Like-Minded

Externalism and Moral Psychology

Andrew Sneddon

The MIT Press
Cambridge, Massachusetts
London, England

Contents

Preface ix

1 Introduction: Externalism and Moral Psychology 1

2 The Disunity of Moral Judgment 25

3 Moral Reasoning 71

4 Rethinking the Reactive Attitudes: Attributing Moral Responsibility 111

5 The Production of Action 157

6 Psychological Pluralism, Environmental Sensitivity, and the Bounds of Morality 203

Notes 251
References 263
Index 279

1 Introduction: Externalism and Moral Psychology

1.1 Reasons, Passions, and a Third Option

What are the psychological foundations of morality? What psychological capacities enable us to evaluate actions? To act in accordance with moral norms? To attribute moral responsibility to ourselves and others? Although these have been perennial concerns for philosophers, there has been a flurry of work on them in recent years in a distinctly interdisciplinary vein. Philosophers and psychologists have combined resources to address these questions. The results include both new formulations of familiar positions and genuinely new answers. This book contributes to this interdisciplinary trend.

Historically the chief question for philosophers has been whether the psychological foundations of morality are emotional or rational. The classical protagonists in this debate are well known: David Hume (1740) argued that reason is the slave of the passions, so morality must be based on them, whereas Immanuel Kant (1785) argued that moral law is given by rational agents to themselves in virtue of their rationality. This debate continued through the development of analytic meta-ethics in the twentieth century, and it continues today. Simon Blackburn (1998) is a prominent intellectual descendant of Hume, while Michael Smith (1994, 2004) is arguably the most prominent present-day rationalist. Empirical data have been brought into this debate. For example, Shaun Nichols (2002, 2004a) has argued that empirical studies of psychopathy support a Humean view of morality rather than a Kantian one. Jeannette Kennett (2006) has recently defended moral rationalism from this charge on empirical grounds.

My primary aim is to make a third option plausible. The words 'reason' and 'passion' do not satisfactorily capture all of the important options for

explaining the psychological foundations of morality. A third possibility is, to put it roughly, that these foundations centrally include capacities that enable us to operate within cognitive systems that extend beyond individual agents into the wider world. I call this the *Wide Moral Systems Hypothesis*. This hypothesis fits within the array of positions known as *externalism* about the mind or, rather more catchily, the *Extended Mind Hypothesis*. According to the Extended Mind Hypothesis, at least some cognitive processes extend beyond the individual agent to include worldly resources. These resources are not merely input to cognitive processes that are located within an individual's brain. Rather, they partially constitute the cognitive processes in question. The conventional terminology is to call processes that are partly constituted by environmental resources "wide"—hence the name of the general hypothesis defended in this book. Processes that are located solely within the bounds of agents' bodies are "narrow."

Each of the next four chapters presents specific wide hypotheses about a distinct aspect of our moral psychology. The first topic is moral judgment. This is a traditionally central topic in examinations of moral agency. However, we must be careful with this term. For one thing, it is easy to assume that "judgment" must be the product of a process of judging, and that this in turn is something done consciously by an agent, perhaps analogously to what is done by a legal judge in a courtroom. I wish to avoid these assumptions. For another thing, confusion about this term has arisen. Following Jonathan Haidt (2001), Marc Hauser (2006), and Jesse Prinz (2006a), I use "moral judgment" here to refer to the psychological capacity or capacities by which we evaluate things actions, states of affairs, and persons in moral terms, however this is accomplished. Some people—e.g., Jorge Moll et al. (2005)—define moral judgment such that it is automatically the product of moral reasoning. Doing so eliminates any substantial inquiry into whether the foundation of moral judgment is moral reasoning. I follow many psychologists and philosophers in taking this as a substantial issue, one to be decided through conceptual and empirical inquiry rather than by definition. Accordingly, I do not define moral judgment as the product of moral reasoning.

Besides moral judgment, I will be examining moral reasoning, the production of action, and attributions of moral responsibility. Despite the

judgment-centric approach of most discussions of moral psychology, I am inclined to think that these phenomena are just as central to moral agency as moral judgment and worth just as much attention. Regardless of questions of priority, a good case can be made for thinking that the items on this list come close to exhausting the range of our central moral-psychological capacities. In chapter 6, I look around for topics to add to our view of moral psychology. Lots of work has been done on these topics in both philosophy and psychology, some of it now well known and considered classic. I am going to revisit this work with an eye on what I take to be its overlooked externalist aspects. Again and again I have been struck by the integration of agent and environment either described or hinted at by research on moral judgment, moral reasoning, moral motivation, and moral responsibility. There is a story both familiar and novel here, and I intend to tell it as best I can.

There is work to be done before I turn to moral psychology. In this chapter, I provide some conceptual tools for thinking about cognitive systems that extend beyond the physical boundaries of individual agents. This will illuminate the conceptual possibility of such systems. The bulk of the book will be concerned with establishing the empirical plausibility of the Wide Moral Systems Hypothesis.

1.2 The Extended Mind Hypothesis: Varieties of Individualism and Externalism

A closer look at the individualism/externalism debate is required in order to see just how the Wide Moral Systems Hypothesis is an alternative to the traditional options. Arguably, the reason/passion debate in moral psychology has been about psychological capacities that are attributed to individuals. This is the default approach in both empirical and philosophical psychology: questions are framed in terms of capacities attributed to individuals. However, over since about 1970 philosophical psychology has been marked by the sustained challenge to this approach that has come to be known as the Extended Mind Hypothesis. Defenders of this hypothesis are typically known as "externalists"; those who deny it are "individualists." In very general terms, the debate between individualists and externalists is about how to understand the role of an agent's context in

the agent's psychological functioning. Individualists restrict context to the psychological background. It is a source of input to our psychological processes and it receives output from them. Externalists do not deny that context performs these functions. However, externalists claim that contextual features can also be parts of psychological processes. To refine our sense of the issues, here are some central ways in which externalist theses have been refined and developed.

First Distinction: Content Externalism and Vehicle Externalism

Today's philosophical debate about externalism has its roots in philosophy of mind and language. The seminal thought experiments of Hilary Putnam (1975) and Tyler Burge (1979), with their emphasis on the meaning of utterances and the content of such folk psychological states as beliefs, exemplify this approach. In an assessment of the debates arising from these thought experiments, Mark Rowlands (2003) distinguishes *content* externalism from *vehicle* externalism.

Let's begin with content. The individualist about content holds that the meanings of utterances and mental states are logically independent of environment. The externalist denies this. Putnam and Burge's now-classic thought experiments work by probing our intuitions about what happens when we hold the intrinsic properties of individuals constant and vary their environments. Their claim is that what is revealed by such arguments is that mental and linguistic content turns out to vary with environmental variances despite the constancy of the agents' intrinsic properties.

Content externalism implies nothing about the nature of the items that bear content. Putnam, Burge, and subsequent individualists and externalists about content have not been primarily concerned about the nature of our cognitive architecture. It is perfectly consistent with thoroughgoing content externalism to hold that the bearers of psychological content are intrinsic features of agents. Debate about vehicle externalism calls this directly into question. Physicalist individualists about this issue hold that the vehicles of content are located within the physical bounds of individual agents. Vehicle externalists deny this. Since this issue is distinct from that of content externalism, new arguments are needed to assess the plausibility of these positions. And since the vehicle issue seems to be at least in part about the mechanics of psychological processes, empirical information is more relevant here than it is to content externalism.

Second Distinction: Taxonomic and Locational Externalism

The importance of empirical study to the assessment of vehicle externalism introduces a second way of classifying varieties of externalism. This is because the other way to approach this territory has been from the perspective of philosophical psychology and philosophy of science more generally. Robert Wilson (2003, 2004) takes this approach and distinguishes *locational* externalism from *taxonomic* externalism.

One way Wilson draws the locational/taxonomic distinction is by examining the metaphysics of the relation "realization." Following Sydney Shoemaker, Wilson makes consideration of systems the starting point of his case. For a higher-level property H and a system S in which it is realized, the *core* realization is "a state of the specific part of S that is most readily identifiable as playing a crucial causal role in producing or sustaining H" (Wilson 2001, 8). The *total* realization of H is "a state of S, containing any given core realization as a proper part, that is metaphysically sufficient for H" (ibid., 8). When a system is contained within an individual, an individualistic interpretation of the properties of that system is warranted. However, Wilson draws our attention to the possibility of systems that include individuals as a part and hence extend beyond the physical boundaries of individual agents. Using this possibility, Wilson identifies two sorts of externalist realization. "Wide" realization occurs when there is "a total realization of H whose non-core part is not located entirely within B, the individual who has H" (ibid., 11). "Radically wide" realization involves "a wide realization whose core part is not located entirely within B, the individual who has H" (ibid., 13).

When a property of an individual is widely realized, it must be individuated in reference to the system that extends beyond the boundaries of the individual. This yields a position about taxonomy: taxonomic externalism (Wilson 2003, 276; 2004, 174–178). In contrast, properties with radically wide realizations are not solely properties of the individual one is examining, but are instead located at least partly beyond its physical boundaries. The associated view of externalism is, accordingly, *locational* externalism (Wilson 2003, 276; 2004, 174–178).

These characterizations of externalism apply not only to psychology but to any phenomenon to which the metaphysics of realization of properties by systems applies. Wilson (2004, 114–115) draws examples from biology. The biological property of *being a predator* is one that is properly attributed

to individual organisms, but one that they have by virtue of their role in a predator-prey system. Accordingly, biologists should be taxonomically but not locationally externalist about predators. However, Wilson's central examples of locational externalism come from psychology. There are research programs in cognitive science that describe cognitive tasks as being accomplished between individuals, or via individual-environment interaction. Wilson discusses Edwin Hutchins's work on how navigational tasks are performed (1995) and Rodney Brooks's work designing robots (1991) as examples of such research programs. The contention is that the cognitive processes in question are located partially beyond the physical boundaries of the individuals participating in the systems, so we should be locationally externalist about them.

Let's return to the content/vehicle distinction. Taxonomic externalism is equivalent to content externalism only if principled scientific psychological taxonomy is done only in terms of the content of psychological states. Taxonomy by content is undeniably important. However, whether it is the only way scientific psychology can characterize the elements of its domain seems to be an open question. As a broad possibility, perhaps some items in psychological explanations ought to be individuated in functional terms, i.e., in terms of their relationships to input and output. If this is the case—and whether it is seems to be an empirical issue—then taxonomic externalism is distinct from content externalism.

In contrast, locational externalism is equivalent to vehicle externalism only if the only way in which the bearer of content can extend beyond the intrinsic boundaries of an individual is for it to be *realized* by a *system* of which the relevant individual is a part. If realization is not the only relation relevant to the nature of bearers of content, or if there are principled ways of addressing bearers of content independent of their possible or actual roles in systems, then locational externalism is distinct from vehicle externalism. Again, these appear to be empirical issues.

Putting these nuances aside, it is reasonable to see the content/vehicle and taxonomic/locational distinctions as deeply related. That said, Wilson's turn from the traditional concerns of philosophy of mind and language to empirically informed metaphysics is both a genuine step forward in the development of externalism and a minor obstacle to a clear view of the possible implications of externalism. It is a step forward in that it both acknowledges and makes clear the connection

between this issue and empirical work in various sciences, particularly psychology. It obscures matters because of its emphasis on the metaphysics of realization, which is at least one step removed from the practical concerns of practicing psychologists and opaque with regard to its relevance to these concerns.

The focus on the practice of psychology brings us to a distinction that, although implicit in content/vehicle distinctions and (especially) in taxonomic/locational distinctions, has gone undeveloped. For psychological externalism to be empirically assessed, psychological hypotheses must be framed in terms that are relevant to differences between individualistic and externalist interpretations of psychological phenomena. The most straightforward way for this to happen is for the hypotheses themselves to be framed in explicitly externalist terms. How are we to know what topics call for externalist hypotheses? I have no thoroughgoing answer to this question, but Wilson's work suggests a starting point. Externalist hypotheses are warranted for any psychological phenomenon that exhibits *systematic* individual-environment relations. I am inclined to think that the question of when an individual-environment system is present is one that must be answered *a posteriori*, and that particular sciences may justifiably have differing working notions of conditions that must be satisfied for the presence of a system. Nevertheless, Wilson's work provides us with a rough notion of what I shall call "systemicity" that can be used as a rule of thumb. (Note well: The following is not offered as an analysis of "system" in necessary and sufficient conditions.)

In refining the concepts central to Developmental Systems Theory, Wilson (2005, 153) characterizes developmental systems as follows: "Developmental systems must be causally and functionally integrated chains of developmental resources, and these, individually and collectively, must play a replicable causal role in ontogeny and inheritance." If we strip this of content peculiar to developmental systems, we have a general schema for systemicity:

_____ systems must be causally and functionally integrated chains of _____ resources, and these, individually and collectively, must play a replicable causal role in _____.

At present we are interested in particular kinds of psychological systems, so the resources in question must be *cognitive* ones, broadly understood as

informational input and output domains and mechanisms. To justify the description of something as a psychological system, these resources must play a replicable causal role in the production and the execution of particular psychological phenomena. To instantiate a *moral*-psychological system, there must be the appropriate sorts of cognitive resources connected in the appropriate ways to produce and sustain particular aspects of moral cognition.

The possibility of systems that are distributed between an individual and that individual's environment—i.e., of *wide* systems—is delivered by the possibility of the requisite causal and functional integration. The higher the degree of causal and functional integration there is between an individual and aspects of the individual's environment, the greater the reason there is to think that the individual and those aspects of the environment constitute a system. In the chapters that follow, our attention will be on exactly this sort of individual-environment integration with regard to our central moral-psychological capacities.

Using the notions of systems and systematic individual-environment interaction as our starting point, we can distinguish two forms that externalist hypotheses can take:

A psychological hypothesis is *shallowly* externalist when it begins with psychological items attributed to an individual regardless of environmental integration and construes them widely.

A psychological hypothesis is *deeply* externalist when it begins with systematic individual-environment interaction and attributes psychological items to the individual as needed to participate in the given wide system.[1]

The difference between shallow and deep externalism is one of initial presuppositions. Shallow externalist hypotheses are to be expected when they are framed as reinterpretations of individualistic hypotheses. In these cases, it is reasonable to interpret what one encounters as a relatively superficial modification of one's understanding of something with an individualistic basis. This is one way of understanding debates about content externalism: propositional attitudes are attributed to individuals, which individualists had construed as logically independent of context but which externalists reconstrue as logically dependent on certain contextual features. In both cases, exactly the same psychological items are attributed to

individuals. In contrast, deep externalist hypotheses are framed from an externalist starting point, rather than as a reinterpretation of previously individualistic ideas.

The traditional reason/passion debate is reasonably interpreted as consisting in the examination of either individualistic or shallowly externalist hypotheses. Reason and passion are understood as psychological items that can be attributed to individuals regardless of context. In contrast, the Wide Moral Systems Hypothesis (WMSH) is a deeply externalist position. My primary aim is to make plausible the idea that the psychological foundations of morality should be understood, at least partly, in terms of cognitive systems that extend into the environment beyond the physical bounds of individual agents. Psychological items will be attributed to individuals on the basis of such systematic agent-environment interaction, where it is found. In short, context is treated as integral to our central moral-psychological capacities in the WMSH.

Although the particular nature of the psychological items attributed to individuals in deeply externalist hypotheses must depend on the details of the case, two things can be said in general. First, these items are for the individual to participate in the wide system; they are not for replicating whatever psychological functions the wide system performs.[2] The idea is that some psychological job P is performed once by the wide system, not twice (once by the wide system *and* once by narrow systems that an individual happens also to have). Sometimes we will encounter cognitive redundancy, where P can be performed, and may even actually be performed, by both wide and narrow systems. However, there is no *a priori* reason to require individuals to narrowly perform P while participating in wide systems that also perform P. In fact, if this were taken to be a point about systems generally, there would be an *a priori* reason against it, as it would generate an infinite regress: for a system to perform P, there would have to be some other system to do so. For this second system to perform P, there would have to be a third system, *ad infinitum*. Second, the psychological items attributed to an individual to participate in a wide cognitive system need not themselves be wide in every sense. To be precise, they need not be locationally wide. They will be taxonomically wide, insofar as they have to be classified in terms of the wide system in which they play a role. But they themselves can, quite comfortably, be located within the

physical boundaries of individual agents. There is good reason to think that locationally narrow but taxonomically wide psychological capacities provide important underpinnings for our moral psychology.

1.3 What about Twin Earth?

Let us briefly return to content externalism and the famous arguments of Putnam and Burge. They employed a specific sort of argument that many will associate with discussion of externalism in general. This type of argument asks us to compare pairs of linguistic contexts. Crucially, we are to compare people in these contexts whose intrinsic, individualistic properties are identical. The contexts themselves differ in some specific way. For instance, Burge presents two people who are individualistically identical. Their contexts vary with regard to the meaning of the word 'arthritis'. In one context, the word applies only to certain disorders of joints; in the other, it also applies to disorders in body parts other than joints. When a person in the first context complains of arthritis in his thigh, he speaks incorrectly. When an identical person in the second context makes the same complaint, he is truly speaking of arthritis in his thigh (Burge 1979, 77–79). Putnam asks us to compare Earth and "Twin Earth," in the process giving rise to the tradition of referring to such arguments as Twin-Earth arguments. On Earth, water is H_2O, but on Twin Earth it has a different chemical constitution, which Putnam calls XYZ. When Oscar_1 on Earth uses the word 'water', he means H_2O even if he has no idea what elements constitute water. When Oscar_2—who is intrinsically identical to Oscar_1—uses the word 'water' on Twin Earth, he means XYZ even if he has no idea what the constituents of water are (Putnam 1975, 139–141). For both Burge and Putnam, the point is that mental and linguistic content are at least partly determined by the contexts in which agents find themselves. Agents' individualistically construed properties do not suffice to determine the content of their thoughts and their utterances.

Twin-Earth arguments are so closely tied to debates about externalism and individualism that some readers will expect to find some in this book. You will not find any. There are two reasons for this. The first is that, as we have just seen, Twin-Earth arguments concern, first and foremost, questions of content. Content is not the topic of this book, so the Twin-Earth considerations of Putnam and Burge find no natural application here.

Some may find this lame. Surely the construction of Twin-Earth arguments for topics other than content is not impossible. It takes patience, imagination, and hard work, not magic, meaning that their omission is suspicious. Let us admit for the purposes of argument that Twin-Earth arguments can be constructed for topics other than content. There is a second and more important reason for omitting this way of arguing in the present book. I suspect that Twin-Earth arguments are useful for developing and evaluating shallow externalism. However, I am impressed by the more radical possibilities offered by externalism.[3] This calls for deeply externalistic hypotheses, but Twin-Earth arguments are not useful tools for devising such hypotheses. Twin-Earth arguments work by holding some pre-specified feature or features of agents constant and varying the contexts in which the agents function. This invites a conservative approach to the description of agents. The reason is that there is a tendency toward individualism in both folk and scientific psychology. This bias results in the description of agents in terms that are usually used individualistically. In Twin-Earth arguments these descriptions are subsequently re-imagined widely. This is shallow externalism. If I am correct that the theoretical possibilities of externalism run deeper, then we will do well to avoid ways of arguing that invite shallowly externalistic hypotheses. Consequently there will be no more visits to Twin Earth in this book.

1.4 Objections to Psychological Externalism

Of course, there have been objections to the Extended Mind Hypothesis. The positions of Putnam and Burge have long been resisted. Indeed, Burge's seminal 1979 paper is largely constructed around potential objections to the very idea of externalism about mental content. In his 1995 book, Wilson scrutinizes two decades of work favoring individualism and finds it wanting. The more recent varieties of externalism have met equally persistent opposition. Frederick Adams and Kenneth Aizawa (2001, 2008) and Robert Rupert (2004, 2009) offer significant challenges. Important responses can be found in Clark 2008 and in Wilson and Clark 2008. I shall not delve into the details of the discussion about the objections made by Rupert and by Adams and Aizawa, since I think that the decisive responses have already been made. A more recent objection that has not

yet been adequately answered is that of Mark Sprevak (2009). Sprevak's case deserves some attention before we turn to empirical issues.

Sprevak's argument has two ideas at its core: the Parity Principle and the Martian Intuition. The Parity Principle, which Sprevak calls the "fair treatment" principle (2009, 505), comes from the famous formulation of externalism by Andy Clark and David Chalmers. The Parity Principle is designed to focus attention on our judgments of what systems, states, and processes are cognitive and to divert attention from putatively misleading side issues such as whether a system is located solely within the physical bounds of an organism:

> *The Parity Principle* If, as we confront some task, a part of the world functions as a process which, were it to go in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process. (Clark and Chalmers 1998, 8)

It should be clear that the specific issue is locational externalism. For present purposes, the Parity Principle provides a partial rule of thumb for deciding whether a process is cognitive (or mental—Clark and Chalmers address both). Whether it is the only idea relevant to such decisions is one of the things Sprevak examines with his argument.

The Martian Intuition is that it is conceivable for creatures with mental states to exist even if they are physically and biologically different from humans (Sprevak 2009, 507). This idea has long had a role in philosophy of mind: it has been a famous part of arguments leading from identity theories to functionalism (ibid., 509). Blood, skin, hearts, and the like are, *prima facie*, inessential to mentality. So perhaps are brains, spines, and nerves. According to functionalist deployments of the Martian Intuition, the important thing is what these substances do, not the substances themselves. This implies that creatures made of silicone or mud or tin cans (use your imagination) could have minds if these substances do the same things that neurons do for us. Such non-humans with minds are the "Martians" in question.

The Martian Intuition has three roles in Sprevak's argument. First, following Clark, Sprevak uses it to answer objections to locational externalism. Rupert (2004) and Adams and Aizawa (2008) object to externalism by appealing to fairly fine-grained features of putatively brain-bound or body-bound human cognition that extended processes do not share. Sprevak claims that arguments of this sort violate the Martian Intuition: it is

conceivable for creatures to have minds yet not to share the fine-grained features of human thought offered by the critics of externalism. Hence the objections turn on inessential features of mentality and are unduly chauvinistic.

The second role of the Martian Intuition in Sprevak's case is as the core of an argument for locational externalism from functionalism. Functionalism offers the functional organization of systems as the essential feature of mentality. The functional role of putatively mental phenomena must be specified to account for the nature of the phenomena. However, such specification can be done in myriad ways. Parameters must be provided to constrain such descriptions such that they provide all and only the relevant information. The Martian Intuition provides one of these parameters: functional roles must be specified in a sufficiently coarse-grained manner in order to allow for the possibility of minded creatures whose minds are realized in ways or substances different from the ways and substances that realize human minds. Sprevak argues that if the "grain parameter" is set at least coarse enough to allow for the possibility of Martian minds, then it will also allow for extended cognition in humans. The reason is that these cases of extended cognition will be as similar to brain-bound or body-bound human cognition as the Martian cognitive processes are. If we combine the Martian Intuition with the Parity Principle, then allowing for Martian minds but not extended human minds is unduly chauvinistic. Thus, if functionalism preserves the Martian Intuition, it also implies locational externalism.

Sprevak focuses on the version of externalism offered by Clark and Chalmers (1998). Besides functionalism, Clark and Chalmers specify three conditions that they think human-world processes must meet in order to count as cognitive. But, as with the objections of Rupert and those of Adams and Aizawa, Sprevak wields the Martian Intuition: these conditions are too fine-grained because we can imagine creatures with minds who do not share them. When combined with the Parity Principle, the implication of the cases generated by the Martian Intuition is that analogous cases which happen to extend into the world should count as cognitive. Hence, the constrained externalism of Clark and Chalmers is, again, unduly chauvinistic. Functionalism delivers *radical* externalism instead: *any* human interaction with a worldly resource suitable for use in a cognitive system constitutes an extended cognitive system. This is "radical" in that the constitution of

extended cognitive systems is unconstrained, and hence it is very easy for human-world systems to count as cognitive. Clark and Chalmers's constrained externalism is more modest because it sets limits on what sorts of systems can count as cognitive. Here is the third role of the Martian Intuition: radical externalism sets the bar of the mental so low that it allows phenomena that are, *prima facie*, non-mental to count as mental. For instance, by acquiring a book, the agent comes to believe everything contained in the book (Sprevak 2009, 517). The reason is that we can imagine Martians who encode beliefs using ink within the bounds of their bodies, are born with innate beliefs, and do not necessarily access these beliefs. The result is a Martian with the physical and functional equivalent of a book within its head which contains its non-accessed innate beliefs. If this system counts as mental, then, according to the Parity Principle, so should a system constituted by a person who has just acquired a book. Other examples: by stepping into a library I acquire millions of beliefs; by browsing the Internet I acquire billions of beliefs (Sprevak 2009, 518). Sprevak argues that, because it is implausible to count such processes as mental, radical externalism should be rejected. The Martian Intuition works against constraining externalism, so the problem can be traced back to functionalism itself. Externalism and functionalism probably contain insights into the mind and thereby provide useful material for developing their replacements (Sprevak 2009, 527), but if Sprevak is correct they are false.

The crux of Sprevak's critical argument is the construction of implausible cases of extended cognition using the combination of the Parity Principle and the Martian Intuition. However, the Parity Principle is subtler than the discussion so far suggests. As formulated by Clark and Chalmers and accepted by Sprevak, this principle functions by directing our attention to internal processes which we confidently count as cognitive and then extending this status, and the correlative confidence, to processes that extend beyond the bounds of agents into the wider world. But this is not the only place we find such confidence. We should see the familiar Parity Principle as the first part of a two-part principle. Here is the second part:

(PP2) If, as we confront some task, there is a process in the head that, were it to extend into the world beyond the agent, we would have no hesitation as accepting as not cognitive, then that process in the head is not a cognitive process.

We can be just as confident about what is not cognitive as we can about what is cognitive. PP2 codifies this confidence to guard against chauvinism in judging cases.

PP2 complicates Sprevak's argument. The reason is that the implications of our confident judgments about what processes count as respectively cognitive and not cognitive can be inconsistent. Consider Sprevak's confidence in the non-cognitive status of a system constituted by a person with a new book and his confidence in the cognitive status of the Martian processes constituted by ink-encoding of beliefs in the head, innate beliefs, and partial access to these beliefs. We have seen that Sprevak uses his confidence in the Martian case plus the Parity Principle to generate a judgment of "cognitive" for the person-plus-book system, which he finds to be unacceptably implausible. But we can run this argument the other way around. Let's begin with the person-plus-book system. Suppose that we confidently judge that this is a non-cognitive system. When this judgment is combined with PP2, the implication is that an analogous system that is located within the physical bounds of an agent should also be judged to be non-cognitive. The Martian system of ink-encoding of beliefs in the head, innate beliefs, and partial access to these beliefs is such an analogous system. Therefore we should see this system as non-cognitive. Strictly speaking, we should redescribe this case, insofar as 'belief' is a cognitive term that no longer applies here.

The ideas behind the Parity Principle turn out to be more nuanced than was expected, and the result in the present context is a clash of intuitions. Run one way, the Parity Principle generates a challenge to locational externalism and functionalism. Run another way, PP2 challenges our imagination about Martian cases. In general we could argue about which intuition is stronger, but this would not help Sprevak's argument, as he is committed to both the cognitive status of the Martian-ink case and the non-cognitive status of the book case. Moreover, such battles of intuitions are invariably unsatisfying. What is preferable is a principled way of adjudicating this clash.

Generally, and in a manner particularly germane in the present context, functionalism provides the tools for making progress with the (non-) cognitive status of these cases. The crucial question is whether these systems involve beliefs. Are there beliefs in the ink-Martian case? Does a person who buys a book thereby automatically come to believe the ideas

contained in the book by realizing a person-plus-book cognitive system? Functionalism directs us to things that are relevant to answering these questions. Here is Sprevak's characterization of functionalism: "Functionalism preserves the Martian intuition by claiming that what makes an organism have a mental state is the organism's functional organisation. This is typically understood in terms of the notion of a causal role, which in turn is understood as a pattern of typical causes and effects." (2009, 509) To assess whether our cases involve beliefs, we must have access to some specification of the typical causes and effects of beliefs. Too fine-grained a specification will render our account of belief chauvinistic. Too coarse-grained a specification will render our account too inclusive. Doing without any specification leaves us unable to make determinate judgments about cases. To see what might be in such a specification, consider two of the ways in which we can develop the book case:

(A) I buy a book. It is written in a language that I understand, but I have not read it yet. The topics are familiar to me, but I have not yet formed firm beliefs about them. If I were to read the book, I would be inclined to assert the ideas it contains. I would adjust my conduct in accordance with those ideas.

Here we have reason to think that the person-plus-book system really does involve something much like beliefs, albeit ones that are not yet deployed in on-line processing. The representations in question are accessible to my higher thought functions. If used on line, my earlier exposure to information about these topics would combine with these representations to deliver clear cases of beliefs. Potential access and responsiveness to rational processes informed by evidence are plausible hallmarks of belief. So are potential assent and suitability for guiding conduct. The pieces of a plausible characterization of the typical pattern of causes and effects of belief found here license the judgment that something much like beliefs, by functionalist standards, is found in this case. But now the case stands not as an implausible challenge to locational externalism, but as a plausible but surprising case of a cognitive system by functionalist standards.

(B) I buy a book. It is in a dead language for which no translating procedure exists, nor is there any reasonable hope of finding a "Rosetta Stone" key to its syntax and vocabulary. I have no familiarity with the topics. If

I were to read the book (which I cannot do), I would not be inclined to assent to its contents, nor would I adjust my conduct accordingly.

In this version of the case, virtually no elements of the pattern of the typical causes and effects of belief are found. Accordingly, we have no particular reason to see the person-plus-book system as a cognitive system involving beliefs. Again, this runs counter to Sprevak's argument: no implausible challenge to locational externalism is found here.

One might worry that the appeal to functionalism is question-begging in the present context, supposing that Sprevak claims that the cases generated by the Parity Principle and the Martian Intuition function as a challenge to functionalism. However, such a supposition would mistake the nature of Sprevak's argument. Sprevak argues from functionalism and the Parity Principle to the problematic cases. The present argument claims that Sprevak's argument omits important details about how the Parity Principle and functionalism work. That is, the argument claims that Sprevak's premises deserve more scrutiny. Once these details are taken into account, we see that Sprevak's argument does not go through. The problematic cases that are offered as a challenge to functionalism are not actually generated in the manner portrayed by Sprevak's argument.[4]

Let's put general objections to externalism behind us. From this point on, the important points will be more specific wide and narrow hypotheses concerning moral cognition. To prepare for these hypotheses, let's attend to wide systems themselves in more detail.

1.5 A Model for Thinking about Wide Cognitive Systems

A system takes inputs and delivers outputs, both of specific kinds. The relations between inputs and outputs are governed by if-then rules.[5] These rules need not be codified, of course; for some systems they are codified and for others they are not. For instance, I have an alphabetization-and-date system for my record collection: I take a musician's name as input, and the output is a particular place on my shelves for storing recordings by that musician. In the case of multiple recordings by the same musician, I take the date of recording as the input to a subsystem that also delivers a shelving order as output: earlier recordings are stored before later ones. I had never articulated these rules for this system until writing these words,

but in my case this particular system has existed for more than two decades.

Systems come in many kinds. Our present interest is in cognitive systems—that is, in the cognitive processes by which inputs are correlated with outputs. Moreover, since this book is an exercise in psychological theorizing about actual humans, the topic is causally realized cognitive systems that, by hypothesis, actually implement our thought about morality, rather than, e.g., merely formal systems. On paper my record-storing system is merely formal. When I put away batches of records, this system is causally efficacious in producing my actions. It is typical to think of cognitive systems as contained within the physical boundaries of individual people or other organisms. This assumption is challenged by the Extended Mind Hypothesis. Some wide systems will use cognitive resources, such as symbolically encoded symbols—e.g., printed letters and numbers. Other wide systems will use cognitive resources of other kinds. In the following chapters, I will argue that other people play a particularly important cognitive resource for moral psychology. To clarify the issues involved in thinking of psychological systems that exist between individuals, here is a simplified model.

Think of birds traveling in a flock (a tricky phenomenon to explain). Craig Reynolds (1987) has famously provided a computer simulation of the flocking of birds.[6] Reynolds calls his computer creatures "boids." Boids exhibit very realistic flocking behavior. This is achieved using three rules for steering:

Separation: Steer to avoid crowding local flockmates.

Alignment: Steer toward the average heading of local flockmates.

Cohesion: Steer to move toward the average position of local flockmates.

These rules require that boids monitor their immediate neighbors. The more complex phenomenon of flocking—that is, moving as a unit, dividing and recombining, and changing direction together—emerges from the behavior of individual boids as they track their local circumstances. They do not have a plan to form a group, or to follow a specific leader. This is very suggestive about how actual birds might accomplish their complex flocking behavior.

Now imagine a group of birds that act in accordance with the steering system codified above for boids. Actual birds do things other than fly

together—for example, they also seek food and avoid predators.[7] Let's add the cognitive capacities to find food and predators to our imaginary birds; for present purposes these need not be specified in any detail. Imagine the flock traveling through the air. The east-most bird sees food on the ground. This food is hidden from the west-most bird. The east-most bird heads toward the food. Nearby birds adjust their behavior in response to both the east-most bird and the food and subsequently follow. The west-most bird, in accordance with the three steering rules, adjusts its motion to keep up with the flock. As a result, the west-most bird ends up at the source of food.

Let's suppose, as seems quite plausible, that the behavior of the east-most bird can be explained in terms of psychological capacities located completely within that bird's physical boundaries. It takes the information about food as input, and the output is flying toward the food. The relevant systems are, by hypothesis, locationally narrow. How should we understand the behavior of the west-most bird? One possibility is that the behavior should be understood solely as a local response to the movements of its neighbors. Another possibility is that we should construe the bird's behavior as a response to the food and also as a response to its neighbors. One might balk at this interpretation on the grounds that the bird did not actually encounter the food, and so its behavior could not be a response to the food. However, externalist ideas give us a way to make sense of this: Perhaps the bird is part of a wide cognitive system. The input to the system is the information about the food. This information is taken in by, primarily, the east-most bird, which produces the output of turning toward the food. This information is taken as input by the intermediary birds and is subsequently processed via their responding movements, until it can produce the flying behavior of the west-most bird. Note that it is not required that the west-most bird *realize* that there is food to be had, or *know* about the food, or anything of the sort. To require that would be to suppose that the information about the food would have to be taken explicitly as input by the west-most bird in order for it to play a role in producing the bird's behavior. But such is not the case: the wide system, not the west-most bird, processes the information about the food. The west-most bird need only be able to play a role in this system of a sort suited for the processed input to produce the relevant behavior.

Thus we have two interpretations of the behavior of the west-most bird. How should we decide between them? One pertinent question, if not the

crucial question, is whether we are warranted in seeing the relations between the birds as systematic. To answer this, recall the schema for systemicity extracted from Wilson's work:

_____ systems must be causally and functionally integrated chains of _____ resources, and these, individually and collectively, must play a replicable causal role in _____

The issue of a "replicable causal role" can be put aside if we assume that this overall phenomenon is typical flocking behavior. The resources in question are, in the specific case, the information about the food, the steering capacities of the birds, the birds' movements that the steering capacities track, and the birds' food-detection capacities. What should we say about the causal and functional integration of these resources? Let's begin with causal integration. By hypothesis, the birds have specific steering rules for tracking their neighbors and responding to their whereabouts. That is, the behaviors typical of flocking are not by-products of more general procedures for moving or for tracking features of the environment. Thus, it seems to me that we are warranted in seeing the birds as exhibiting the requisite degree of causal integration. (N.B.: 'Degree' is the correct word here, as there is no specific line that, once crossed, divides systemicity from non-systemicity.)

These remarks about causal integration make reference to what the cognitive capacities in question are for—that is, they raise the issue of functional integration. In the case of boids, we can say that they exhibit functional integration because they were deliberately designed by humans to track each other in specific ways. Our imaginary birds are importantly different, insofar as nobody designed them. In this case the question of functional integration has to be addressed from a thoroughly naturalistic perspective. The natural way to address this, if not the only way, is to ask about the evolution of the birds' cognitive capacities. Without getting into the complex debate about the nature of natural functions,[8] here is a suggestion: if the finding of food via the following of nearby birds has contributed to the cross-generational persistence of the steering capacities by increasing reproductive fitness, then we have reason to think that flying in groups is not the only function of these capacities, and that finding food is another of their functions. By extension, birds' movements transmit information not only about themselves but also about the location of

food in the wider world.[9] Evidence about the evolutionary descent of such capacities might be difficult to gather, but the conjecture that finding food has contributed to the persistence of birds' navigational capacities strikes me as initially quite plausible, so I think we have *prima facie* reason to see the birds' movements, steering capacities, and environmental opportunities as functionally integrated to the requisite degree. That is, in this case we have reason to think that we find wide systemicity.[10]

Two *very* general things can be said about cognitive prerequisites for participating in wide systems, at least with regard to humans. First, the birds in this hypothetical case have the capacities to track the movements of their neighbors, but arguably other capacities are implicit in this example. The birds we have been considering are a sociable bunch: they are content to be around each other, and no conspecific hostility features in the case. This general state of affairs is important for participation in at least some wide systems that involve the use of some organisms as cognitive resources by other organisms. If the west-most bird were unwilling to pay attention to its neighbors, it would not be able to participate in the information processing that they make available. Second, although I have just described the birds as sociable, they are not nearly as social as humans.[11] The birds in this example participate in a wide cognitive system by tracking movements. In contrast, and in the spirit of much research into human sociality in general, I conjecture that many important wide systems in which humans participate require that we track each other's thoughts. If this is correct, then so-called mind-reading capacities are going to be required for individual humans to get access to wide cognitive resources.[12]

Let's return to the birds. If the wide interpretation is correct, then the west-most bird processes both information about the location of its neighbors and information about the location of food; it is aware, at most, of only the former. Thinking of the birds as taking part in a wide cognitive system allows us to distinguish three types of potential input. First, there is input to the individual's psychological capacities alone. Let's call this *unmediated* input. We can presume, for now, that this is an apt way to characterize the east-most bird's encounter with the food. Second, there is input that an individual does not directly encounter at all, but that is processed by the wide system. Let's call this *mediated* input. The west-most bird's processing of the information about the location of the food is mediated. Finally, there is information that is processed both directly by an

individual and by the wide system. Let's call this *dual* input. When an intermediary bird both sees the food and responds to the movements of the east-most bird, which is moving toward the food, it is dealing with dual input.

Dual input is tricky. It should give us pause with regard to how we think of unmediated input, at least for social creatures such as ourselves. Real birds not only follow each other and share food; they also compete for food and other opportunities. Humans are no different. But human social life is massively complex precisely because of the opportunities for manipulating each other for physical and social gain.[13] Thus, when two or more humans share input, we should be careful to include cognitive capacities for assessing and dealing with competition in our account of the psychological processes at work. This idea has at least two general implications. First, it makes the general openness to wide systems and resources an even more important prerequisite. Suppose that another person or some other organism is competing with me for food, status, and other opportunities, and that that individual has ideas about what it deserves, what its status is and should be, and what it can do to protect itself and to get ahead. It will be important for our generally agreeable co-existence that I appear to the other individual to have largely the same ideas. If I have different ideas—for example, that I, rather than he, she, or it, deserve X, and that my status is more important—then I pose a significant threat.

So far this point has been made in terms of two individuals and a limited number of topics of thought. But human life is far more complex than that. We interact with vast numbers of people, and about a relatively open-ended group of topics. We stand to each other in complex relations of power, status, threat, entitlement, and opportunity. Thus, it is not only important that I appear to agree with individual A about topic P; it is important that I generally fit in with most people, about an open-ended number of issues. This imposes on individuals a general, complex pressure to conform. This should set us up in particularly good position to realize wide cognitive systems with other individuals.

Consider how conformity might be psychologically implemented. Suppose that, regardless of how one actually thinks, there is reason to appear to agree with the views of others. One way to do that is to have mechanisms that suppress one's own contrary judgments and produce conforming behavior. But another way is to have mechanisms that conform

one's judgments to those of others. I see no reason to think that we do not have both sorts of mechanisms. If this is right, then thinking about competition and dual input has implications for how we think about the processing of unmediated input. If we have judgment-conforming processes, then the absence of other people in a particular situation is not significant: the effects of conformity extend all the way in, so to speak. Research into psychological heuristics provides some support for just such a phenomenon—see, for example, Gigerenzer 2008, 24; Richerson and Boyd 2005. Thus, even when dealing with input that is isolated from other individuals, it should be predicted that we will act as social animals. The availability (or the unavailability) of wide cognitive systems will be relevant even to the processing of unmediated input.

In light of the above, the term I use to describe people is *Like-Minded*. We are like-minded in two respects. First, as social animals we are under significant pressure to conform our views of the world to those of our conspecifics. Second, insofar as we participate in wide cognitive systems, partly in virtue of the psychology of conformity, there is an important sense in which we literally share psychological processes with other people. We think the same partly because it is prudent and partly because we use the same token systems to think. Sometimes we enter these systems as autonomous equals, sharing information through dialog and reasoning together to form judgments, solve problems, and generally figure things out. At other times, these systems are constituted not by explicit intersubjective reasoning but in other, less obvious ways. This is one of the lessons of the hypothetical birds: although they do not reason together explicitly, they nonetheless think together via subtler wide cognitive systems. The fact that these systems are relatively inconspicuous helps to explain why philosophers and psychologists tend to overlook them. Nevertheless, I am inclined to think that these less obvious ways of thinking together are the more important ones.