moral action, whether morally correct or incorrect. Thus, for instance, from the substantive perspective we distinguish matters of etiquette—how to place the table settings—from moral matters such as not harming an innocent person. From the functional perspective, an action is functionally moral, if, for instance, it is done with adequate knowledge and intention.

I take moral agency to be a complex of interacting capacities concerned with actions about substantively moral matters and performed in a functionally adequate fashion (Rottschaefer, 1998). Given these assumptions, we can ask questions about the nature of these capacities, how they are acquired and activated, and whether and how they are functioning properly. I shall take moral learning to involve the processes by which we acquire our moral capacities and skills and successfully put them to work. A consequence of successful moral learning is well-functioning moral capacities and a well-functioning moral agent. The moral learning hypothesis is that inquiry into moral learning reveals the kinds of moral capacities that we have and the ways in which they function well or not. I argue that findings of successful moral learning, as indicated by moral internalization and the management of anger and aggression, are indicative of the presence of reliable perceptual, cognitive, emotional, and motivational capacities in a moral agent. Such capacities enable moral success.

Despite the continuing disagreements about what are the proper substantive and functional criteria for moral action, psychologists of different persuasions have settled upon some operational substantive and functional criteria for successful moral action. They agree that voluntary behaviors intended to benefit another, such as helping, sharing, and comforting behaviors—behaviors that are prosocially motivated, that is, motivated by sympathy for others or by the desire to adhere to internalized norms—are within the moral realm both substantively and functionally (Eisenberg, 1992).

¹ Work on moral learning through the regulation of verbal behavior (Hayes et. al., 1998) and self control (Rachlin, 2000) offer complementary, fruitful sources for the investigations for moral learning. I thank Jack Marr for pointing this out to me.
Beginning in the 1980s, an upsurge in the study of infants and very young children, as young as one to two years of age, showed that moral development begins much earlier than moral cognitive developmental psychologists such as Piaget and Kohlberg had supposed (Grusec & Kuczynski, 1997). These results have led to a refocusing on the moral learning in which parents and caregivers play an important role. Researchers generally agree that parents and siblings are the primary agents within the family for the promotion of moral learning. Focusing on parents and children, research in this area has promoted the investigation of a number of different causal factors that potentially play a role in moral learning. These include parental discipline techniques, styles of parenting, the mutual influence of parent and child upon each other, modeling, and instruction. I focus on discipline techniques.

**Parental Discipline Techniques and Reliable Mechanisms of Moral Action**

Starting with research done in the 1950s and coming to focus in the 1970s, developmental psychologists have studied the relative effectiveness of three distinctive parental discipline techniques that caregivers use in bringing about what psychologists call moral internalization (Hoffman, 1970, 1975, 1977). Moral internalization is a way that developmental psychologists often describe what moralists have discussed as the development of conscience (Hoffman 2000; Kochanska & Thompson, 1997). It designates a psychological state, and its development, in which one feels or believes that he or she has an obligation to act in accord with moral norms. Developmental psychologists mean various things by moral norms, but one accepted and non-biasing version is that a moral norm is a norm that enjoins one in a specific situation to act for the welfare of another. An agent achieves moral internalization when, in situations where there is a conflict of interests between the welfare of another and her own interests, she consistently acts to promote the welfare of another person rather than to attain social approval or egoistic aims.

The three prominent parental discipline techniques are: (1) assertions of power, (2) withdrawal of love, and (3) induction (Eisenberg, 1992; Hoffman, 1988; Macoby, 1992). Assertions of power involve such measures as the use of force, deprivation of privileges, threats, and commands. Love withdrawal includes expressions of parental anger and disapproval. In the use of inductive techniques, caregivers point out to the child, either directly or indirectly, the effects of the child’s behavior on others, provide information about moral norms, and communicate their values regarding the consideration of others. The association of a moral norm both with empathic feelings, particularly empathic distress, and with guilt feelings makes it a “hot cognition,” one that has motivational power. It can thus enter into future considerations as a motivator independently of any concerns about approval or disapproval or fear of punishment.

Both naturalistic and experimental studies since the late 1950s and the early 1960s indicate that the most effective means of moral internalization are inductive
techniques (Eisenberg, 1992; Hoffman, 1970, 1977, 2000; Macoby, 1982; Moore & Eisenberg, 1984; Radke-Yarrow et. al., 1983; Zahn-Waxler, Radke-Yarrow, & King, 1979, 1982, 1990). However, there is some evidence that the occasional use of power assertion techniques by nurturing parents who usually employ inductive techniques plays a positive role in moral internalization by, for instance, letting the child know that the parent feels strongly about something or by controlling a child’s defiant behavior (Zahn-Waxler et. al., 1979). Love withdrawal, on the other hand, contributes to the child’s inhibition of anger.

For example, in a naturalistic study, Zahn-Waxler and her colleagues (1979) studied the responses of trained mothers to both the prosocial behaviors of 1- and 2-year-olds and their failures to act in a properly prosocial fashion. In the latter situation, some mothers corrected their children by pointing out to them in an emotionally expressive fashion that they had harmed or hurt someone and that it was wrong to do so. Other mothers corrected their children by emphatically commanding them to stop their behavior. The former children were later deemed to be more prosocial and more concerned to make up for the wrongs they had done, while the latter showed fewer prosocial and reparative behaviors.

The general message of these studies is that love withdrawal and power assertion techniques may play a role in getting the child’s attention, while inductive techniques serve to point out to the child, for instance, the harmful consequences of her actions. These help to engage the child’s empathic capacities and can also lead to feelings of guilt. The most successful moral internalization occurs as the result of repeated use of inductive techniques in varied circumstances of moral learning.

Researches have identified four components of conscience that make for successful internalization (Kochanska & Thompson, 1997). These are cognitive awareness and understanding, emotions, self-control, and motivational capacities. They usually appear in a child’s second and third years. Cognitive awareness and understanding concern the toddler’s non-reflective knowledge of herself as a causal agent whose actions are subject to evaluation. They also concern the toddler’s abilities to represent and remember standards of conduct. It is speculated that these abilities include the capacity to form prototypical structures or scripts concerning action scenarios that incorporate the use of simple rules and behavioral standards and that enable the prediction of the outcomes of different actions in a given scenario. Emotions are directed toward oneself and others. They include, respectively, on the one hand, anxiety, shame, embarrassment, guilt, remorse, and pride and, on the other, upset (anger), interest, amusement (joy), and empathy. The emotions provide the resources for both motivation and evaluation. They tract environmental factors that indicate actions to be performed or avoided, and they also motivate their performance or avoidance. Self-control refers to the capacity of the toddler to avoid prohibited actions and perform socially desirable ones without surveillance or immediate external reinforcers. Motivational capacities refer to the toddler’s relatively enduring stance toward their caregiver’s values and standards as well as their openness to socialization. These are associated with a toddler’s temperament and the influences of parenting styles and approaches. The
development and acquisition of these capacities lead to internalization and thus to the successful performance of right moral actions by the toddler herself. A key element in the success of inductive techniques is the role the infant’s and child’s empathic and sympathetic abilities play. These enable the child to discern the observable indicators of harm and hurt and to be moved by them. The success of inductive techniques in bringing about moral internalization provides strong evidence for the presence of reliable cognitive and motivational mechanisms of moral learning in infants and young children, in particular the infant and child’s empathic and sympathetic capacities (Rottschaefer, 1999).

Managing Conduct Disorder and Reliable Mechanisms of Moral Learning

Research on conduct disorder lends further support for the important role of parenting and the development of perceptual, cognitive, and emotional skills for effective moral learning.

Developmental psychologists have examined a range of behaviors generally classified as aggressive and antisocial. These contrast with the helping and prosocial behaviors that are the object of moral internalization. Coie and Dodge (1998, p. 781) define aggressive behavior as behavior that is aimed at harming or injuring another person or persons, including the threat of force in inflicting loss or damage to property. Conduct disorder is a psychiatric term that is applied on the basis of the frequency of such problems over a period of time, for example, three or more over a period of six months.

Two effective types of treatment of conduct disorder point toward reliable mechanisms that are similar, on the one hand, to the inductive parenting techniques discussed earlier and, on the other hand, to the emotional and cognitive mechanisms employed by children who have successfully internalized the moral lessons of their parents. They are Problem-Solving Skills Training (PSST) and Parent Management Training (PMT) (Kazdin, 1993, 2003).

There is solid support for the claim that cognitive processes such as perceptions, statements to oneself, attributions, and problem solving skills play a major role in conduct problems (Kazdin, 1993). For instance, aggression is triggered not only by environmental events but also by how such events are perceived and cognitively appraised. These appraisals concern the situation, including other individuals, anticipated reactions of participants in the situation, and the agent’s views about herself in relation to the situation. Attribution of intent turns out to be critically important in understanding aggressive behavior. Aggressive young people have a predisposition to attribute hostile intent to others, especially in social situations in which the cues of actual intent are ambiguous. Perceptions of hostile intent are more likely to lead to hostile reactions than are perceptions of neutral or friendly intent. Aggressive children have also been shown to be lacking in some interpersonal problem solving skills, for instance the ability to generate alternative solutions to problems, to relate means to ends, and to effectively consider consequences. Although there is still much work to be done on
the precise nature of these cognitive deficits, enough is known that it can be used to develop ideas about treatment and treatment plans.

PSST aims at developing interpersonal problem solving skills. The many varying treatment plans have some common features. First, they emphasize the cognitive strategies that the children should use in approaching problematic situations, including importantly discerning expressions, gestures, and behaviors indicative of intent. Second, children are instructed on how to go in a step-by-step fashion to solve interpersonal problems. In doing so, they formulate for themselves key aspects of the problem and the ways they can address them. Third, the treatment involves structured tasks such as games, academic activities, and stories. As their problem-solving skills develop, children begin to apply them more and more to real-life situations. Fourth, therapists model the cognitive processes that the children are learning. Finally, treatment usually involves several modes of learning. Besides practice and modeling, the children role-play. They are also reinforced for their successes and mildly punished, for instance by the loss of tokens or points. Several outcome studies show impulsive, aggressive, and antisocial children and adolescents have reduced these problems significantly (Kazdin, 1993).

PMT provides instructions to parents for altering their child’s behavior at home. Parents learn the fundamentals of social learning theory and how to apply procedures based on that theory to increase prosocial and decrease antisocial behaviors in their child. The hypothesis is that parents have, in their discipline practices, inadvertently taught their child antisocial behaviors. They have negatively reinforced such behaviors by ceasing to insist on their discipline practices when the child increases his disruptive behavior. Thus, the child has learned that increasing his disruptive behavior will lead to parents giving in to him and, thereby, enabling the child to avoid the parent’s discipline. In turn, children negatively reinforce the parents, when upon achieving what they wanted, they cease their disruptive behavior. Parents learn that they can avoid the child’s disruptive behaviors by relenting to him.

Common features of PMT include instruction of parents without direct intervention by the therapist with the child. Parents are taught to specify carefully the problem behaviors. They also learn the principles of social learning theory and the procedures following from it. These include the use of positive reinforcers, such as the use of social praise and tokens, to promote prosocial behavior, and of mild punishment such as time out from reinforcement or loss of privileges. Parents observe how the techniques are put to work, practice the techniques, and review their own successes and failures in their use. The parents are taught to proceed in a step-by-step fashion, moving from less severe to more severe problems and extending their focus to other situations in which the problematic behaviors manifest themselves.

Researchers have studied PMT extensively with children of varying ages that have conduct problems with different degrees of severity. Several controlled studies have shown that children disciplined by parents using PMT have improved by decreasing antisocial behaviors and increasing prosocial behaviors (Kazdin,
This improvement is evidenced after one year and even 10 years later. PMT also has other positive consequences such as the reduction of deviant behaviors in siblings and a reduction of stress and depression in parents.

Kazdin (2003) has extended both lines of research, testing the results of using PMT and PSST alone and in combination. He has found in a controlled study that the combination treatment leads to more improvement than either treatment alone.

Detailed analyses of cognitive and motivational mechanisms invoked by PSST and PMT, as well as of the features of the natural and social environments to which they are attuned, are still required. Nevertheless, the findings are sufficient to lend further support to the moral learning hypothesis. In particular, one of the skills acquired by children is the ability to discern the observable indicators of the emotions, intentions, and other psychological states of those with whom they are interacting and to respond to these indicators in morally effective ways.

**Some Lessons for Studying Successful Moral Agency and Its Achievement through Moral Learning**

The power and significance of the types of research programs, empirical findings, and confirmed hypotheses that I have discussed lie in several features. First, the findings are supported by controlled studies in real-life situations. The ecological validity of these studies contrasts favorably with the data gathered by second philosophical, experimental philosophers who explore issues in normative and metaethics by means of surveys, paper and pencil tests, computer-assisted studies of hypothetical situations and dilemmas, and experimental games. Second, the studies focus on moral behaviors in contrast with a continuing attention to folk moral concepts, hypothetical moral judgments, and artificial moral dilemmas. Thirdly, they focus on the means by which successful moral agency is established in contrast to less-successful means. I mention these contrasts not to denigrate the findings frequently used by the new experimental philosophers, but only to argue that these findings need to be supplemented if an adequate account of successful moral learning and successful moral agency is to be achieved.

To highlight some of the features of the work that I have been describing, let me present in schematic form some lessons they provide for the investigation of moral learning and successful moral agency. These suggestions may be helpful not only to students of metaethics, especially moral psychology, who are already persuaded of the importance of scientific findings and theories for their work, but also to first philosophers who admit that such results are significant, though in a limited fashion, for their endeavors. I focus on aspects of the studies discussed and others of a similar sort that, I contend, would enable progress on understanding moral learning and would do so without bringing in a number of important, but more controversial, metaethical issues that are of concern to both first and second philosophers—issues such as moral realism and anti-realism that arise in metaethical discussions of moral learning. I suggest, though I do not argue for it here, that these and other empirical findings and theories concerning moral learning will be important in settling these metaethical controversial issues because
they offer a commonly shared understanding of the moral phenomena—something that is often lacking in these discussions. These suggestions offer ways of achieving some consensus about the moral phenomena (successful and unsuccessful moral actions) concerning which metaethicians can test their semantic, ontological, metaphysical, and epistemic theories.

- Consider, as cases of successful moral behavior, types of behaviors that are thought to be prototypically moral.
- Use functional and substantive criteria for what counts as moral that are derived from common-sense markers of the moral and from commonalities agreed upon by empirical researchers from different research traditions.
- Examine studies that focus on causal factors that are under the control of the agent in the sense that they are not primarily biological, whether genetic or developmental, neuropsychological, or constitutive of the social environment.
- Seek a level of description of the behavior that is, as far as possible, neutral to competing standard normative ethical theories such as consequentialism, deontology, or virtue theory.
- To the extent possible, fix the description of the target behavior, both substantively and functionally, at a level that is neither that of a particular case nor so general that it embraces several different moral values or functional patterns, or becomes vacuous.
- Make use of psychological studies that investigate how and why the target behavior is or is not successfully accomplished.
- Ideally, employ controlled studies that compare competing hypotheses along with a control group.
- Ideally, rely on studies that make use of the best sorts of scientific standards.
- To the extent possible, prescind from metaethical assumptions about the ontological or metaphysical status of moral phenomena and from epistemic assumptions about the nature of the grasp of moral phenomena, whether cognitive or non-cognitive.

I contend that the studies that I have examined concerning moral internalization and aggression management meet these criteria and that a focus on studies that do so will provide second philosophers, as well as interested first philosophers, with a wealth of substantially agreed upon results. These results will offer an empirically grounded basis for studying successful moral agency in more detail and for addressing the more controversial metaethical questions about the semantic, metaphysical, ontological, and epistemic characteristics of such agency. In turn, they will enable metaethicians to assess the relative epistemic quality of their theories.
Repliyng to Some Objections

Let us consider a range of objections and problems that, no doubt, will be raised by metaethicicans in the first philosophy tradition to my proposal.²

First philosophers often find that appeals to empirical findings are interesting but either fundamentally irrelevant to their concerns or, at best, indicative of minimal sorts of constraints on an enterprise that proceeds by means of armchair methods. I shall address some of these reactions in a series of replies to objections often raised by first philosophers to the sort of empirically based second philosophical approach that I have been advocating. My aim is to nudge the dialogue further along toward a second philosophical perspective.

Reasons cannot be causes. We understand people to be rational agents, agents who act on reasons. Reasons provide an understanding of actions—an account of why they occur. They do not furnish causes for actions—an account of how they occur. Thus, we should take explanations in terms of reasons (rationalizations) to be the use of an interpretive framework that offers an understanding of action, not a causal explanation of it.

Reply: Some reasons are causes insofar as they make a difference in what actions occur. A causal reason is a belief/desire complex, the knowledge of which enables predictions and explanations of a person’s actions and an understanding of a person’s action in terms of her mental states (Davidson, 2001). Actions are contingently connected to belief/desire complexes. Thus, rather than being understood as the application of an implicit analytic definition of a rational agent, reason accounts of human action are better understood as claims about what sorts of belief/desire complexes produce what sorts of action in specified circumstances. This account of action explanations in terms of reasons as belief/desire complexes fits both common sense accounts of reason explanations as well as scientific psychological accounts.

Moreover, they are superior to armchair philosophical accounts, for such armchair accounts, if meant to be analyses of linguistic usage, are irrelevant to descriptions of the phenomena under discussion. If, however, they are intended to capture common sense concepts, they face the problem that the intuitions and reflections used to support them need to be shown to be commonly shared and trustworthy. But recent empirical work concerning such intuitional methods indicates that there is at least a prima facie case against the claim that the armchair techniques are reliable or that they capture genuine commonalities (Alexander, Mallon, & Weinberg, 2009; Nagel, 2007; Alexander & Weinberg, 2007). Moreover, even if it were shown that such intuitions are reliable and capture genuine commonalities about our conceptions, it is a further matter to establish that the hypothesized common concept of rationality actually applies to the

² Stephen Finlay (2007) provides a very helpful survey of current major metaethical positions set in the context of moral realism and anti-realism.
phenomenon of human action. On the other hand, ordinary knowledge and scientific findings understand an agent’s reasons to be causes of her action in so far as they are difference—making factors in the coming about of an action. Like causes, reasons (understood as belief/desire complexes) can sometimes make a difference in what people actually do. Thus, there is solid support for thinking that reasons do serve as causes of human behavior rather than merely interpretive devices for promoting understanding.

**Causes cannot be normative.** But even if one grants that reasons as belief/desire complexes can serve as causes, such complexes lack the normative component that is an essential part of reason explanations. Thus, such causal reasons are not sufficient for the kind of reason explanations that we are seeking in moral psychology. To see this, assume a reason explanation takes the form:

If agent A wants item I and believes that means M will bring about I, then, *ceteris paribus*, A *ought* to do M.

On the other hand, causal explanations have the following form:

If agent A wants item I and believes that means M will bring about I, then, *ceteris paribus*, A *will do* M.

Thus causal reason explanations may have predictive or explanatory power, but they lack normative power.

**Reply:** It should be granted that reason explanations do have normative power, but it need not be conceded that explanations in terms of reasons understood as belief/desire complexes lack such power. In the first place, desire can serve a normative function, being the indicator of either instrumental or intrinsic ends. And second order desires can serve to normatively order first order desires, if necessary. Alternatively, one can add an empirically based hypothesis that people are rational agents (Smith, 1998, 2000). Doing that we can get the following sort of normative analysis:

If A is a rational agent, then if A wants item I and believes that means M will bring about I, then, *ceteris paribus*, A *ought* to do M.

And if A acts rationally, then A *will do* M.

Consequently, reasons explanations can have a normative component, but, nevertheless, are causal.

**Causes cannot be justifiers.** But rational causal reasons (rationalizations) can provide justifications only if the right sort of desire or right sort of reason is operative. Otherwise, they provide at best normatively aberrant causal
explanations, where the causal factor is a belief/desire complex. There is no guarantee that an appropriate rationale will be among the rational causal reasons that account for an action, predict it, or give it normative force. The right sort of normative force is required.

**Reply:** Some rationalizations can also be justificatory since some provide an appropriate rationale for an action. The norm invoked by the action may be the right one for the situation and the action.

**Rationalizations cannot be moral justifiers.** Rationalizations may provide some appropriate rationales for some actions, but they cannot do so for *moral* actions. For instance, they may provide appropriate rationales for following the rules of etiquette, the rules of a game, or that of an institution. But the first two norms are irrelevant to moral matters and the last may involve the rules of an immoral institution, for instance slavery.

**Reply:** Sometimes the rationale provided by rationalizations is appropriately moral if it is related to a moral action in the right way. For instance, in helping someone in need an agent may respond to a spontaneous desire to do so or act on a second order desire to do her duty or to increase the welfare of the person in need. When such moral norms are involved in rationalizations, rationalizations can serve as moral justifiers (Kennett & Fine, 2009; Woodward & Allman, 2007).

**The rationalizations cited in the cases of moral internalization and anger management cannot serve as moral justifiers.** Even if it is granted that in some cases rationalizations can provide appropriate moral rationales for actions, the cases of moral internalization and anger management are not illustrative of such cases. In the cases of moral internalization and anger management we do not have the right sort of relationship between the rationale and the action because it has not been established that the action is morally right. Thus, the empirical findings that you use to make your case are irrelevant.

**Reply:** It is highly plausible that anyone examining the actions resulting from successful moral internalization and anger management would agree that the actions in question are at least *prima facie* morally right. I expect that even for the first philosopher they appear intuitively—and even upon reflection—the *prima facie* right thing to do. Moreover, many—if not all—normative moral theories (Utilitarian, Kantian, Natural Law, and Virtue Theory) would deem them so, at least in a *prima facie* fashion.

**Agreement is not sufficient to establish moral rightness.** It is clear that agreement is never a sufficient condition for establishing the truth of a claim. Even massive consensus is fallible. The actions in question may not be morally right, and it is clear that they have not been shown to be morally correct.
Reply: It is true that consensus is not a sufficient criterion of truth or justification, but the issue here does not concern the ultimate moral justification of the types of action in question. The issue is whether the mechanisms of moral internalization and the strategies of anger management provide satisfactory explanations and justifications of *prima facie* successful (that is, correct) moral actions and beliefs. Given the adequacy of the scientific findings, the explanations are satisfactory and justificatory, for they both exhibit the causal factors that bring about the moral actions in the right sort of way and provide the right sort of rationale for the action.

**A case of the genetic fallacy.** But one cannot provide the right sort of justification if the explanatory factors are only about the causal origin of the action since issues of origin and justification are distinct. And that is the case with respect to the two sorts of moral learning we are considering. Thus the genetic fallacy has been committed: an appeal to causal origins has been substituted for a justification.

Reply: Sometimes causal sources also provide justification, for instance, perceptual processes provide justification for perceptual beliefs. In the two cases that we have examined we have learning processes that enable the development of capacities whose activities generally lead to moral ends. They are facilitative of morally correct actions and they demonstrate that children can attain moral capacities to achieve certain moral goals that they ought to bring about. They show that the child is making use of reliable means for attaining a moral end.

**The rationalizations are insufficiently reflective.** Even granting that the means employed are reliable, moral actions require something more—sufficient reflection and deliberation. The inductive capacity employed by the child is entirely unreflective, and the anger management techniques employed by the children and young people are not sufficiently reflective, especially those of PMT. But sufficient reflection is a necessary condition for justification.

Reply: In both cases we can say that the child is at least *prima facie* morally justified in her action, though she may not be able to show that she is so justified. Indeed, any one of the major competing normative theories could be used to show that the inductively tutored child’s action or the cognitively behaviorally trained child’s action is at least *prima facie* morally right. Reflective processes may be necessary for *showing justification*, but they are not necessary for *being justified*.

**Genuine justification requires reflective processes.** Achieving genuine justification requires the ability to show that one is justified and that requires the capacity for reflective activity, the ability to consider one’s reasons for acting and explaining why they are appropriate ones. But neither the process of moral internalization nor those involved in anger management make use of this sort of reflective ability.
Reply: Requiring reflective ability for justification is too stringent a condition of justification. It makes the possession of reflective capacities a necessary condition for moral evaluation. But moral practice does not set up such a requirement for moral judgment or assessment. Very young children are introduced to moral practice well before they are capable of moral reflection, and adults are assessed morally even though it is not clear that many are capable of such reflective practices. It seems better and more in accord with moral practice to view moral assessment on a continuum. Reflection may be an ideal, but it does not seem to be a necessary condition for the claim that someone has a justified moral belief or has acted in a morally justified fashion. Indeed, there is evidence that in complex situations, moral intuitions based in primary processes provide results superior to those achieved by reflective and deliberative processes, so-called secondary processes (Woodward & Allman, 2008). This is so because, among other reasons, the latter processes are able to handle only a limited number of items in reflection, but complex moral situations often require attunement to multiple morally relevant factors.

Adult moral agency requires reflective capacities. Adults need to be able to assess what they have done or intend to do in the light of moral principles and thereby ascertain whether they have justification and be able in the proper circumstances to show that they are justified. Thus, the findings you examine are irrelevant to adult moral assessment. They deal only with children. Children are only potential moral agents.

Reply: Clearly there are differences between children and adults in terms of both capacities and requirements, but if the findings that we have discussed are correct, then adults will need to be able to appeal to these or other learned reliable mechanisms of moral knowledge, motivation and action in reflectively justifying their actions. For it seems implausible that the adult will have to appeal to entirely new reliable mechanisms of moral knowledge, motivation, and action—ones unconnected with their childhood learning—in order to show justification. Moral practice reveals continuity in the understanding of moral assessment that favors the less demanding view of justification, that is, the view that does not require for genuine justification the ability for reflective justification. Reliable mechanisms of moral belief formation and moral action are sufficient for the right sort of psychological states and processes that lead to successful moral action, and that is sufficient for justification. Reflective capacity may improve justification, but does not constitute justification.

Moreover, there is an important ambiguity in the notion of reflection that should be cleared up. First philosophers often think of reflection in terms of a meta-level cognitive process, like introspection, that makes cognitively available to them the types of first-level processes involved in judgment, belief formation, or action. Reflection then is the process by which these internally accessible first-level processes are assessed for their epistemic and moral adequacy. Call this internalist reflection. However, there is another understanding of reflection that
allows for the fact that many first-level, reliable cognitive processes may not be introspectively available to meta-level. Indeed, there is much evidence that this is the case (Haidt, 2001; Woodward & Allman, 2007). Nevertheless, knowledge of the nature and reliability of these processes can be acquired through scientific investigation. The moral agent can then make use of this sort of knowledge to assess the reliability of the means used to form moral judgments and bring about moral actions. Call this externalist reflection. The second philosopher, then, can endorse the notion that reflective justification often enhances knowledge about the bases of one’s justification, thus enabling one to show that and why one is justified, without, however, conceding that use of such reflection is in some way a concession to the practices of first philosophy.

**Appeals to reliable mechanisms do not solve any of the problems of metaethics.** But an appeal to the mechanisms that you have discussed is still unsatisfactory because it does not answer the fundamental problems of metaethics, those involving semantic, ontological, metaphysical, and epistemic issues. For instance, it does not answer moral nihilists and skeptics, who deny that there are any moral values at all or who deny that we are able to know that there are such values, even if they exist.

**Reply:** This is true, but one should not expect these findings to address directly the issues raised by moral skeptics and nihilists (Sinnott-Armstrong, 2006). For the findings assume that there are some moral values and actions that are morally right or wrong. After all, the *explananda* are taken to be instances of genuine morally right actions. Moral skeptics and nihilists deny that there are any such actions. Saying that, however, is not to say that those findings might not play some role in an adequate answer to skeptic or nihilist. The *explananda* in question are considered by all parties other than the moral nihilists and skeptics to be moral *explananda*. While one might object that this assumption begs the question against moral nihilists or skeptics, one might argue, on the other hand, that the agreement about the phenomena renders the skeptic and nihilist position problematic. The burden of proof seems to be on the moral skeptic and nihilist to show why these empirically established findings are to be rejected as off track—but further discussion of this matter will have to wait for another day.

**The metaethical issues between moral realists and anti-realists or between cognitivists and non-cognitivists all remain unsettled.** Even if I grant you that one ought not require that the findings you have presented solve fundamental epistemic issues concerning moral skepticism and nihilism all on their own, you have to grant that they also leave semantic, ontological, and metaphysical and epistemic issues unsettled. For it seems clear that the findings that you have presented, even if one grants that they provide justifying explanatory reasons for the moral actions in question, do not answer these larger metaethical questions. Even if one grants that the empirical findings that you have made use of have relevance for issues concerning moral psychology and functional moral agency,
topics addressed in metaethics, they tell us nothing about other major issues in metaethics.

Reply: The short answer is that your assessment is correct, but that does not imply that the findings I have discussed, as well as other well-established results, do not have important implications for these issues. Recall that I have distinguished between functional and substantive metaethical questions. I take it that the findings lead us to some functional metaethical conclusions, namely, that the skills learned through inductive parental moral instruction and cognitive behavioral techniques endow moral agents with reliable means of performing moral actions and that as reliable, they also render the agent justified in her actions. Though there is not space here to go into these metaethical issues, let me conclude by making a few comments about the implications of the findings about functional morality for them.

Semantic issues have to do with the question of whether moral claims can be true or false, and if so, in what sense of the term. Ontological issues have to do with the truth makers of moral claims, if such claims can be true. Metaphysical issues have to do with the nature of those truth makers. My claim is that examination of successful moral agency, as illustrated in the two cases discussed, will be able to provide evidence for or against cognitivist and non-cognitivist theories. And given, for instance, that the evidence favors cognitive theories, such findings will aid in sorting out what sort of factors make moral beliefs true and what the nature of these truth makers are. Indeed, given that a phenomenon of major interest has been pinned down—successful and unsuccessful moral behaviors in various sorts of situations—one would expect progress on these metaethical issues. Contrast this with attempts to make progress on these issues by means of armchair methods. One problem mentioned already is that the commonality and trustworthiness of these methods is questionable. Another is that even if proponents of differing metaethical views agree that their methods produce a result that can be taken as a genuine explanandum, for instance, the results of thought experiments about pleasure machines or the assumption that ordinary people believe that moral values are objective, the ecological validity of these explananda is often left completely open. On the other hand, my suggestion is that these problems are assuaged if metaethicians rely on reliable empirical findings, results, and theories. But this suggestion remains, as Wilfrid Sellars would say, a promissory note to be paid off in the future.

Some Challenges

Lastly, I shall note some major challenges that remain for those metaethicians, whether of the first or second philosophy persuasion, who find what I have been suggesting initially plausible enough that they would seriously consider pursuing the above roughly outlined research project.
Challenge I: Ecological Validity
Even though some of the studies referred to above have demonstrated successful moral learning in actual situations, it would be good to have even greater evidence of ecological validity and in many more areas of moral action.

Challenge II: Cross-Cultural Validity
Work needs to be done to test the cross-cultural validity of claims about successful moral learning. For instance, there do seem to be some cross-cultural differences in effective parenting techniques. In addition the work of Haidt and Bjorklund (2008) and others has shown that the substantive realm of the moral has sometimes been too narrowly conceived.

Challenge III: Identifying Reliable Processes
A lot of effort needs to be extended in distinguishing, identifying, and delineating the various factors and processes that go into the performance of a correct moral action and distinguishing the factors and processes key to well functioning. In both cases that I have discussed, especially in the instances of PMT and PSST, some progress has been made. But in both of them, and in other areas where reliable methods of moral learning seem to have been found, much still needs to be done.

Challenge IV: Rapprochement
Though second philosophers may be uneasy about the *a priori* methods of first philosophers, much work in metaethics done by first philosophers is, I believe, extremely rich and suggestive in its theoretical exploration of the nature of moral agency. This work offers possibly fruitful hypotheses for second philosophical exploration. For example, David Copp’s (2007) recent work on the psychological basis of the normativity of self-grounding reason echoes some of the empirical and theoretical work of social cognitive theorists on the role of notion of the self in an account of successful agency (Bandura, 1986, 1997). On the other hand, though first philosophers contend that normative issues are necessarily distinct from empirical ones, empirical work, like that concerning moral internalization and aggression management, can provide first philosophers with well-grounded findings on human moral functioning. They might be able to use these findings to supplement their conceptual analyses, thought experiments, and intuitions about how agents act morally as they construct their non-empirically grounded normative and metaethical theories.

Conclusion
I conclude that the moral learning hypothesis has solid scientific support and that it provides an important way to understand and explain successful moral agency. Moreover, successful pursuit of the moral learning hypothesis is a significant addition to the growing efforts to provide a scientific basis for this central issue in metaethics. As such, it stands as a good example of efforts in a
philosophical research program aimed at transforming metaethics from a first to a second philosophy enterprise.

References

MORAL AGENCY AND MORAL LEARNING


