
Mindreading and the Core Architecture of Moral Psychology

Shaun Nichols
Department of Philosophy
College of Charleston
Charleston, SC 29424
and
Center for Cognitive Science
Rutgers University
Piscataway, NJ 08854
nichols@cofc.edu

Contents

---

1. Introduction

Throughout this century, philosophers and psychologists have tried to explain features of our moral psychology by appealing to features of our capacity for understanding other minds or "mindreading". Perhaps the most widely known treatment goes back to Piaget's early work in developmental moral psychology (Piaget 1932). Piaget and his followers placed enormous weight on the ability for perspective taking, of imagining oneself to have the mental states of another (e.g., Kohlberg 1984, Selman 1980, Damon 1977; see also Rawls 1971, chapter 8). Over the last two decades, there has been considerable empirical and conceptual progress in work on moral psychology and in work on mindreading. The moral psychology tradition has looked at the nature and development of two basic moral capacities: the capacity for altruistic motivation (e.g. Batson 1991, Blum 1994, Eisenberg 1992, Hoffman 1991, Sober & Wilson 1998, Zahn-Waxler & Radke-Yarrow 1982), and the capacity for moral judgement (e.g., Blair 1995, Goldman 1993, Nucci 1986, Smetana & Braeges 1990, Turiel et al. 1987). The mindreading tradition has explored the capacity for attributing mental states to others and predicting others' behavior (e.g., Baron-Cohen et al. 1985, Bartsch & Wellman 1995, Currie & Ravenscroft forthcoming, Goldman 1989, Gopnik & Wellman 1994, Gordon 1986, Harris 1992, Leslie 1994, Nichols & Stich forthcoming, Stich & Nichols 1992). Although each tradition has flourished, work on moral psychology and work on mindreading has been pursued largely independently. Advances in both fields put us in an excellent position to begin charting the relations between these two capacities and to develop

a more detailed picture of the core architecture of moral psychology.

Within the last decade, several philosophers and cognitive psychologists have begun to suggest cognitive accounts of altruism and moral judgement (e.g., Batson 1991, Blair 1995, Blum 1994, Currie 1995, Darwall 1998, Deigh 1995, Goldman 1993, Gordon 1995, Sober & Wilson 1998). The role of mindreading is a central issue in all these accounts. Roughly, the proposals about mindreading and moral psychology fall into two camps. Some (e.g., Blair 1995, Sober & Wilson 1998) maintain that basic capacities of moral psychology do not require any mindreading ability at all. Others (e.g., Batson 1991, Goldman 1993) maintain that basic capacities of moral psychology depend on the capacity for perspective taking.

In this paper, I'll consider the recent cognitive accounts of altruistic motivation and moral judgement. In section two, I'll review the recent work on altruistic motivation. I'll argue that, contrary to the prevailing views, altruistic motivation depends on a minimal capacity for mindreading and also on an affective system, a "Concern Mechanism" that generates the motivation. The third section will focus on recent cognitive accounts of the capacity for moral judgement. I'll argue that the capacity for moral judgement, like the capacity for altruistic motivation, depends on a minimal capacity for mindreading. Furthermore, recent evidence suggests that moral judgement also depends on the affective mechanism underlying altruistic motivation. Thus, I'll maintain that at the core of moral psychology is a Concern Mechanism, which is crucial both to altruistic motivation and to the capacity for basic moral judgement.

2. Cognitive mechanisms in altruistic motivation

The literature on altruism is simply enormous, and it spans several disciplines including philosophy, social psychology, developmental psychology, and evolutionary biology. Although I'll draw on work from all of these areas, the goal of the present section is restricted to the project of determining the cognitive mechanisms underlying basic altruistic motivation. I'll consider in some detail recent proposals about the mechanisms underlying altruistic motivation. I'll argue against the radical view that mindreading capacities are unnecessary for altruistic motivation. Then I'll sketch the more prevalent proposal, that altruistic motivation depends on the capacity for perspective taking. I'll maintain that none of the arguments for the perspective taking account is convincing and that there is considerable evidence that altruistic motivation does not depend on such sophisticated mindreading capacities. Rather, I'll suggest that altruistic motivation depends on a Concern Mechanism that requires only minimal mindreading capacities, e.g., the capacity to attribute distress to another. But before we get to that, I need to get clearer about the operative notion of altruistic motivation.

2.1. Core cases of altruistic motivation

To begin a discussion of altruism, it will be useful to set out some core cases of altruistic behavior. In science in general, it's not always clear at the outset what the core cases are, and new evidence and arguments might alter our conception of what should be included as core cases. The situation is no different in studying altruism, and we may want to revise our view about what the core cases are. Philosophical discussions in this area tend to rely on hypothetical cases of altruism. But since the present goal is to give an account of the cognitive mechanisms implicated in actual cases of altruistic behavior, it is important to begin with real cases.[1] To his credit, philosopher Lawrence Blum takes this strategy and offers real examples of helping behaviors that he suggests need to be accommodated by an adequate theory of altruism (Blum 1994). Blum's cases all come from young children. For present purposes, it

will suffice to recount just a few of the examples:

1. Sarah at 12 months retrieves a cup for a crying friend (Blum 1994, p. 186).

2. Michael at 15 months brings his teddy bear and security blanket to a crying friend (Blum 1994, p. 187).

3. A two-year old accidentally harms his friend (another two-year old) who begins to cry. The first child looks concerned and offers the other child a toy (Blum 1994, p. 187).

The clearest real-life examples of altruistic behavior in adults come from work on helping behavior in social psychology. Perhaps the best known research on adults' helping behavior is the work on the 'bystander effect' by Latane & Darley (1968). They found that when there are numerous bystanders, subjects are relatively unlikely to offer assistance to those in need. This finding is often used to draw a rather bleak picture of human altruism. However, focusing on these studies obscures the pervasiveness of human altruism. For it turns out that if subjects perceive unambiguous distress cues and there are no bystanders, virtually *everyone* helps. For instance, in one study, Clark & Wood (1974) had each subject engage in a distracter task and as the subject left the experiment, he passed a room in which a man (the experimenter's accomplice) made a sharp cry of pain and then feigned unconsciousness apparently as a result of being shocked by an electronic probe. The researchers found that when the accomplice was no longer touching any of the electronic equipment, *all* of the subjects offered help. And even when the accomplice was still touching electronic equipment (thus presenting potential danger to the helper), over 90% of the subjects offered help (Clark & Wood 1974, p. 282). An adequate account of altruistic motivation should explain these kinds of helping behaviors.

This list of core cases is, admittedly, rather short. It excludes possible cases of altruistic motivation that do not involve helping others in need. Sometimes people are generous to strangers who aren't in need, and I don't mean to suggest that such behaviors can't be altruistic. However, I think that by focusing on a more limited range of cases, we are more likely to make progress on cognitive theories of altruism. The cases of comforting or helping others in distress form a plausible core because such cases emerge so early in children and they appear to be pervasive among adults. Furthermore, although I'm focusing on a very short list of core cases, these cases already present a fairly daunting task. Devising an account of altruistic motivation that would capture both the child cases and the adult cases would be a considerable advance. Of course, it's possible that the examples from children and the examples from adults cannot be captured by a single motivational system. But all else being equal, an account of altruistic behavior that can capture both of these cases would be preferable to an account that captures only one. I'll argue that a close look at the role of mindreading in these cases will provide us with a unified account.

2.2. Altruism without mindreading?

One account of the role of mindreading in altruistic behavior is to deny that mindreading plays *any* essential role in altruistic motivation. There are two versions of the view that are discussed in the recent literature. In this section, I'll argue that neither account is at all plausible.

Emotional contagion

Perhaps the most common explanation of the basis for altruistic behavior is empathy. For instance, Goldman writes, "empathy . . . seems to be a prime mechanism that disposes us toward altruistic behavior" (1993, p. 358). However, it is important to distinguish between two different capacities that get labeled as "empathy". Most generally, empathy is regarded as a "vicarious sharing of affect" or an emotional response in which the emotion is "congruent with the other's emotional state or situation" (Eisenberg & Strayer 1987, pp. 3,5). There are two rather different ways that one might arrive at a "vicarious sharing of affect". One way is by perspective taking, i.e., imagining oneself to have the other person's mental states. I will consider an empathy-based account of altruism along these lines in section 2.3. A quite different way that we arrive at the same affect is by emotional contagion, when we "catch" another's affect. Some capacity for emotional contagion is present at birth as shown by the fact that infants will cry when they hear the cries of another infant (Simner 1971). The capacity for emotional contagion thus does not require the capacity for perspective taking. Indeed, since the capacity for emotional contagion is present at birth, this capacity is presumably completely independent of mindreading capacities. There is some dispute about when mindreading capacities become available, but all sides agree that newborn babies cannot engage in mindreading.

The capacity for emotional contagion suggests a natural and simple account of altruistic motivation. If the distress of another causes oneself to feel distress, this may provide a motivation to relieve the distress of the other – it will thereby relieve one's own distress. This view has a certain elegance, but it is not easy to find a prominent advocate for the view. Although Goldman maintains that altruistic behavior is generated by empathy, Goldman also maintains that emotional contagion is not genuine empathy (1993). Indeed Goldman's simulation account of empathy is quite implausible as an account of emotional contagion (see Nichols et al. 1996), so it's unlikely that Goldman thinks that altruism derives from emotional contagion. Martin Hoffman, one of the most influential figures in empathy research, has been read as proposing something like the simple emotional contagion view in the following passage: "Empathic distress is unpleasant and helping the victim is usually the best way to get rid of the source. One can also accomplish this by directing one's attention elsewhere and avoiding the expressive and situational cues from the victim" (Hoffman 1981, p. 52, quoted in Batson 1991, p. 48). It's not clear that Hoffman is really committed to this view, but it is instructive to consider the account in any case.

Notice that on the emotional contagion account of altruistic motivation, mindreading isn't essential to altruistic motivation. As noted above, emotional contagion needn't implicate mindreading processes at all. The distress cues are like bad music that you try to turn off. It requires no knowledge of electronics to be motivated to figure out how to stop the offensive stimuli coming from a stereo – one simply experiments with the various knobs and switches. Failing that, one can just leave the room. Similarly, then, one might find the cries of an infant offensive, so one might try to figure out how to stop the stimuli. To be sure, mindreading can provide useful tools for stopping the unpleasant stimuli. But on this account, mindreading needn't be essential to the motivation to stop the crying.

This story has a prima facie virtue – we know that this capacity is well within the abilities of young children who provide some of our core cases of altruistic motivation. So, the emotional contagion account provides an extremely simple explanation of altruistic motivation, and we know that children have the capacity for emotional contagion. Hence, it would seem that our problem is solved. Altruistic motivation doesn't depend on mindreading at all. Rather, it depends on the rather primitive capacity for emotional contagion.

Things are not so simple, however. For consider that, at least in the core cases of altruism from adults, one way to rid oneself of the unpleasant cues is to *leave the situation*. But this is not what

happened in the core cases noted above. Although the subjects could have eliminated contagious distress by fleeing the situation, almost none of them did so (Clark & Word 1974). The fact that adults often help when they could perfectly well escape has now been extensively explored in the work of C. Daniel Batson and his colleagues (Batson et al. 1981; Batson et al. 1983; Batson 1990, 1991). This research provides powerful evidence that some core cases of altruistic motivation cannot be accommodated by the simple emotional contagion account.[2]

Batson has the broader agenda of defending a perspective-taking account of altruism, which we will consider in section 2.3, but for present purposes, it will suffice to see how his data undermine the emotional contagion account. In classic social psychological fashion, Batson and his colleagues set up a mock shock methodology. Subjects were told that they would be in a study with another person and that one of them would be picked at random to be the worker and the other would be the observer. The worker would perform tasks while being given electric shock at irregular intervals, and the observer would watch the person performing the task under these aversive conditions. Of course, the real subjects always ended up in the observer condition, and the "worker" was really a confederate. The subjects were then told that they would view the "worker" via closed-circuit television (though it was really a videotape). The experiment manipulated the *ease of escape* for the subjects. In the easy-escape condition, subjects read "Although the worker will be completing between two and ten trials, it will be necessary for you to observe only the first two"; in the difficult escape condition, subjects read "The worker will be completing between two and ten trials, all of which you will observe" (Batson 1991, p. 114). The subjects subsequently viewed the worker endure two trials (of the ten trials that the worker had agreed to) in which the worker exhibited considerable discomfort. Subjects were given the opportunity to help out the worker by taking over some of her trials. Using this framework, Batson and colleagues also manipulated the degree of "empathy" in the subjects (see section 2.5 for details). Across a wide range of studies, they found that subjects in low empathy conditions were much less likely to help when escape was easy. By contrast, subjects in the high empathy condition were equally likely to help whether it was easy to escape or not.[3]

For our purposes, the crucial point is the following. On the emotional contagion model, one should only help when it's easier to help than it is to escape. However, evidence from Batson and his colleagues suggests that there is an important kind of altruistic motivation that can't be satisfied by escaping the situation. Hence, this kind of motivation can't be captured by the emotional contagion model (see also Batson et al 1981; Batson et al. 1983; Miller, Eisenberg, Fabes, & Shell 1996, Eisenberg & Fabes 1990). More generally, largely as a result of Batson's work, it is now clear that an adequate account of altruistic motivation needs to accommodate the fact that in core cases of altruism, people often prefer to help even when it's easy to escape.[4]

Sober & Wilson on altruistic 'sympathy'

In Sober & Wilson's recent book (1998), they propose an alternative path to altruism that doesn't rely on mindreading or emotional contagion, but rather on a certain kind of sympathy. They suggest that both sympathy and empathy may motivate altruistic behavior (e.g., 1998, p. 232). They then try to distinguish sympathy from empathy in two ways.

First, Sober & Wilson maintain that there is a crucial difference between empathy and sympathy because in sympathy,

> your heart can go out to someone without your experiencing anything like a similar emotion. This is clearest when people react to the situations of individuals who are not experiencing emotions at

> all. Suppose Walter discovers that Wendy is being deceived by her sexually promiscuous husband. Walter may sympathize with Wendy, but this is not because Wendy feels hurt and betrayed. Wendy feels nothing of the kind, because she is not aware of her husband's behavior. It might be replied that Walter's sympathy is based on his imaginative rehearsal of how Wendy would feel if she were to discover her husband's infidelity. Perhaps so – but the fact remains that Walter and Wendy do not feel the same (or similar) emotions. Walter sympathizes; he does not empathize (1998, pp. 234-5).

But of course, this example does not really distinguish sympathy from empathy. As Sober & Wilson seem to anticipate, a sophisticated empathy account can easily accommodate their case by claiming that we use our imagination to empathize with what Wendy would feel if she were to discover the infidelity. Hence, as far as this example is concerned, 'sympathy' is merely a special form of empathy.

The second, and more important, feature of their account is their claim that 'sympathy' doesn't require mindreading. Sober & Wilson maintain that empathy requires that one be a psychologist, but that sympathy does not:

> Empathy entails a belief about the emotions experienced by another person. Empathic individuals are "psychologists"…; they have beliefs about the mental states of others. Sympathy does not require this. You can sympathize with someone just by being moved by their objective situation; you need not consider their subjective state. Sympathetic individuals have minds, of course; but it is not part of our definition that sympathetic individuals must be psychologists (1998, p. 236).

Thus, Sober & Wilson apparently maintain that 'sympathy' does not require any capacity for mindreading.

Taken as an empirical claim, there is no reason to believe that Sober & Wilson's 'sympathy' exists. On the contrary, children only begin to exhibit characteristic signs of sympathy after the first birthday (see section 2.7) and at this age, children probably have some rudimentary mindreading skills (see, e.g., Gergely et al. 1995; Woodward 1998). So, it may well turn out that the capacity for sympathy exists only in creatures that have mindreading capacities and that the capacity for sympathy depends crucially on the capacity for mindreading. Furthermore, even if Sober & Wilson's 'sympathy' exists, they provide no reason to think that it explains anything like the core cases of altruism with which we began. Indeed, as we'll see, children only begin exhibiting comforting behaviors after the first birthday, by which time they probably have some rudimentary mindreading skills.

So, if we take Sober & Wilson's suggestion as an empirical claim, it is distinctly unpromising. However, since Sober & Wilson offer no evidence that 'sympathy' exists without mindreading, perhaps it's better to read their claim as a conceptual claim. Of course, on this reading, Sober & Wilson's treatment of sympathy simply does not engage the issue of what the cognitive mechanisms are that underlie altruistic motivation. Furthermore, if we approach the issue on these kinds of conceptual grounds, it's not clear that we would count 'sympathy'-induced helping behaviors as altruistic. That is, Sober & Wilson's kind of sympathy could motivate "pro-social behavior", which in fact benefits others. However, many cases of pro-social behaviors are not regarded as altruistically motivated (see, e.g., Eisenberg 1992, p. 52). For instance, if John saves a drowning child in order to impress the bystanders, then his action benefits another, but one would hardly consider it altruistically motivated. Similarly, prosocial behavior that is motivated by Sober & Wilson's sympathy would be a poor candidate for altruism. Consider, for instance, the following thought experiment. Imagine that aliens landed on this planet and didn't think that earthlings had mental states, but the cries and moans of earthlings made the aliens have negative affect analogous to Sober & Wilson's sympathy. This might motivate the aliens to engage in prosocial behavior, but, if the aliens really don't recognize that earthlings have minds, it's hard

to regard the aliens' prosocial behavior as altruistically motivated.

In sum, then, neither emotional contagion nor Sober & Wilson's sympathy provides an acceptable explanation of altruistic motivation. It's particularly clear that neither proposal offers a unified account of the core cases of altruistic motivation with which we began. Hence, if we are to have a model of altruistic motivation that can accommodate our core cases, it cannot be one of these models that rejects outright the role of mindreading.

2.3. Perspective taking accounts of altruistic motivation

In the Piaget-Kohlberg tradition, the capacity for perspective taking is thought to be essential to a wide range of moral capacities, including altruistic behavior. Unlike the no-mindreading accounts of altruistic motivation, there is no shortage of advocates for the perspective taking account of mindreading and altruism. In the recent literature, the most prevalent account of mindreading and altruism continues to be that altruistic motivation depends on perspective taking. This view is suggested by several figures including Batson (1991), Blum (1994), Darwall (1998) and Goldman (1993).

Goldman (1992, 1993) is by far the most explicit about the cognitive architecture of perspective taking, so his work provides a useful starting point. As we've seen, Goldman maintains that empathy is central to altruism, and he maintains that genuine cases of empathy depend on perspective taking. His account of perspective taking draws on his earlier work on the off-line simulation account of folk psychology (Goldman 1989, see also Gordon 1986). Goldman maintains that the process of perspective taking is subserved by off-line simulation in the following way:

> Paradigm cases of empathy... consist first of taking the perspective of another person, that is, imaginatively assuming one or more of the other person's mental states.... The initial 'pretend' states are then operated upon (automatically) by psychological processes, which generate further states that (in favorable cases) are similar to, or homologous to, the target person's states. In central cases of empathy the output states are affective or emotional states (1993, p. 351).

Now, if we try to incorporate this account of empathy into an account of altruistic motivation, we get the following account of the processes underlying altruistic motivation when the agent sees another in distress. First, the agent determines the beliefs and desires of the person in distress. Then the agent pretends to have those beliefs and desires. These pretend-states are then operated on automatically, leading to affective states that are similar to the target's state, i.e., distress. These unpleasant affective states then motivate the agent to eliminate the problem at its source, viz., the other person's distress.

Batson's picture is less architecturally explicit, but is still clearly dependent on perspective taking. Batson claims that altruistic motivation derives from "empathy" (1991, p. 83), and as Batson defines it, empathy requires perspective taking. He writes, "Perception of the other as in need and perspective taking are both necessary for empathy to occur at all" (1991, p. 85). The empathic response to perceived need "is a result of the perceiver adopting the perspective of the person in need" (1991, p. 83) and this involves "imagining how that person is affected by his or her situation" (1991, p. 83).

Blum's (1994) view is somewhat more difficult to interpret. He maintains that altruistic behavior, or "responsiveness" requires "that the child understand the other child's state" (1994, p. 197). He rejects the idea that this understanding is limited to cases in which the subject infers "the other's state of mind

from a feeling the subject herself has, or has had, in similar circumstances" (1994, p. 192). Blum rejects this account because it is too "egocentered" (1994, p. 193), and he argues that this can't be the sole cognitive process because "such inference would not account for understanding states of mind different from those one is experiencing or has experienced oneself" (1994, p. 192). Rather, Blum maintains that "understanding others means understanding them precisely *as other* than oneself – as having feelings and thoughts that might be different from what oneself would feel in the same situation" (1994, p. 193). So Blum apparently maintains that altruistic motivation depends on the understanding of others as potentially having different beliefs, desires, and emotions. But he doesn't offer an explicit explanation about how this understanding is achieved.

Although these accounts have important differences, they all share an assumption that altruistic motivation depends on some fairly sophisticated mindreading capacities. First, on Blum's account, and possibly Batson and Goldman's as well, the subject must be able to recognize that the other person might have different mental states than the subject herself would have in a similar situation. Second, for Goldman and Batson, perspective taking requires using the imagination to figure out someone else's mental states and then using the imagination to generate the other person's emotion. As a result, in sharp contrast to the emotional contagion account, the perspective taking accounts of altruistic motivation invoke quite complex mindreading capacities.

2.4. A minimal mindreading account of altruistic motivation

The accounts of altruistic motivation that make no appeal to mindreading run afoul of social psychological findings and commonsense intuitions. However, I think that we can accommodate the data and the intuitions with a much more austere proposal about the role of mindreading than the perspective taking accounts. I want to begin to sketch an account of altruistic motivation that draws on as little mindreading as necessary to accommodate the core cases of altruism, then in the next several sections, we'll consider the relative merits of the minimalist account and the perspective taking account.

The crucial feature in the core cases of altruism from social psychology is the fact that people often help even when it would be easy to escape. If the motivation is caused strictly by immediate situational cues, as in simple emotional contagion, then escape is a good alternative. However, escape is not an adequate alternative if the motivation comes from an enduring *internal* cause. The motivation seems to be relieved when the subject comes to think that the distress has been alleviated (Batson 1991), so a plausible candidate for the internal cause is a *representation of distress*. If altruistic motivation is triggered by a representation of distress, escape isn't an effective solution to the motivational problem since merely escaping the perceptual cues of distress won't eliminate the consequences of the enduring representation that another is in distress.

I suggest, then, that altruistic motivation depends on the minimal mindreading capacity to attribute distress to others.[5] On this view, a person can have the capacity for altruistic motivation even if the person doesn't have or doesn't exploit the capacity for imagining himself in the other's place and having different beliefs, desires or emotions than he himself would have in that situation. However, a person cannot have the capacity for altruistic motivation without the capacity to attribute distress to another. [6]

2.5. Arguments for perspective taking: Batson's evidence

Now that the two proposals are on the table, we can consider the arguments for each account. Although it's widely thought that altruistic motivation depends on perspective taking, it's not easy to find

an argument for the view. The only systematic argument comes from Batson's data. Batson used various methods to manipulate the "empathy" of subjects, creating conditions in which subjects would have either high empathy or low empathy. According to Batson, his evidence indicates that perspective taking is required for altruistic motivation since they found that high empathy subjects were much more likely than low empathy subjects to help in easy-escape conditions (e.g., Batson 1991, p. 87; see also Darwall 1998, p. 273). Batson's data do, I think, provide an important source of evidence against emotional contagion accounts, but they fall far short of establishing that perspective taking is required for altruistic motivation.

To begin, it's important to note that Batson's experiments cannot be decisive evidence for the perspective taking account. For the evidence does not show that altruistic motivation is absent among those with low empathy. A substantial minority of subjects in the low empathy conditions do help – averaging across studies, nearly a third of the low empathy subjects helped (Batson 1991, chap. 8). And it's quite possible that most of the other low empathy subjects had some altruistic motivation, but not enough to outweigh the competing motivation to avoid the pain of electric shock. Submitting to painful electric shock to relieve a stranger is a rather high cost action, and it seems likely that if the altruistic option were low cost (e.g., returning an elderly person's books to the campus library), then the difference between high empathy and low empathy subjects might largely disappear.

Although Batson's evidence hardly counts as a decisive argument for the perspective taking account, it does seem that the perspective taking account provides a natural explanation for why high empathy would lead to higher altruistic motivation. For if altruistic motivation depends on taking the perspective of others, then increased perspective taking might increase the motivation. However, I think that the minimalist account provides equally good explanations for Batson's findings. To see why, we need to consider in a bit more detail Batson's two central empathy manipulations: the perspective-taking manipulation (Batson 1991, p. 120) and the similarity manipulation (Batson 1991, p. 114). In the perspective-taking manipulation, subjects watched a videotape of a student with broken legs. The subjects were either told to "attend carefully to the information presented on the tape" or to "imagine how the person interviewed felt about what happened". Subjects who were told to imagine the other's feelings were more likely than subjects in the other group to help in the easy-escape condition. Although the perspective taking account can explain these results, the minimalist account can explain these results equally well. For in the high perspective-taking conditions, subjects are more likely to focus on the other's distress, and they are more likely to develop elaborate representations of the other's distress. Thus, on the minimalist account, it is hardly surprising that the perspective-taking manipulation facilitates altruistic motivation, since perspective taking implicates representations of the other's distress. In principle, it will be hard to tease apart these two theories using this kind of manipulation since if you increase a subject's perspective taking of a distressed target, you will also increase the subject's representations of the target's distress.

In Batson's other important "empathy" manipulation, subjects were shown a questionnaire putatively filled out by the person who would later need help. One group of subjects saw questionnaires that expressed similar views to those expressed on the subject's own questionnaires. The other group saw questionnaires that expressed dissimilar views. Batson and colleagues found that subjects in their high-similarity group were more likely than subjects in the low-similarity group to help in the easy-escape condition. Batson notes that previous research by Stotland (1969) and Krebs (1975) shows that subjects in high-similarity conditions show increased empathy. But there is a crucial hedge on "empathy" here. What Stotland (1969) and Krebs (1975) found was that subjects in high-similarity conditions showed heightened physiological response and expressed more concern for the other person. The level of *perspective taking* in these tasks was not measured. Nor do the researchers suggest that perspective taking is the crucial mechanism underlying the response of subjects in high-similarity conditions. There is, in fact, a large literature in social psychology suggesting that subjects are more *attracted* to people they think have similar attitudes (e.g., Newcombe 1961; Byrne 1971), and even that people are *repulsed*

by those that they think have different attitudes (Rosenbaum 1986). In light of this, it's hard to see how Batson's similarity manipulation could support the perspective taking account. What his findings do show is that we are more likely to help people who we think have similar attitudes (for a disturbing variation on this, see Tajfel 1981). Coupled with the data on similarity and attraction, we might conclude from this that we are more prone to help people that we like. That's hardly surprising. More importantly, though, it is quite irrelevant to whether altruistic motivation requires perspective taking.

2.6. Arguments against perspective taking

Thus far, we have no reason to think that altruistic motivation depends on the kind of sophisticated mindreading suggested by perspective taking accounts. In this section, I'll argue that the empirical evidence actually weighs against the perspective taking account. In trying to determine the core architecture of a capacity, contemporary cognitive scientists pay close attention to two sources of evidence: evidence from development and evidence from psychopathologies. These sources give us a glimpse into which capacities might be independent from one another and which capacities seem to be inextricably linked. I will argue that evidence from development and evidence from psychopathologies both lead to the same conclusion: altruistic motivation is independent of sophisticated mindreading abilities like perspective taking.

2.6.1. Developmental evidence

The discussion of altruism began with Blum's cases of altruism in young children. Nor are his examples atypical. Blum draws some of his examples from a large body of literature in developmental psychology. This research claims that we start seeing the kind of behavior exemplified in Blum's cases early in the second year. Radke-Yarrow, Zahn-Waxler, & Chapman (1983) found that at 10-12 months, children didn't respond like the kids in Blum's examples, but "Over the next six to eight months the behavior changed. General agitation began to wane, concerned attention remained prominent, and positive initiations to others in distress began to appear" (Radke-Yarrow, Zahn-Waxler, & Chapman 1983, p. 481). In a more recent study, Zahn-Waxler & colleagues traced the development of concern and comforting behaviors in one-year old children. They trained mothers to record their child's emotional and behavioral responses to distress in others. Mothers were also trained to simulate various distress situations. Between 13-15 months, children were reported to respond with concern to 9% of the natural distress situations and 8% of the simulated distress situations. Between 18-20 months, children responded with sad facial expressions or concerned attention to 10% and 23% of natural and simulated distress situations. And by 23-25 months, children responded this way to 25% and 27% of natural and simulated distress situations (Zahn-Waxler et al. 1992, p. 131). So it certainly appears that the capacity for concern or sympathy emerges before the age of 2. Furthermore, between 18-20 months, there is a marginally significant correlation between concern and comforting behavior, and by 23-25 months, there is a very significant correlation between concern and comforting behavior.

Despite this impressive capacity for altruistic motivation, children under the age of two have severely limited mindreading abilities. In particular, they show deficits in the two crucial features of perspective taking accounts. Before the age of 3 years, children are apparently incapable of recognizing that someone else might have a different belief than they do. The most famous result here is the young child's failure on the false belief task. In the classic version of this task, Wimmer & Perner (1983) had children watch a puppet show in which the puppet protagonist, Maxi, puts chocolate in a box and goes out to play. While Maxi is out, his puppet mother moves the chocolate to the cupboard. The children are

asked where Maxi will look for the chocolate. Children under the age of 4 fail this and similar tasks (see also Wellman 1990; Bartsch & Wellman 1995). And, although children begin to pretend by around 18 months, they seem unable to use the imagination to understand other minds until much later (see, e.g., Nichols & Stich 2000, forthcoming). Thus, since toddlers provide core cases of altruistic motivation and they lack the requisite perspective taking capacities, this provides a serious prima facie argument against the perspective taking accounts.[7]

In fact, young children's comforting behaviors offer a striking picture of both altruistic motivation and limited perspective taking. The comforting behaviors of young children tend to be "egocentric". Hoffman notes that young children's helping behaviors "consist chiefly of giving the other person what they themselves find most comforting" (1982, 287). For example, young children will offer their own blanket to a person in distress. Hoffman offers an example of a 13-month old who "responded with a distressed look to an adult who looked sad and then offered the adult his beloved doll" (1982, p. 287; see also Zahn-Waxler & Radke-Yarrow 1982, Dunn 1988, p. 97). Thus, toddlers' comforting behavior seems to be simultaneously altruistic in motivation and egocentric in perspective.

Although much early altruistic behavior is guided by "egocentric" considerations, this is perfectly compatible with the minimalist account. A common interpretation of the fact that toddlers offer their own comfort objects is that it shows that children don't really understand that it is the other person who is in distress. For instance, Hoffman (1982) claims that the fact that children tend to give their own comfort objects to help others indicates that "Children cannot yet fully distinguish between their own and the other person's inner states... and are apt to confuse them with their own" (1982, p. 287). However, the examples of "egocentric" comforting responses provide no reason to think that the child fails to distinguish its own states from the states of others. On the contrary, these responses provide evidence that the child recognizes that the other is in distress. After all, the child is offering the comfort object to the *other* person. Further, the fact that the child offers a comfort object suggests that the child does understand that *distress* is involved. Children don't try to relieve the other's distress by completely bizarre behavior, e.g., pretending that the banana is a telephone. And there's no reason to think that before 18 months, the child experimented with various means of eliminating crying in others (as one might experiment with an unfamiliar piece of electronics). However, the young child has limited mindreading resources at hand and thus relies on egocentric mindreading strategies (Nichols & Stich forthcoming). As a result, the child's knowledge of how *his distress* is relieved guides his thinking about how to relieve the other person's distress. Thus, the toddler's egocentric comforting cases are not only *consistent* with the minimalist account, the cases provide evidence that the child in fact attributes distress to the other person.

2.6.2. Psychopathological evidence: autism and psychopathy

In the mindreading literature, there has been a great deal of research on the mindreading capacities of people with autism (e.g., Baron-Cohen 1995; Frith 1989). On a wide range of mindreading tasks, autistic children tend to perform much worse than their mental aged peers. For instance, most autistic children fail false belief tasks long after their mental age peers can pass such tasks (e.g., Baron-Cohen et al. 1985). In addition to their difficulties with false belief, autistic children fail classic perspective-taking tasks, e.g., determining which gifts would be appropriate for which person (Dawson & Fernald 1987). Further, one of the central characteristics of autism is lack of imaginative activities and spontaneous pretend play (Wing & Gould 1979).

Despite their difficulties with perspective taking and imagination, recent studies show that autistic children *are* responsive to distress in others (Blair 1999; Yirmiya et al. 1992). For instance, in one recent experiment, autistic children were shown pictures of threatening faces and distressed faces, and the

autistic children showed the normal pattern of heightened physiological response to both sets of stimuli (Blair 1999). Thus, although autistic children have a deficit in perspective taking, they do respond to the distress of others. More importantly for our purposes, a recent study suggests that autistic individuals will engage in comforting behaviors. Sigman and colleagues (1992) explored responses to distress in autistic, Downs Syndrome and normally developing children. In one task, the distress was made as salient as possible. The parent was seated next to her child at a small table, and while showing the child how to use a hammer with a pounding toy, the parent pretended to hurt her finger by hitting it with the hammer. The parent then made facial and vocal expressions of distress but didn't utter any words (Sigman et al. 1992, 798). Researchers found that autistic children were much less likely than other children to *attend* to the distress. This fits with a broader pattern of inattentiveness to social cues in autism. For instance, autistic children are much less likely than Down Syndrome children to orient to someone clapping or calling their name (Dawson et al. 1998). Despite the fact that autistic children were less likely to notice or attend to the distress, several autistic children provided comfort to the parent in this experiment. Overall, few children helped, but autistic children helped as often as the children in the other groups.[8]

The fact that autistic children show normal physiological response to distress in others and the finding that autistic children do engage in comforting behaviors suggests that the core architecture for altruistic motivation is intact in autism. This poses a serious problem for the perspective taking account since that account predicts that individuals with serious deficits to imagination and perspective taking would show corollary deficits to altruistic motivation.

So, even though autistic children have a profound deficit in perspective taking, the available evidence indicates that they have no correspondingly serious deficit to altruistic motivation. The complementary question is whether there are individuals who show a deficit in altruistic motivation but no deficit to perspective taking. In fact, it's quite plausible that psychopaths fit this description. The DSM IV claims that individuals with Antisocial Personality Disorder (i.e., psychopaths) "frequently … tend to be callous, cynical, and contemptuous of the feelings, rights, and sufferings of others" (p. 647). "Persons with this disorder disregard the wishes, rights, or feelings of others. They are frequently deceitful and manipulative in order to gain personal profit or pleasure (e.g., to obtain money, sex, or power)…. They may believe that everyone is out to 'help number one' and that one should stop at nothing to avoid being pushed around" (p. 646). Thus, the DSM characterization of psychopathy certainly suggests that psychopaths are significantly less likely than non-psychopaths to exhibit altruistic behavior. Recent evidence provides an explanation for this – unlike autistic and normal children and adults, psychopaths show little or no physiological response to the distress of others (Blair et al. forthcoming). Blair and colleagues (forthcoming) found that non-psychopathic criminals show about the same amount of Skin Conductance Response to another's distress cues as they do to threatening stimuli. Psychopaths, on the other hand, show significantly greater response to threatening stimuli than they do to the distress cues of another. Further, if we look at the standardized scores (corrected for the outliers), Blair's results suggest that "the psychopaths were treating the distress cue stimuli as affectively neutral" (Blair et al. forthcoming, p. 11). Nonetheless, evidence indicates that psychopaths are perfectly capable of perspective taking, and that they perform as well as normal adults on standard perspective taking tasks (Blair 1993).

Hence, there seems to be a double dissociation between perspective taking and altruistic motivation. Young children and autistic children have immature or impaired perspective taking abilities, yet they seem to have the capacity for altruistic motivation. Psychopaths, by contrast seem to have a normal capacity for perspective taking but a severely impaired capacity for altruistic motivation. The

evidence from development and psychopathologies thus counts heavily against the perspective taking account. It seems that altruistic motivation does not require sophisticated mindreading or perspective taking abilities. And it doesn't take any imagination to be an altruist.

Although there is strong evidence against the perspective-taking model, it would be derelict to claim a quick victory for the minimalist account that I've proposed. For there is a less austere alternative that is not excluded by the evidence. By the time toddlers exhibit comforting behaviors, they probably have the capacity to attribute desires that they don't have (see, e.g., Repacholi & Gopnik 1997). So one might maintain that it is *this* mindreading capacity, the capacity to attribute discrepant desires, that is essential for altruistic motivation. This view has not been elaborated and defended in the literature, but it's possible that the view is close to Blum's (1994) account. Recall that Blum maintains that the understanding of others required for altruistic motivation depends on understanding that others might have thoughts and feelings that are "different from what oneself would feel in the same situation" (p. 193). He rejects more austere accounts as too "egocentered" (p. 193).

While this moderate "discrepant desire" position doesn't contravene any of the data, it's unclear why the capacity to attribute discrepant desires (or any other discrepant mental states) should be essential to altruistic motivation. To see this, it's important to distinguish between three different kinds of egocentrism. One kind of egocentrism is just the view that an individual's basic motivations derive solely from that individual's own affective or hedonic states. We might call this view *psychological egoism*. Psychological egoism might be wrong, but the issue belongs to the foundations of cognitive science, not to moral psychology. On the second kind of egocentrism, let's call it *ethical egocentrism*, a person is egocentric if none of the individual's desires are directed at another person's needs, except insofar as the individual thinks that addressing the other person's needs will help him. What's crucial about ethical egocentrism (and what distinguishes it from simple psychological egoism) is that if a person is ethically egocentric, he must go through a process of instrumental reasoning before arriving at a motivation to help another. For he must *think* that helping another will benefit himself. Both of these kinds of egocentrism need to be distinguished from a third kind of egocentrism – *mindreading egocentrism*. To say that someone is egocentric in this sense is to claim that the individual either can't or tends not to grasp that others have different likes and dislikes, different judgements, and different feelings than the individual himself. Notice that ethical egocentrism and mindreading egocentrism make quite independent claims. A person can perfectly well be ethically egocentric without being an egocentric mindreader. That is, a person might know that others have different interests and beliefs than he does, but at the same time, he might not care in the least about the interests of others, except insofar as he thinks it will affect him. Psychopaths seem to fit this characterization fairly well. Conversely, a person could be an egocentric mindreader without being ethically egocentric. That is, a person might be oblivious to the fact that others have different desires and thoughts than she does, but she might care about trying to help others in need, even if she doesn't think that doing so will serve her own interests. Of course, if she is an egocentric mindreader, she may not be very effective at helping others, because she won't be sensitive to the variation in desires, feelings and thoughts that actually exist among those she tries to help. Now, finally, we can get to the point of drawing these distinctions – if someone is an egocentric mindreader, that provides no reason to conclude that she lacks altruistic motivation. The kind of egocentrism that undermines the claim for altruistic motivation is *ethical* egocentrism, not *mindreading* egocentrism. As we've seen, when toddlers offer comfort, they often offer their own comfort objects to others. The fact that these children are using egocentric mindreading strategies does not undermine the claim that these children are altruistically motivated. Even if children turned out to be completely egocentric mindreaders, I see no reason to conclude that their attempts to comfort adults with their dolls and blankets would not be the product of altruistic motivation. Thus, although the discrepant desire view fits with the available evidence, it's not at all clear why we should prefer this account to the simpler minimalist theory.

2.7. Affect and altruistic motivation

I've argued that altruistic motivation depends only on the minimal mindreading capacity for distress attribution, but I've said nothing about how attributing distress to another leads to altruistic motivation. In keeping with most other accounts, I will assume that altruistic motivation is mediated by an affective response (see e.g., Eisenberg 1992, Goldman 1993, Hoffman 1991). So, on the account I'm suggesting, the attribution of distress triggers an affective response that generates the motivation to help the person in distress. However, there are a couple of importantly different possibilities for the character of the affective response. One possibility is that the representation of the other's distress produces a distinctive emotion of sympathy or concern for the other person and this emotion is not homologous to the emotion of the person in need. The sympathy view has some support from an emerging body of research which ties altruistic behavior to a distinctive facial expression labeled "concern" (Roberts & Strayer 1996, 456; Eisenberg et al 1989, p. 58; Miller et al 1996, 213) There is also a bit of evidence that sympathy might have distinctive physiological characteristics (Eisenberg & Fabes 1990, p. 140; Miller et al. 1996). Facial expression and physiological signs are the kind of features that have been used to delineate "basic emotions" (e.g., Ekman 1992). The exciting possibility here is that sympathy is a genuine, distinctive basic emotion with a distinctive facial expression and physiological profile and that this emotion is the motivation behind altruistic behavior. Darwin himself actually made a similar suggestion: "Sympathy with the distresses of others, even with the imaginary distresses of a heroine in a pathetic story, for whom we feel no affection, readily excites tears…. Sympathy appears to constitute a separate or distinct emotion" (Darwin 1872, p. 215). But Darwin seems to have had a somewhat different notion of sympathy in mind since he thinks that we can sympathize with the happiness of others.

The possibility that altruistic motivation derives from a distinctive basic emotion of sympathy is exciting, but it has turned out to be difficult to get unequivocal data correlating the postulated features of sympathy with altruistic behavior. There are several different measures – e.g., self-report, facial expressions, physiological measures. The findings suggest that some of these features are correlated with altruistic behavior some of the time. For example, Eisenberg & Fabes (1990) showed preschoolers a film of children who were injured and in the hospital, and the preschoolers were "given the opportunity to assist the needy others by packing crayons in boxes for the hospitalized children rather than playing with attractive toys." (Eisenberg & Fabes 1990, p. 140). Although children's self-reports were unrelated to their helping behavior, the physiological measure of sympathy (heart-rate deceleration) was positively correlated with higher levels of helping (Eisenberg & Fabes 1990, pp. 140-1). Further, facial expressions of concerned attention have been significantly correlated with greater helping in boys, but the findings are much weaker for girls (Eisenberg & Fabes 1990, p. 141). And there is a bit of evidence that there is a correlation between these emotions and the conditions set up in Batson-style experiments (Eisenberg et al. 1989).

Notice that if the above account of the affect is right, sympathetic motivation for altruism doesn't count as empathy at all. Rather, altruistic behavior is motivated by a distinctive emotion that is not homologous to the emotion felt by the person in need, or indeed homologous to any other emotion.[9] This would entail that a certain class of empathy-based accounts is thoroughly mistaken. If empathy is a vicarious feeling of the emotion that the target is feeling (caused by perspective taking or emotional contagion), then the empathy account is wrong not just about the mindreading involved in altruistic motivation but also about the affect. For on the sympathy account, the emotion driving altruistic behavior does not parallel any other emotion. So, except in the iterative case of empathizing with someone feeling sympathy, empathy will not produce the emotion that generates altruistic behavior.

Although the idea that a distinctive emotion of sympathy underlies altruism is theoretically

appealing, there is another possibility. The distress attribution might produce a kind of second order empathic distress in the subject. For example, representing the sorrow of the target might lead one to feel sorrow. This would provide a kind of empathic motivation for helping. And the motivation would be effective even when escape is easy. For the cause of the emotion is still the representation of the other's mental state and as a result, one is motivated not simply to escape the situation since that would not rid one of the representation. As a result, this story would provide an equally effective explanation of Batson's data. And some of the above research on sympathy actually provides support for this alternative story. For instance, Eisenberg and colleagues (1989) found that the strongest predictor of helping in adults was not facial sympathy, but facial sadness (Eisenberg et al. 1989, 61). The available evidence doesn't really decide between these two accounts of the affect underlying altruistic motivation. Indeed, perhaps both affective mechanisms are operative.[10]

2.8. The Concern Mechanism

For present purposes, what is really crucial is not the character of altruistic affect (whether it's a distinctive emotion or homologous to some other emotion) but the broader characterization of the cognitive mechanisms implicated in altruistic motivation. We are now in a position to state the proposal about the core architecture a bit more precisely. Altruistic motivation depends on a mechanism that takes as input representations that attribute distress, e.g., *John is experiencing painful shock*, and produces as output affect that *inter alia* motivates altruistic behavior. To avoid the terminological difficulties with 'sympathy' I'll use a slightly less problematic term and call this system the Concern Mechanism.

Given this account, it's likely that the Concern Mechanism is a "module", on at least some construals of modularity. To be sure, the Concern Mechanism has many of the features of modules as set out by Fodor (1983). It's plausible that the mechanism is fast and that its operation is largely mandatory. The evidence on development and psychopathology indicates that it has a characteristic ontogeny and a characteristic pattern of breakdown. And, on certain conceptions of modules (e.g., Baron-Cohen 1995), that suffices for modularity. However, one additional feature that is regarded as crucial for *Fodorian* modules is "encapsulation" (Fodor 1983, forthcoming), and the relationship between affective systems and encapsulation is far from clear in the current literature. A cognitive mechanism is encapsulated if it has little or no access to information outside of its own proprietary database. The currently preferred experimental methodology for demonstrating that a mechanism is encapsulated is to show that there is information that is obviously relevant and available in the cognitive system as a whole, but the mechanism ignores it and thus gets the wrong answer on certain tasks (see, e.g., Hermer & Spelke 1996). This approach to arguing for encapsulation does not convert unaltered to the study of affective systems, since it's less clear what it is for an affective system to make a mistake (see, e.g., D'Arms & Jacobson forthcoming). In arguing for encapsulation, the most relevant instances of "mistakes" of affective systems are plausibly cases in which the affective response is resistant to practical reason. For one of the hallmark debits of an encapsulated system is that such systems resist our preferences: You can't make the Muller-Lyer illusion disappear by wanting it to go away. It's likely that the Concern Mechanism is similarly resistant to our preferences and to the dictates of practical reason. We might think that it would be best, all things considered, not to feel concern in some circumstances when we know about another's distress, but it's likely that wanting not to feel concern in these situations is often not sufficient to stop the system from producing the affect. For instance, I might think it's best, all things considered, not to feel concern when my daughter gets inoculated because any show of concern on my part might intensify her anxiety about inoculations. Nonetheless, it might be extremely difficult to suppress concern in these circumstances. Of course, I don't have any non-anecdotal evidence that the Concern Mechanism resists preferences in this way. But if, as seems likely, the Concern Mechanism does resist preferences and practical reason, then that suggests that the mechanism is encapsulated to some non-trivial extent.

If the basic story about mindreading and the Concern Mechanism is right, it has a particularly interesting implication for the possibility of altruism in non-human animals. For if human altruism requires so little mindreading, it becomes quite possible that the mechanisms underlying helping-behavior in some non-human animals are analogous to the mechanisms underlying altruistic motivation in humans. Although it's hotly debated at present, some non-human animals may well have the mindreading capacity to attribute distress to another. There is some evidence, for instance, that chimpanzees can attribute goals (Premack & Woodruff 1978; Uller & Nichols forthcoming). Research also suggests that non-human primates are sensitive to a conspecific's distress signals (e.g., Miller et al. 1963).

Apart from its intrinsic interest, the possibility that altruistic motivation might be present in non-humans is of some importance to an evolutionary approach to altruism. If altruistic motivation in humans is an adaptation that depends on sophisticated mindreading abilities like perspective taking, then altruistic motivation must have been shaped after the evolution of our sophisticated mindreading abilities. If so, the mechanisms for altruistic motivation must have emerged relatively recently in evolutionary time since, by most accounts, humans are the only species with sophisticated mindreading abilities. The Concern Mechanism account of altruistic motivation, on the other hand, needn't be committed to the view that altruistic motivation is a recent adaptation since on this view the requisite mindreading mechanisms are minimal and may well have been present in our more distant phylogenetic ancestors.

3. Cognitive mechanisms in moral judgement

In addition to altruistic motivation, the other basic capacity of moral psychology that has been intensively studied recently is the capacity for moral judgement. Moral judgement has been at the center of research in moral psychology for both philosophers and psychologists for decades. Within psychology, this capacity has perhaps been most directly approached empirically by exploring the basic capacity to distinguish moral violations (e.g., hitting another person) from conventional violations (e.g., playing with your food). This tradition in psychology began with the work of Elliott Turiel and has flourished over the last two decades (see, e.g., Turiel 1983, Nucci 1986, Turiel et al. 1987, Dunn & Munn 1987, Smetana & Braeges 1990, Blair 1993). However, only recently have there been attempts to characterize the cognitive mechanisms underlying moral judgement (e.g., Blair 1995, Goldman 1993). One central question is the extent to which moral judgement depends on the capacity for mindreading. Blair (1995) maintains that moral judgement is independent of the capacity for mindreading and depends on a Violence Inhibition Mechanism. Several others (e.g., Goldman 1993, Gordon 1995, Deigh 1995) suggest that moral judgement depends on the capacity for perspective taking. In this section, I'll argue that none of these theories is adequate. Rather, I maintain that some capacity for mindreading is essential to moral judgement, but not the sophisticated capacity for perspective taking. However, recent evidence indicates that this can only be part of the story about moral judgement. I maintain that moral judgement also depends on the Concern Mechanism that is central to altruistic motivation. So, on the account I'll develop, moral judgement depends both on some capacity for understanding other minds and on a certain affective mechanism.

In his influential work on moral judgement, Turiel explicitly draws on the writings of several philosophers, including Searle, Brandt and Rawls to draw the moral/conventional distinction. Turiel

characterizes the distinction as follows: "Conventions are part of constitutive systems and are shared behaviors (uniformities, rules) whose meanings are defined by the constituted system in which they are embedded" (Turiel et al 1987, p. 169). For example, the prohibition against chewing gum in class is a conventional rule. Moral rules, on the other hand, are "unconditionally obligatory, generalizable, and impersonal insofar as they stem from concepts of welfare, justice, and rights" (Turiel et al 1987, pp. 169-170). The prohibition against pulling hair in class is an example of a moral rule.

The research program generated by Turiel's work indicates that people distinguish moral violations from conventional violations along several dimensions. One might question whether the data really show that people adhere to Turiel's characterization of the moral and conventional domains, but there is no doubt that people do distinguish central examples of moral violations from conventional violations in several ways. This is the basic capacity that I want to explore.

3.1. Core cases of moral judgement

Rather than embark on an attempt to define the moral and conventional domains, the easiest way to see the import of the data on moral judgement is to consider how subjects distinguish between prototypical examples of moral violations and prototypical examples of conventional violations. Hitting another person is a prototypical example of a moral violation used in these studies. For instance, in Smetana & Braeges, subjects are shown a colored drawing and told "This child is hitting this child" (Smetana & Braeges 1990, p. 334). Other frequently used examples of moral violations are pulling hair, stealing, and pushing another child. The examples of conventional violations that have been studied are much more varied. Some of the examples are violations of school rules, e.g., not paying attention during storytime or talking out of turn. Some of the examples are violations of etiquette, e.g., drinking soup out of a bowl. Other examples are violations of family rules, e.g., not clearing one's dishes. What is striking about this literature is that, from a young age, children distinguish the cases of moral violations from the conventional violations on a number of dimensions. For instance, children tend to think that moral transgressions are generally less permissible and more serious than conventional transgressions. And the explanations for why moral transgressions are wrong are given in terms of fairness and harm to victims, whereas the explanation for why conventional transgressions are wrong is given in terms of social acceptability. Further, conventional rules, unlike moral rules, are viewed as dependent on authority. For instance, if at another school the teacher has no rule against chewing gum, children will judge that it's not wrong to chew gum at that school; but even if the teacher at another school has no rule against hitting, children claim that it's still wrong to hit. Indeed, a fascinating study on Amish teenagers indicates that moral judgements are not even regarded as dependent on *God's* authority. Nucci (1986) found that 100% of a group of Amish teenagers said that if God had made no rule against working on Sunday, it would not be wrong to work on Sunday. However, more than 80% of these subjects said that even if God had made no rule about hitting, it would still be wrong to hit. These findings on the moral/conventional distinction have turned out to be quite robust. They have been replicated numerous times using a wide variety of stimuli (see Turiel et al. 1987 and Smetana 1993 for reviews). Thus, it seems that, like the capacity for altruistic motivation, the capacity for drawing the moral/conventional distinction is part of basic moral psychology.

Most of the above research on the moral/conventional distinction has focused on moral violations that involve harming others, and that will be my main focus as well. However, it's clear that harm-centered violations do not exhaust the moral domain. To take one obvious example, adults in our society make moral judgements about distributive justice that have little direct bearing on harm. Furthermore, recent evidence indicates that the moral domain may not even be cross-culturally stable (e.g., Miller et al. 1990; Haidt et al. 1993). In a clever study by John Haidt and colleagues, they found that on several different dimensions, low-SES subjects treated disgusting actions (e.g., eating your dog or having sex with a dead chicken) as they did moral violations, whereas high-SES subjects did not treat disgusting

actions in this way (Haidt et al. 1993). Thus, Haidt and colleagues suggest that it is parochial to think that harm is central to drawing the moral/conventional distinction (e.g., Haidt et al. 1993, p. 625). However, although there may be some relativity in the moral domain, the cross-cultural work also indicates that in all cultures, canonical examples of moral violations involve harming others (see, e.g., Hollos et al. 1983; Nucci et al. 1983; Song et al. 1987). Indeed, even Haidt and colleagues found that the different SES groups did not show a difference in their judgements about violations involving harm – e.g., they thought that a girl who pushes a boy off a swing should be punished or stopped.

Thus, even though the moral/conventional distinction seems to show up in violations that do not involve harm, it's quite plausible that judgements about harm-based violations constitute an important core of moral judgement. For the appreciation of harm-based violations shows up early ontogenetically (as we will see in section 3.3), and it seems to be cross-culturally universal. Brian Scholl and Alan Leslie make a related point about theory of mind (Scholl & Leslie 1999). They note that, although there are cross-cultural differences in theory of mind, all cultures seem to share a core theory of mind which emerges early ontogenetically, what they call "early theory of mind" (p. 14 [ms]). Something similar might be said about the findings on moral judgement – although there may be cross-cultural differences in moral judgement, the evidence indicates that all cultures share an important core, what we might call "early moral judgement". The capacity to distinguish harm-based moral violations from conventional violations seems to be an important part of this early moral judgement, and this will be the sense of "moral judgement" that is intended throughout this paper.

3.2. Blair's VIM-account

Armed with a dazzling series of experiments, James Blair has developed the most detailed cognitive account of moral judgement in the recent literature. Blair maintains that moral judgement derives from the activation of a Violence Inhibition Mechanism (VIM). The idea for VIM comes from Lorenz' (1966) suggestion that social animals like canines have evolved mechanisms to inhibit intra-species aggression. When a conspecific displays submission cues, the attacker stops. Blair suggests that there's something analogous in our cognitive systems, the VIM, and that this mechanism is the basis for our capacity to distinguish moral from conventional violations. On Blair's account the process seems to go as follows. The VIM is triggered by distress cues or by associations to distress cues; this VIM activation is experienced as aversive; and events that activate VIM are accordingly judged as bad (Blair 1993, 83, 88; Blair 1995, 7).[11]

One important feature of Blair's account is that it proposes that moral judgement is independent of other capacities including, crucially, the capacity for mindreading. According to Blair, since VIM is independent of mindreading capacities, one can draw the moral/conventional distinction even if one lacks the ability to represent mental states. Blair tries to support this claim by appealing to his data on autism. As noted earlier Blair (1999; see also Yirmiya et al. 1992) found that autistic children do show normal physiological responses to distress cues. Blair also found that autistic children were able to make the moral/conventional distinction. For instance, autistic children judged that conventional rules are more modifiable than moral rules (Blair 1996, p. 577). Further, although some autistic children do pass the false belief task, Blair notes that in his experiment, "level of ability on false belief tasks is not associated with the tendency to distinguish moral and conventional transgressions" (Blair 1996, p. 577). Blair suggests that this evidence shows that the capacity for mindreading or 'mentalizing' is entirely dissociated from the capacity to draw the moral/conventional distinction: "Children with autism have been demonstrated to be incapable of 'mentalizing' (e.g., Baron-Cohen, Leslie & Frith 1985)" and so, they are incapable of "forming a representation of the mental state of the other" (Blair 1995, 22). He maintains that his VIM theory explains how autistic children can make the moral/conventional distinction

even though they can't mentalize: "While children with autism may not be able to represent a mental state of another's distress, this distress, as a visual or aural cue, will activate their VIM" (Blair 1995, 22).

Blair's VIM proposal has several shortcomings. First, Blair's attack on the role of mindreading is unconvincing. Claiming that autistic children can't "mentalize" or that they can't represent the mental states of others overstates their deficit. There is reason to think that autistic children can represent some mental states. Autistic children can attribute perceptual states to others, e.g., they can specify which object another person will identify as "in front" (Tan & Harris 1991). Further, autistic children are capable of attributing simple desires and emotions (e.g., Tan & Harris 1991; Yirmiya et al. 1992). They understand that people can have different desires and "that someone who gets what he wants will feel happy, and someone else who does not get what he wants will feel sad" (Baron-Cohen 1995, 63). Furthermore, studies of spontaneous language use in autistic children indicate that these children use the term 'want' appropriately and often (Tager-Flusberg 1993). Thus there is good reason to think that the capacity for attributing desires is largely intact in autistic children (see also Nichols & Stich forthcoming). As a result, the fact that these children can distinguish moral from conventional violations does not provide evidence that the capacity for making this distinction is entirely independent from mindreading.

Not only is Blair's rejection of mindreading in moral judgement unsupported, Blair's no-mindreading account faces serious problems. First, it's not clear how VIM-activation is related to moral judgement. On the most obvious reading of the proposal, the idea is that distress cues activate VIM and when VIM is activated, it produces aversive experience that leads to the judgement that the event which caused VIM-activation is morally bad. But if this is the proposal, it seems obviously wrong. For the class of moral violations is clearly not the same as the class of events that trigger the Blair's VIM. On Blair's theory, the VIM is triggered by distress cues. But in many cases, we see distress cues and have aversive experience without drawing moral judgements. For instance, when we witness victims of a natural disaster or when we see people accidentally hurt themselves or others, we often have an aversive response to the distress cues. But in at least many of these cases, we do not draw moral judgements. We don't, for instance, judge that the person committed a moral violation by tripping on a rock. In addition, we often have aversive experience on perceiving obviously superficial signs of distress. Blair's own method for testing for VIM is to show subjects a photograph of a crying child, and if subjects show heightened physiological response, that indicates VIM-activation (Blair 1999). This strategy would presumably work just as well if paintings were used rather than photographs. So, VIM can be activated whether or not one believes that the other person is in distress. Indeed, this is crucial for Blair's view on autism and moral judgement. According to Blair, even though autistic children cannot represent distress, they have an intact VIM, and this is the basis for their capacity for moral judgement. However, the production of artificial distress cues (e.g., by playing a tape of simulated crying) is not commonly taken to be a moral violation. Even though artificial distress cues can lead to aversive experience, such experiences will not lead us to make moral judgements if we don't think that someone has actually been caused distress.

So, VIM can't do all the work of picking out the class of moral violations. For in cases of natural disasters, accidents, and artificial distress cues, the VIM will be activated but we will not draw a moral judgement. Perhaps the most plausible way to remedy this problem is to maintain that there is a body of information about which events count as moral transgressions, and this body of information excludes the non-moral cases. However, the most obvious way to exclude the problematic cases would appeal to facts about whether someone is really being caused distress and whether the distress is caused intentionally. But, of course, Blair can't appeal to these sorts of strategies since they clearly implicate the capacity for mindreading. As a result, it's quite unclear how the VIM account could explain moral judgement without appealing to the capacity for mindreading. And since Blair's argument from autism is unconvincing, there's no reason to adopt that problematic position.

In addition to the difficulty of explaining how VIM can subserve moral judgement without mindreading, there's a further problem with the VIM part of Blair's proposal. For it's not at all clear that VIM, as Blair characterizes it, exists. Blair adopts Lorenz' view that animals in some social species have a mechanism designed to inhibit aggression in response to submission cues, and Blair maintains that an analogous mechanism is present in humans. However, Lorenz' account of the aggression-inhibition mechanism is now widely rejected on evolutionary grounds. According to Lorenz, the mechanism for inhibiting aggression evolved because having such a mechanism would benefit the species, which would otherwise lose many of its members (Lorenz 1966). Thus the view is committed to an untenably strong form of group selection (see, e.g., Williams 1966). A more widely accepted view of submission cues and the responses they elicit is that such cues play a crucial role in establishing rank within a group (e.g., Tomasello & Call 1997). For example, even in the absence of an attack, a submission cue from a rival male chimpanzee will lead to reconciliation between the chimpanzees (see, e.g., de Waal 1996). But of course, if this work in comparative psychology is right, the mechanism that happens to cause violence inhibition does not have violence inhibition in response to distress cues as its special purpose or core function. Indeed, there well may be no special purpose mechanism for aggression inhibition in any animal.

So, Blair's VIM account of moral judgement has some serious flaws. His claim that mindreading isn't required for moral judgement is unsupported, and it's quite unclear how VIM is supposed to underwrite moral judgement without recourse to mindreading. Indeed, there's reason to doubt that VIM even exists.

3.3. Perspective-taking accounts of moral judgement

Since Blair's no-mindreading account of moral judgement is pretty clearly inadequate, we need to explore which mindreading capacities might be required for moral judgement. Unfortunately, the positive proposals about mindreading in moral judgement have not been nearly so crisp as Blair's account. But there are some hints in the literature that perspective taking is required for moral judgement. After sketching empathy as perspective taking, Goldman cites Schopenhauer as an advocate of the view that empathy is "the source of moral principles" (1993, p. 355). One plausible interpretation of this is that perspective taking is essential for moral judgement, at least if we construe "moral judgement" broadly. Certainly, the view that perspective taking is vital to moral judgement (broadly construed) has deep roots in developmental psychology stretching back to Piaget. More recently, the view has been elaborated in philosophy by John Deigh (1995). Deigh claims that in order to grasp right and wrong in the deeper sense, one needs mature empathy, which involves *inter alia,* "taking this other person's perspective and imagining the feelings of frustration or anger" (Deigh 1995, p. 758). Robert Gordon offers a more sophisticated strategy for determining whether an action is wrong, suggesting that we "imagine being in X's situation, once with the further adjustments required to imagine being X in that X's situation and once without these adjustments. If your response is the same in each case, approve X's conduct; if not, disapprove" (Gordon 1995, p. 741).

Now, surely people sometimes use perspective taking in making moral evaluations. And the above authors aren't sufficiently precise about which kinds of moral judgements depend on perspective taking to allow us to determine whether they would maintain that the basic capacity to make the moral/conventional distinction depends on perspective taking. But the work on the moral/conventional distinction currently provides the clearest way to explore the basic capacity for moral judgement, so it will be of interest to see how a perspective taking account of this capacity fares against the evidence in any case.

3.4. Arguments against perspective taking

As noted, it's not at all clear whether Goldman (1993), Gordon (1995) or Deigh (1995) are committed to the view that drawing the moral/conventional distinction depends on the capacity for perspective taking. There is certainly no systematic argument in the recent literature for the view that perspective taking is required for drawing the moral/conventional distinction.[12] But it is much clearer that any attempt to defend that position will face some serious obstacles.

3.4.1. Developmental evidence

Not surprisingly, we don't find a full-blown moral/conventional distinction at the age when children begin engaging in altruistic behavior. On everyone's account (except, perhaps, Blair's), the capacity to draw the moral/conventional distinction is more cognitively demanding than the capacity to be concerned about another. Nonetheless, children begin to appreciate features of the moral/conventional distinction surprisingly early. Smetana & Braeges (1990) found that at 2 years and 10 months, children already tended to think that moral violations (but not conventional violations) generalized across contexts when asked, "At another school, is it OK (or not OK) to X?" (p. 336). Further, according to Smetana and Braeges, after factoring in corrections for language, the results suggest that children generalize moral violations in this way shortly after the 2nd birthday, and they recognize that conventional violations but not moral violations are contingent on authority at 2 years and 10 months (Smetana & Braeges 1990, p. 342). So, there are some pretty impressive indications that young children can make these distinctions in controlled experimental settings. In addition, studies of children in their normal interactions suggest that from a young age, they respond differentially to moral violations and social violations (e.g., Dunn & Munn 1987; Smetana 1989). The developmental evidence thus provides some reason to be skeptical of the perspective taking account. Although at 2 years and 10 months, children have some mindreading capacities, their perspective taking abilities are still quite limited. It's especially unlikely that they are able to determine the other's beliefs and desires and then pretend to have those beliefs and desires.

3.4.2. Psychopathological evidence: autism and psychopathy

The evidence on children is hardly conclusive, but Blair's evidence on autism and moral judgement is rather more compelling against perspective taking accounts. Although Blair overstates the deficits found in autism, he is right to note that the data on autism pose a significant problem for perspective taking proposals (Blair 1993). For there is no doubt that autistic children have deficits in perspective-taking and other sophisticated mindreading capacities. Hence, Blair's data on autistic children suggest that sophisticated mindreading abilities are not required to draw the moral/conventional distinction. Since imagination is impaired in autism, the data also indicate that the capacity to draw the moral/conventional distinction does not depend on imagining oneself to be in the other's situation.

Again, we might wonder whether there are individuals who can engage in perspective taking but fail to distinguish moral from conventional violations. Blair's work suggests that this is the case in psychopathy. Blair tested psychopaths' capacity to make the moral/conventional distinction by exploring their understanding of a variety of violations (e.g., one child hitting another child; two children talking in class) (Blair 1995, pp. 16-17). Since the pool of psychopaths was drawn from a prison population, Blair used non-psychopathic prison inmates as a control. Blair used found that control criminals made a

significant moral/conventional distinction on permissibility, seriousness and authority dependence; psychopaths, on the other hand, didn't make a significant moral/conventional distinction on any of these dimensions. Further, the psychopaths were much less likely than the control criminals to justify rules with reference to the victim's welfare. Rather, psychopaths typically gave conventional-type justifications for all transgressions.[13] Although psychopaths have difficulty with the moral/conventional distinction, as we saw above, psychopaths seem to have no serious deficiency in perspective taking (Blair 1993).

Thus, the empirical evidence suggests that the capacity to distinguish moral from conventional violations does not depend on the capacity to engage in perspective taking. Furthermore, as with altruistic motivation, it's unclear why the basic capacity for moral judgement should depend on the capacity to understand that others might have beliefs and desires that are different from those oneself would have in the same situation. One of the ways that children distinguish moral violations from conventional violations is by claiming that moral violations are universally wrong. It's always wrong to pull hair; it's always wrong to steal from another; and it's always wrong to hit. Of course, as adults, we know that these particular claims are oversimplified -- for example, perhaps it's not wrong to hit a masochist under certain conditions. Most children don't know about masochists; they assume that everyone has the desire not to be hit. But it would be perverse to think that, because the child doesn't appreciate that others might have different desires in this domain, the child isn't really drawing the moral/conventional distinction.

3.5. A Minimal-mindreading alternative

Thus, the empirical evidence suggests that the capacity to distinguish moral from conventional violations does not depend on the capacity to engage in perspective taking. Leaning on the earlier proposal about altruistic motivation, I suggest that the capacity for drawing the moral/conventional distinction in these tasks depends on only minimal mindreading including, crucially, the capacity to attribute distress to others. So, the idea is that a subject must have the capacity to attribute distress in order to have the capacity to distinguish conventional violations from moral violations in the classic tasks. However, on the minimalist account, drawing the moral/conventional distinction does not require the capacity for perspective taking.

I hasten to add that I'm not claiming that the capacity to attribute distress is *sufficient* for the capacity to draw the moral/conventional distinction. On the contrary, it's likely that young children have a considerable body of information guiding their judgements about moral violations, a normative "theory". One crucial feature of this body of information about moral violations is that it probably can't be captured by a simple rule like *an action is wrong if it causes distress*. Clearly sometimes a person can cause distress without eliciting negative moral judgements. For example, some actions of dentists probably fit this description. Pre-school children do understand that causing distress isn't always wrong. So it's likely that there is a body of information that provides the basis for distinguishing wrongful harm from acceptable harm, and this may be present quite early in development. The important point for our purposes is that this body of information, this normative theory, depends on a minimal capacity for mindreading. Hence, the capacity for minimal mindreading is *necessary* for the capacity for early moral judgement.

Although the evidence on development and autism poses a serious problem for the perspective taking account, that evidence fits comfortably with the minimal-mindreading account. By the time children distinguish moral and conventional violations, they are certainly capable of distress attribution.

And as noted earlier, it's likely that, despite their deficits in perspective taking, autistic children are capable of distress attribution. Further, appealing to minimal mindreading capacities will suffice to avoid the problems that arose for Blair's outright rejection of the role of mindreading in moral judgement. For it only requires minimal mindreading to have the capacity to distinguish accidents from intentional actions and to distinguish superficial distress cues from real distress cues.

3.6. The Concern Mechanism and moral judgement

Although the minimalist account captures much of the data better than the alternative proposals, the minimalist account alone, like the perspective taking account, provides no explanation for Blair's finding that psychopaths fail to draw the moral/conventional distinction. Clearly, psychopaths have the capacity for minimal mindreading, so their apparent inability to draw the moral/conventional distinction requires some other explanation. Blair's own explanation is that psychopaths lack VIM. He found that psychopaths showed significantly less physiological response to distress in others than autistic children and normal adults (Blair et al. forthcoming), and he uses this as evidence that VIM is absent in psychopaths. However, it's not clear how the absence of VIM in psychopaths would explain their performance on the moral judgement task. Furthermore as argued above, it's not clear that VIM exists in *anyone*.

A more plausible interpretation of these data, as argued in section 2, is that the Concern Mechanism is defective in psychopathy. On this interpretation, Blair's findings suggest that there is a correlation between lacking the Concern Mechanism and failing to draw the moral/conventional distinction. Hence, a prima facie plausible hypothesis is that the Concern Mechanism plays a crucial role in the capacity to draw the moral/conventional distinction on these tasks. But, of course, this does not address *how* the Concern Mechanism might be implicated in drawing the moral/conventional distinction. I'll offer a tentative and speculative suggestion in what follows.

To give a detailed account of the relation between affective mechanisms and moral judgement would require a serious consideration of the philosophical treatments of moral discourse. There is a huge body of literature in metaethics on this topic (see, e.g., the essays in Darwall et al. 1997), and I couldn't possibly do justice to the issue in the present context. Rather, I want to try to explain the crucial difference between the responses of psychopaths and control criminals in Blair's moral judgement experiments (1995). Notice that the difference between psychopaths and control criminals is *not* that psychopaths don't grasp normative judgements at all. In some sense, psychopaths know the difference between right and wrong. They correctly note that the child shouldn't talk out of turn or hit another child. What they fail to do is distinguish between these two kinds of normative judgements. That is, what Blair's findings indicate is that, while psychopaths know the difference between right and wrong, they don't appreciate the difference between (conventional) wrong and (moral) wrong.

By Blair's reckoning, the telling feature about the psychopaths' performance is that they offer conventional-type justifications for moral violations rather than justifications in terms of harm to the victim (Blair 1995, 24). But presumably psychopaths are well aware of the fact that hitting another falls under the general category of harm-based violations. For instance, if you presented psychopaths with a novel case of causing harm to another, they would likely be able to generalize and say that it's wrong. Similarly, control criminals are well aware that hurting others is socially discouraged. So both control criminals and psychopaths presumably have available to them both social-convention explanations and harm-based explanations of the moral violations. Yet when asked why it's wrong to hit another child, control criminals and psychopaths offer different explanations. Why is that? To answer this question, I think we need to appeal to a bit of pragmatics. For Gricean reasons (the first maxim of Quantity, to be

precise [Grice 1975, 45]), we can expect subjects to try to give the most informative answer appropriate to the question. For the control criminals (and other non-psychopaths), the Concern Mechanism gives a special salience and priority to distress in others. As a result, when control criminals are asked why it's wrong to hit another child, the fact that distress in others activates the Concern Mechanism leads them to explain the wrongness in terms of the victim's distress. Again, the control criminals know perfectly well that hitting others is socially discouraged, but their emotional response to distress leads them to appeal to the victim's distress as a more informative or deeper explanation for their judgement. For psychopaths, on the other hand, since they don't have an intact Concern Mechanism, the distress of others does not have the special salience it does for the control criminals; so for the psychopaths, the most informative explanation that is appropriate to the question is the social-convention explanation. Again, this does not mean that the psychopaths are unaware of the fact that the general category that is discouraged is the category of harming others. However, lacking an intact Concern Mechanism for processing distress attributions, they presumably regard the social-convention explanation as more informative than the fact that the transgression involves harming. Thus, I suggest that the Concern Mechanism plays an essential role in distinguishing conventional violations from moral violations.[14]

This account has an important parallel with part of Blair's account. Blair argues that his data on moral judgement show that there is a double dissociation between VIM and Theory of Mind *tout court* (e.g., Blair 1995). I think that this is almost right – his data do suggest an important double dissociation. But it's not between VIM and Theory of Mind *tout court*. Rather, the apparent double dissociation is between the Concern Mechanism and a certain class of mindreading abilities that includes perspective taking. Psychopaths seem to have sophisticated mindreading capacities, including the capacity for perspective taking, but they apparently lack the Concern Mechanism. Autistic children, by contrast, lack the requisite mechanisms to solve a wide range of theory of mind and perspective taking tasks. However, they apparently have an intact Concern Mechanism.

Appealing to the Concern Mechanism here has some significant advantages over Blair's appeal to VIM. For it is doubtful that humans actually have a VIM. As noted above, recent work in comparative psychology makes it unclear whether *any* species actually has a special purpose mechanism for aggression inhibition. By contrast, the evidence from altruism indicates that Concern Mechanism is present in humans. Further, since the Concern Mechanism is tied to altruistic motivation rather than violence inhibition, the Concern Mechanism account need not be connected to the implausible group selection story associated with VIM. Rather, to determine the evolutionary function of the Concern Mechanism, we can rely on the much more promising work on the evolutionary function of altruistic motivation (e.g., Frank 1988).

On the model of moral judgement that I've suggested, then, there are two quite different mechanisms underlying moral judgement. First, there is a body of information, a normative "theory" that underlies moral judgement, and this theory depends on minimal mindreading capacities, including the capacity to attribute distress. Second, there is a Concern Mechanism, which takes attributions of distress as input and produces as output an affective response that, *inter alia*, motivates altruistic behavior. In order to make the kinds of moral judgements that are explored via the moral/conventional task, both mechanisms must be intact.

My focus in this section has been to give an empirically adequate account of the cognitive mechanisms underlying moral judgement. But it's important to note that the account of moral judgement I've offered, if right, would have significant ramifications for meta-ethics. Although the issues are subtle and complex, one of the abiding concerns in moral philosophy is the extent to which moral judgement depends on desire or emotions (See, e.g., Smith 1995; Darwall et al. 1997). If the account I've developed

here is close to right, it suggests that there is a surprisingly direct relationship between the emotions and a basic form of moral judgement. Individuals who lack certain kinds of emotional responses fail to make certain basic kinds of moral judgements. Hence, if the account I've offered is right, the empirical evidence provides a new argument for the Humean view that moral judgement depends on the passions.

4. Conclusion

At this point, the research on mindreading and moral psychology suggests a strikingly simple model of the core architecture underlying basic moral psychology. The evidence on altruistic motivation and the evidence on moral judgement converge, I've argued, suggesting that there is a single mechanism, the Concern Mechanism, that is implicated in both capacities. Further, I've maintained that altruistic motivation and moral judgement depend on only minimal mindreading capacities. Of course, the account I've offered in this paper is hardly a full account of the cognitive mechanisms implicated in moral psychology. The network of capacities underlying mature moral psychology is no doubt magnificently complex, and I have only tried to sketch one piece of this puzzle. These are early days for the study of moral psychology in philosophy of mind and cognitive science, but the growing body of work gives us every reason to be optimistic that this approach will deeply enrich our understanding of our moral capacities and their origins.

Acknowledgements: I would like to thank Trisha Folds-Bennett, James Blair, Justin D'Arms, Dan Jacobson, Chris Knapp, Jaime Leiser, Ron Mallon, Elizabeth Ronquillo, and Steve Stich for discussion and comments on earlier versions of this paper. This research was supported by NIH grant PHST32MH19975.

References:

American Psychiatric Association 1994. *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed. Washington, D.C.: American Psychiatric Association. (DSM-IV).

Baron-Cohen, S. 1995. *Mindblindness,* Cambridge, MA: MIT Press, Bradford Books.

Baron−Cohen, S., Leslie, A.M., & Frith, U. 1985. "Does the Autistic Child Have a "Theory Of Mind"?" *Cognition*, 21, 37 46.

Bartsch, K. & Wellman, H. 1995. *Children Talk about the Mind,* Oxford University Press.

Batson, C. 1990. "How Social an Animal?: The Human Capacity for Caring". *American Psychologist*, 45, 336-346.

Batson, C. 1991. *The Altruism Question,* Hillsdale, NJ: LEA.

Batson, C., Duncan, B., Ackerman, P., Buckley, T., & Birch, K. 1981. "Is Empathic Emotion a Source of Altruistic Motivation?". *Journal of Personality and Social Psychology*, 40, 290-302.

Batson, C., O'Quin, K., Fultz, J., Vanderplas, M. & Isen, A. 1983. "Self-reported Distress and Empathy and Egoistic versus Altruistic Motivation for Helping". *Journal of Personality and Social Psychology*, 45, 706-718.

Blair R. 1993. *The Development of Morality*. Unpublished Ph.D. thesis, University of London.

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

Blair, R. 1996. "Brief Report: Morality in the Autistic Child". *Journal of Autism and Developmental Disorders*, 26, 571-579.

Blair, R. 1999. "Psychophysiological Responsiveness to the Distress of Others in Children with Autism". *Personality & Individual Differences*, 26, 477-485.

Blair, R., Jones, L., Clark, F., Smith, M. forthcoming. "The Psychopath: A Lack of Responsiveness to Distress Cues?" Ms.

Blum, L. 1994. "Moral Development and Conceptions of Morality". In *Moral Perception and Particularity*, L. Blum. Cambridge University Press.

Byrne, D. 1971. *The Attraction Paradigm,* NY: Academic Press.

Currie, G. 1995. "The Moral Psychology of Fiction". *Australasian Journal of Philosophy*, 73, 250-259.

Currie, G. and Ravenscroft, I. forthcoming. *Recreative Minds: Image and Imagination in Philosophy and Psychology*. Oxford University Press.

Damon, W. 1977. *The Social World of the Child,* San Francisco: Jossey-Bass.

D'Arms, J. and Jacobson, D. forthcoming. "The Moralistic Fallacy: An Equivocation on the 'Appropriateness' of Emotion". *Philosophy and Phenomenological Research*.

Darwall, S. 1998. "Empathy, Sympathy, Care". *Philosophical Studies,* 89, 261-282.

Darwall, S., Gibbard, A., and Railton, P. (eds.) 1997. *Moral Discourse and Practice*. Oxford University Press.

Darwin, C. 1872. *The Expression of the Emotions in Man and Animals,* University of Chicago Press. [Reprinted 1965.]

Dawson, G. & Fernald, M. 1987. "Perspective-Taking Ability and Its Relationship to the Social Behavior of Autistic Children". *Journal of Autism and Developmental Disorders*, 17, 487-498.

De Waal, F. 1996. *Good Natured,* Harvard University Press.

Deigh, J. 1995. "Empathy and Universalizability". *Ethics*, 105, 743-763.

Dunn, J. 1988. *The Beginnings of Social Understanding,* Harvard University Press.

Dunn, J. & Munn, P. 1987. "Development of Justification in Disputes with Mother and Sibling". *Developmental Psychology*, 23, 791-798.

Eisenberg, N. 1992. *The Caring Child,* Harvard.

Eisenberg, N., Fabes, R., Miller, P., Fultz, J., Shell, R., Mathy, R. and Reno, R. 1989. "Relation of Sympathy and Personal Distress to Prosocial Behavior: A Multimethod Study". *Journal of Personality and Social Psychology*, 57, 55-66.

Eisenberg, N. & Fabes, R. 1990. "Empathy: Conceptualization, Measurement, and Relation to Prosocial Behavior". *Motivation and Emotion*, 14, 131-149.

Ekman, P. 1992. "An Argument for Basic Emotions". In N. Stein & K. Oatley (eds.), *Basic Emotions*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Flanagan, O. 1991. *Varieties of Moral Personality: Ethics and Psychological Realism,* Harvard University Press.

Fodor, J. 1983. *Modularity of Mind,* Cambridge, MA: MIT Press, Bradford Books.

Fodor, J. forthcoming. *The Mind Doesn't Work That Way*.

Frank, R. 1988. *Passions within Reason,* Norton.

Frith, U. 1989. *Autism: Explaining the Enigma,* Oxford: Blackwell.

Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. 1995. Taking the intentional stance at 12 months of age. *Cognition*, 56, 165-193.

Goldman, A. 1989. "Interpretation Psychologized". *Mind and Language,* 4, 161-185.

Goldman, A. 1992. "Empathy, Mind, and Morals". *Proceedings and Addresses of the American Philosophical Association, 66*, No. 3, pp. 17-41.

Goldman, A. 1993. "Ethics and Cognitive Science". *Ethics* 103, pp. 337-360.

Gopnik, A. and Wellman, H. 1994. "The Theory-Theory". In L. Hirschfeld and S. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture*. New York: Cambridge University Press. Pp. 257-293.

Gordon, R. 1986. "Folk Psychology as Simulation". *Mind and Language*, 1, 158-171.

Gordon, R. 1995. "Sympathy, Simulation, and the Impartial Spectator". *Ethics* 105, pp. 727-742.

Grice, H. 1975. "Logic and Conversation". In P. Cole and J. Morgan (eds.), *Syntax and Semantics*, vol. 3. New York: Academic Press.

Haidt, J., Koller, S. and Dias, M. 1993. "Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?" *Journal of Personality and Social Psychology*, 65, 613-628.

Harris, P. 1992. "From Simulation to Folk Psychology: The Case for Development". *Mind and Language*, 7, 120-144.

Hermer, L. and Spelke, E. 1996. "Modularity and Development: The Case of Spatial Reorientation". *Cognition*, 61, 195-232.

Hoffman, M. 1976. "Empathy, Role-taking, Guilt and Development of Altruistic Motives". In T. Lickona (ed.), *Moral Development and Behavior: Theory, Research and Social Issues*. New York: Holt, Rinehart & Winston.

Hoffman, M. 1981. "The Development of Empathy". In J. Rushton and R. Sorrentino (eds.), *Altruism and helping behavior: Social, personality, and developmental perspectives* (pp. 41-63). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoffman, M. 1982. "Development of Prosocial Motivation: Empathy and Guilt". In N. Eisenberg, ed., *Development of prosocial behavior*, pp. 281-313. New York: Academic Press.

Hoffman, M. 1991. "Empathy, Social Cognition, and Moral Action". In W. Kurtines and J. Gewirtz (Eds.). *Handbook of Moral Behavior and Development*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hollos, M., Leis, P., & Turiel, E. 1986. "Social Reasoning in Ijo Children and Adolescents in Nigerian Communities". *Journal of Cross-Cultural Psychology, 17*, 352-374.

Kohlberg, L. 1984. *The Psychology of Moral Development: The Nature and Validity of Moral Stages,* Harper & Row.

Krebs, D. 1975. "Empathy and Altruism". *Journal of Personality and Social Psychology,* 32, 1134-1146.

Kunda, Z. 1999. *Social Cognition: Making Sense of People,* Cambridge, MA: MIT Press, Bradford Books.

Kurdek, L. 1980. "Developmental Relations Among Children's Perspective Taking, Moral Judgement, and Parent-Rated Behaviors". *Merrill-Palmer Quarterly*, 26, 103-121.

Latane, B. & Darley, J. 1968. "Group inhibition of bystander intervention in emergencies". *Journal of Personality and Social Psychology*, 10, 215–221.

Leslie, A. 1994. "ToMM, ToBY and Agency: Core Architecture and Domain Specificity". In L. Hirschfeld & S. Gelman (eds.) *Mapping the mind*. Cambridge University Press, 119-148.

Lorenz, K. 1966. *On Aggression,* New York: Harcourt, Brace, Jovanovich.

Miller, J., Bersoff, D. Harwood, L. 1990. "Perceptions of social responsibilities in India and the United States: Moral imperatives or personal decisions?" *Journal of Personality and Social Psychology*, 58, 33-47.

Miller, R., Banks, J., & Ogawa, N. 1963. "Role of Facial Expression in 'Cooperative-avoidance Conditioning' in Monkeys". *Journal of Abnormal and Social Psychology*, 67, 24-30.

Miller, P., Eisenberg, N., Fabes, R., & Shell, R. 1996. "Relations of Moral Reasoning and Vicarious

Emotion to Young Children's Prosocial Behavior Toward Peers and Adults". *Developmental Psychology*, 32, 210-219.

Newcombe, T. 1961. *The Acquaintance Process,* New York: Holt, Rinehart & Winston.

Nichols, S. MS. Is It Irrational to Be Amoral? How Psychopaths Threaten Moral Rationalism.

Nichols, S., Stich, S., Leslie, A., and Klein, D. 1996. "Varieties of Off-Line Simulation". In P. Carruthers & P. Smith (eds.) *Theories of Theories of Mind*. Cambridge University Press.

Nichols, S. and Stich, S. 2000. "A Cognitive Theory of Pretense". *Cognition,* 74, 115-47.

Nichols, S. and Stich, S. forthcoming. *Mindreading,* Oxford: Oxford University Press.

Nucci, L., Turiel, E., & Encarnacion-Gawrych, G. 1983. "Children's Social Interactions and Social Concepts: Analyses Of Morality and Convention in The Virgin Islands". *Journal of Cross-Cultural Psychology, 14,* 469-487.

Nucci, L. 1986. "Children's Conceptions of Morality, Social Conventions and Religious Prescription". In C. Harding, (ed.), *Moral Dilemmas: Philosophical and Psychological Reconsiderations of the Development of Moral Reasoning*. Chicago: Precedent Press.

Piaget, J. 1932. *The Psychology of Moral Development: The Nature and Validity of Moral Stages,* M. Gabain, trans. New York: Free Press. (Translation published 1965.)

Premack, D. & Woodruff, G. 1978. "Does the Chimpanzee Have a Theory of Mind?" *Behavioral and Brain Sciences*, 1, 516-526.

Radke-Yarrow, M., Zahn-Waxler, C., & Chapman, M. 1983. "Children's Prosocial Dispositions and Behavior". In P. Mussen (ed.), *Handbook of Child Psychology, Vol. 4, Socialization, Personality, and Social Development*. New York: Wiley.

Rawls, J. 1971. *A Theory of Justice,* Harvard University Press.

Roberts, W. & Strayer, J. 1996. "Empathy, Emotional Expressiveness, and Prosocial Behavior". *Child Development*, 67, 449-470.

Rosenbaum, M. 1986. "The Repulsion Hypothesis: On the Nondevelopment of Relationships". *Journal of Personality and Social Psychology 51*, 1156-1166.

Scholl, B. and Leslie, A. 1999. "Modularity, Development, and 'Theory of Mind'". *Mind and Language*, 14, 131–153.

Selman, R. 1980. *The Growth of Interpersonal Understanding,* New York: Academic Press.

Sigman, M., Kasari, C., Kwon, J., & Yirmiya, N. 1992. "Responses to the Negative Emotions of Others by Autistic, Mentally Retarded, and Normal Children". *Child Development*, 63, 796-807.

Simner, M. 1971. "Newborn's Response to the Cry of Another Infant". *Developmental Psychology 5*, 136-150.

Smetana, J. 1985. "Preschool children's Conceptions of Transgressions: Effects of Varying Moral and Conventional Domain-related Attributes". *Developmental Psychology*, 21, 18-29.

Smetana, J. 1993. "Understanding of Social Rules". In M. Bennett (ed.) *The Development of Social Cognition : The Child as Psychologist*. New York: Guilford Press, 111-141.

Smetana, J. & Braeges, J. 1990. "The Development of Toddlers' Moral and Conventional Judgements". *Merrill-Palmer Quarterly*, 36, 329-346.

Smith, M. 1995. *The Moral Problem,* Oxford: Blackwell.

Sober, E. and Wilson, D. 1998. *Unto Others,* Harvard University Press.

Song, M., Smetana, J., Kim, S. 1987. "Korean Children's Conceptions of Moral and Conventional Transgressions". *Developmental Psychology*, 23, 577-582.

Stich, S. and Nichols, S. 1992. "Folk Psychology: Simulation or Tacit Theory". *Mind & Language*, v. 7, no. 1, 35-71.

Tager-Flusberg, H. 1993. "What Language Reveals about the Understanding of Minds in Children with Autism". In S. Baron-Cohen, H. Tager-Flusberg & Donald Cohen (eds.) *Understanding Other Minds: Perspectives from Autism*, 138-157.

Tan, J., and Harris, P. 1991. "Autistic Children Understand Seeing and Wanting". *Development and Psychopathology*, 3, 163-174.

Tajfel, H. 1981. *Human Groups and Social Categories,* Cambridge: Cambridge University Press.

Tomasello, M. and Call, J. 1997. *Primate Cognition,* Oxford University Press.

Turiel, E. 1983. *The Development of Social Knowledge: Morality and Convention,* Cambridge: Cambridge University Press.

Turiel, E., Killen, M., & Helwig, C. 1987. "Morality: Its Structure, Functions, and Vagaries". In J. Kagan & S. Lamb (eds.) *The Emergence of Morality in Young Children*. Chicago: University of Chicago Press, 155-244.

Uller, C. and Nichols, S. forthcoming. "Goal Attribution in Chimpanzees". *Cognition*.

Williams, G. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought,* Princeton: Princeton University Press.

Wing, L., and Gould, J. 1979. "Severe Impairments of Social Interaction and Associated Abnormalities in Children: Epidemiology and Classification". *Journal of Autism and Developmental Disorders,*

9, 11-29.

Woodward, A. 1998. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69, 1-34.

Yirmiya, N., Sigman, M., Kasari, C., & Mundy, P. 1992. "Empathy and Cognition in High-Functioning Children with Autism". *Child Development, 63*, 150-160.

Zahn-Waxler, C., & Radke-Yarrow, M. 1982. "The Development of Altruism: Alternative Research Strategies". In N. Eisenberg-Berg (ed.), *The Development of Prosocial Behavior*. New York: Academic Press.

Zahn-Waxler, C., Radke-Yarrow, M., & Brady-Smith, J. 1977. "Perspective-taking and Prosocial Behavior". *Developmental Psychology*, 12, 87-88.

Zahn-Waxler, C., Radke-Yarrow, M., & King, R. 1979. "Child Rearing and Children's Prosocial Initiations toward Victims of Distress". *Child Development*, 50, 319-330.

---

[1] Of course, some might claim that this begs the question since there may be no genuine cases of altruism. For instance, Sober & Wilson (1998) suggest that 'altruistic' desires driven by empathy or sympathy might not be truly altruistic since the desires may be instrumental rather than "ultimate": "Perhaps empathy and sympathy are able to evoke altruistic desires because people don't like experiencing these emotions and therefore wish to do what they can to extinguish them" (232). However, since the present goal is to outline actual cognitive architecture, we must look at actual cases of human motivation, and if no actual cases meet with Sober & Wilson's notion of altruism, that just means that Sober & Wilson's notion is not useful for charting the cognitive architecture.

[2]Actually, Batson addresses a somewhat broader category, the "aversive-arousal reduction" model of altruism, according to which "becoming empathically aroused by witnessing someone in need is aversive and evokes motivation to reduce this aversive arousal" (Batson 1991, 109). Emotional contagion provides perhaps the most obvious mechanism for producing aversive arousal, and the emotional contagion model outlined here is a version of the aversive-arousal model.

[3]Although Batson's work has been widely known and discussed in psychology for over a decade, philosophers have only recently come to appreciate its significance (Darwall 1998; Sober & Wilson 1998).

[4]Darwall (1998) uses Batson's evidence to shore up the view that sympathy presents a categorical justification for preventing another's woe. Sympathy, according to Darwall, connects us to 'person-neutral value':

> sympathetic concern presents itself as of, not just some harm or disvalue *to* another person, but also the *neutral disvalue* of this personal harm owing to the value of the person himself. In feeling sympathy for the child, we perceive the impending disaster as not just terrible for him, but as neutrally bad in a way that gives anyone a reason to prevent it. We experience the child's plight as mattering categorically because we experience the child as mattering... sympathy's emotional presentation is *as of* the neutral disvalue of another's woe, and hence, as of a categorical justification for preventing it (p. 275).

However, whether sympathy presents itself this way is not shown by any of Batson's data. What the social psychological evidence (and commonsense) suggests is that our emotional response to a child's plight prompts us to help the child rather than flee. So *the person in need* does play a crucial role here. But one would need quite different evidence to show that people think that a child's plight gives *anyone* reason to help the child. It's important to distinguish here between what people *expect* others to do and what people think is *categorically* justified. It's probably true that most people *expect* others to be moved by the plight of children (as we tend to expect others to share many of our attitudes [Nichols & Stich forthcoming]); however, it's not clear that most people think that a child's suffering matters categorically. As far as I know, there is simply no evidence on this issue, and it's quite possible that there is considerable variance on the issue across individuals and across cultures.

[5]I'm focusing on distress, but this is merely for ease of exposition. I don't mean to exclude the possibility that representations of other negative affective states (e.g., grief, fear, sorrow) will produce altruistic motivation.

[6] Of course, like the perspective-taking account, this is only a partial account of altruistic motivation, since it doesn't explain the process that goes from mindreading to motivation. As will be discussed below (2.7), on both the perspective-taking account and the minimalist account, a natural assumption is that the representations generated by mindreading produce an affective response that produces the motivation.

[7]Of course, one might deny that toddler comforting behaviors count as core cases of altruism. Rather, one might claim that such cases should be construed as ersatz altruism. However, one would need an argument for excluding these cases. For if we focus on the underlying motivation, the evidence suggests that altruistic concern in toddlers is continuous with altruistic concern in later childhood and adulthood (e.g., Zahn-Waxler et al. 1992; Eisenberg & Fabes 1990; Eisenberg et al. 1989).

[8]6 out of 29 autistic children helped; 7 out of 30 mentally retarded children helped; and 3 out of 30 normally developing children helped (Sigman et al. 1992, 800).

[9]As we saw in section 1.2, Sober & Wilson (1998, pp. 234-5) maintain that sympathy doesn't require that the sympathizer and the target feel the same emotion simultaneously. But that doesn't really distinguish sympathy from sophisticated accounts of empathy. The psychological work, however, really does raise the possibility of a profound distinction. Feelings of sympathy may not parallel *any* other feeling.

[10]Another possibility is that affect plays no role in altruistic motivation. Rather, perhaps altruistic motivation follows directly from an attribution of distress. Something like this might be Sober & Wilson's view (1998, pp. 312 ff.). They suggest that evolution built a mechanism for altruistic motivation that does not rely on hedonic or affective states. However, they do not explain how that

mechanism might have evolved in the existing motivational systems of our ancestors. The standard models of motivation in psychology are 'hot' models, on which affect plays the central role in basic motivation (see, e.g., Kunda 1999). To make a 'tepid' model of altruistic motivation plausible would require a broader defense of the idea of tepid basic motivation. Furthermore, there is some evidence suggesting important correlations between affect and altruistic behavior. As we've seen, the developmental data suggest a correlation between affective response and helping behavior in children, and the social psychological data suggest a similar correlation in adults. In addition, the evidence on psychopaths indicates that their lack of helping behavior might be correlated with a deficit to the capacity for responding to others' distress.

[11]Blair writes that the activation of VIM produces an aversive experience and that it is "this sense of aversion to the moral transgression" that results in the act being "judged as bad" (1995, p. 7; see also 1993, p. 83).

[12] There are arguments about perspective taking and moral judgement (broadly construed) in the older developmental literature. This tradition developed detailed models and collected extensive data on perspective taking and moral reasoning (e.g., Piaget 1932, Kohlberg 1984, Selman 1980, Damon 1977; for a useful review of this tradition, see Flanagan 1991). But the arguments are for the somewhat different claim that as perspective taking abilities improve, so do moral reasoning abilities. That does not, of course, show that perspective-taking abilities are required for drawing the moral/conventional distinction. In addition, even in this tradition it is not clear that perspective taking drives moral reasoning capacities. For example, Kurdek (1980) was surprised when his results suggested the opposite: "The present results indicate that the preponderance of causation was in the direction of children's early moral judgement abilities operating as a cause of their latter cognitive perspective taking ability.... the preponderance of causality seemed to be in the direction of moral judgement causing later developments in cognitive perspective taking" (1980, p. 116, 118).

[13]Although psychopaths failed to draw the moral/conventional distinction, they did so in a surprising way. Contrary to Blair's predictions, psychopaths rated both moral and conventional transgressions as impermissible, serious and not dependent on authority. Thus, it may seem that psychopaths regard all transgressions as moral. However, Blair notes, with some plausibility, that the important feature is that the justifications offered for why the moral transgressions are wrong were consonant with conventional-type justifications (Blair 1995, 24). According to Blair, whose subjects were British, a typical justification was "it's not the done thing" (personal communication).

[14] It's important to note that I am not making a *conceptual* claim here. A great deal of work in ethics focuses on the idea that it's conceptually possible that someone might have mastery of moral concepts without having any concomitant motivation to act morally or any concern for others. I think this is perfectly right and of some significance for evaluating certain moral rationalist claims (Nichols MS). However, the issue in the present paper is the empirical question – what mechanisms are *in fact* involved in moral judgement. My claim is that the Concern Mechanism is in fact implicated in normal moral judgement. The fact that it's conceptually possible for an amoralist to be fluent with moral discourse (or to be trained to be so) is a separate issue.