Philosophical Perspectives, 23, Ethics, 2009

MORAL CONCEPTS AND MOTIVATION[1]

Mark Greenberg
UCLA

1. Introduction

Abner's mother works for a regulatory agency. Abner overhears her talking with a colleague about an imminent regulatory action. He realizes that he could make money by buying stock in certain companies. You try to convince him that it would be morally wrong for him to do so, and it seems that you are successful. But Abner's apparent recognition that buying stock would be morally wrong has no impact on his decisions, intentions, or actions. He goes right ahead, without any struggle between conflicting motivations, and buys the stock.

Assuming Abner is sincere, it is natural to doubt his grasp of the concept of wrongness. We might express this doubt by saying that if Abner really understood what it meant for something to be morally wrong, he would at least have some tendency not to buy the stock. Examples like this one are therefore used to motivate what I will call *moral internalism* (*internalism*, for short) — the view that there is an internal (or necessary, conceptual, etc.) connection between moral facts or judgments and the motivation of action.[2]

Cases such as Abner's seem to be possible, even actual. It seems that people can recognize moral facts without having any tendency to be motivated accordingly.[3] Examples like this one are therefore used to motivate what I will call *moral externalism* (*externalism*, for short) — the view that it is *not* the case that there is an internal connection between moral facts or judgments and the motivation of action.[4]

Obviously, there is something peculiar about the fact that internalists and externalists appeal to the same kind of example to motivate their respective views. I will suggest that each side gleans a sound insight from the examples, but each side takes away the wrong bottom line. The sound internalist insight is that there is something conceptually defective about someone who recognizes a moral fact or makes a moral judgment, yet is not appropriately motivated. The sound externalist insight is that people can recognize moral facts or

make moral judgments without having even a disposition to be motivated accordingly.

These two insights are incompatible given the presupposition, shared by most participants in the debate, that *if* concepts are individuated by their role in reasoning — specifically by canonical inferences or, more generally, mental transitions — then for a thinker to have a concept is for the thinker to have a disposition to make the concept's canonical transitions. (As I use the term, to *have a concept* is to have, or be able to have, thoughts in whose content the concept figures.)[5] I have just referred to this conditional as a presupposition, but it might be more accurate to say that the view in the antecedent is not carefully distinguished from the view in the consequent. The critical result is that it is taken for granted that if a mental transition is individuative of a concept, a thinker cannot have the concept without having a disposition to make that mental transition. For example, Michael Smith argues that because they lack the disposition to move from moral judgments to appropriate motivational states, "amoralists do not have mastery of moral terms, and they therefore do not really make moral judgments."[6] (Of course, externalists use the presupposition to argue contrapositively: because amoralists do make moral judgments and therefore must have mastery of moral concepts, such mastery must not require a disposition to be appropriately motivated to act.)

I will develop a new type of example to argue that standard versions of internalism are false. I will argue, however, that more general considerations in the theory of content suggest that the implication of this type of example is not that there is no conceptual connection between moral judgments and motivational states, but that we need a different view of what is required to have a concept. According to this view, thinkers can have thoughts involving a concept without having a disposition to make the concept's canonical mental transitions. This suggestion yields a way of reconciling the insights of internalism and externalism within a cognitivist account of moral judgments.[7]

The position deserves to be called a form of internalism because it holds that a connection to motivation is built into moral concepts. It honors the internalist insight that one who *fully* grasps a moral concept will be motivated: mastery[8] of a moral concept consists in part of having the appropriate motivational disposition. At the same time, the position honors the externalist insight because it allows that thinkers can recognize moral facts and make moral judgments without having mastery of moral concepts. It therefore is immune to externalist arguments that thinkers seem to be able to make moral judgments without being motivated. Similarly, the position is not susceptible to prominent objections, such as John Mackie's famous argument from queerness and the objection based on the Humean view that beliefs are motivationally inert.

My arguments will mostly focus on moral concepts, but similar points apply to normative concepts generally, though I won't have space to say much about them. Also, my arguments will mostly concern the connection between moral judgments and motivation, but similar points apply to the connection between

moral judgments and judgments about reasons and to the connection between judgments that certain (nonnormative) fact patterns obtain and judgments involving specific moral concepts. That is, the arguments show how these connections can be built into moral concepts, despite the fact that thinkers need not have the corresponding dispositions.

In section 2, I develop the examples that challenge standard versions of internalism. Next, in section 3, I turn to considerations in the theory of mental content to determine what lesson should be drawn from the examples. Finally, in section 4, I briefly explore the way in which the argument can be extended to the (non-moral) normative concept of what one ought to do, all things considered.

2. Moral Judgments without Motivation

Internalism, defined as the view that there is an internal (or necessary, essential, etc.) connection between moral judgments and motivation, is vague and abstract. It is standardly interpreted in a more precise and concrete way. According to one version of this standard interpretation, necessarily, anyone who recognizes that an action has a moral property is motivated to act in the appropriate way — at least if the question of the thinker's taking that action arises.[9] I will call this version the *strict version* of the standard interpretation of internalism — or, for short, the *strict thesis*. The basic thought is, for example, that someone who judges that an action is cruel or wrong has to be motivated against taking that action. The motivation need only be pro tanto — it may be outweighed or overridden by other motivations — so it may not lead to action.

Recognizing that there seem to be cases in which people make moral judgments without being motivated, many proponents of internalism understand it in a more relaxed way — as the claim that, necessarily, anyone who makes a moral judgment is *disposed* to be motivated or is motivated *other things being equal*. This *dispositional version* of standard internalism (*the dispositional thesis*) is consistent with the possibility that, in a particular case, someone might make a moral judgment but fail to be motivated to act (even *pro tanto*) because of some interfering factor, such as depression or weakness of the will.

Another common claim is that, necessarily, one who makes a moral judgment is motivated accordingly, if she is rational. This *rationality thesis* can express very different positions — for instance, a version of the dispositional thesis — depending on the conception of rationality involved.[10]

In this section, I will develop a new type of example that challenges all versions of the standard interpretation of internalism, including the rationality thesis on some ways of understanding it. The examples involve thinkers who apparently make moral judgments yet, as a consequence of their unusual beliefs, lack even a disposition to be motivated accordingly.[11]

I will discuss one main example involving a specific moral concept and will indicate how other examples involving moral concepts could be developed in a

parallel way. I will also more briefly discuss an example involving the normative but nonmoral concept of what one ought to do, all things considered, in part because this example shows how far the type of example can be extended.

A few words about methodology. The attribution of content to thoughts, such as beliefs and intentions, is a pre-theoretical phenomenon. The point of my examples is to appeal to our ordinary practices of attributing contentful mental states in order to generate the pre-theoretical data. We may ultimately find theoretical reasons for rejecting or reinterpreting that data. But since the point is to determine whether our ordinary standards attribute thoughts involving the relevant concepts, it is important, in evaluating the examples, to put aside theoretical preconceptions or general views about what is required to have thoughts involving a concept.

One example much discussed in the literature is that of the so-called amoralist. The amoralist's views and positions are typically not much fleshed out. Rather, we are simply told that he is able to and does use moral terms to classify actions (character traits, states of affairs, etc.) in the same way as other people, but lacks any motivation to act appropriately. Such underdescribed examples make it hard to avoid stalemate. Some philosophers assert that there is no reason to deny that the amoralist has moral concepts; others insist that he cannot have them, that his apparent uses of, for example, *wrong* are best understood as something like *what others mean by wrong*. In order to move forward, we need examples that give us an understanding of the agent's psychology.

Consider Alice, a moral and political philosopher with a strong libertarian streak. She develops an elaborate moral theory according to which liberty is the fundamental value, and equality is not a value at all. In working out the consequences of the theory, she forms the hypothesis that considerations of fairness are really considerations of equality in another guise. She finds some arguments that support the hypothesis. Eventually, she comes to question whether fairness is a moral virtue, indeed whether it is a reason for action of any kind. Accordingly, she loses the belief that fairness is a reason for action.

Alice believes that it is important to bring one's motivations and actions into line with one's beliefs, and she is good at accomplishing this in her own case. Because of her doubt about whether fairness is a reason for action, she apparently loses any disposition to be motivated to perform fair actions *qua* fair. (Of course, many actions that are fair also happen to be actions that Alice is motivated to perform for other reasons.) It is not that she is motivated to some extent to perform fair actions, but does not act on that motivation because, for example, it is overridden by other motivations. Rather, in line with her doubt about whether fairness is a reason for action, she is not motivated to perform fair actions to any extent. She also apparently loses any disposition to feel resentment or indignation at unfairness.

Alice's theoretical views about liberty, equality, and fairness are no stranger than views that have been held by many philosophers. For example, many utilitarians think that fairness has no moral relevance. The independent moral

value of equality has often been questioned. For present purposes, the details of Alice's arguments do not matter. I will assume that Alice's doubts are in fact misplaced, and that fairness is a reason for action.

Alice's worry is not that the concept of fairness has no application. She apparently continues to have fairly standard views about which procedures, rules, people, etc. are fair. She continues to use the word "fair", and she continues to be adept in classifying things as falling under the term.

On the face of it, it seems hard to deny that Alice believes that fairness is not a moral virtue, that fairness is just equality in another guise, and so on. Since such beliefs involve the concept of fairness, it follows that she is able to have thoughts involving the concept of fairness — or, in my terms, she has the concept of fairness.

What about specific judgments about which things are fair? A theorist with Alice's beliefs and doubts *could* reasonably decide not to use the concept *fair* to make specific judgments, instead substituting judgments involving a concept such as *what others consider fair*. That is, rather than judging that a particular procedure is fair, such a theorist would judge that the procedure is what others consider fair.

In fact, however, Alice does not decide to distance herself from the concept in this way, and does not intend to stop making specific judgments about what is fair. One reason is that she knows that other people take fairness to be an important feature. She therefore often finds it useful to think about and talk about what is fair. For example, she tries to convince her Dean that it would be fair to raise her salary. Similarly, when her department chair instructs her to find a fair procedure for allocating scarce places in her class, she (apparently) is able to make judgments about which procedures are fair, though of course she thinks that their fairness provides no reason to adopt them. Moreover, out of long habit, she continues spontaneously to judge that actions are fair and unfair — or, so as not to beg any questions, that is at least how she would characterize her judgments. In sum, it appears that Alice makes judgments about what is fair.

The example seems to show that a good-willed person can make moral judgments without having a disposition to be motivated accordingly (and also without judging that there are corresponding reasons to act). At an abstract level, there are two ways to dispute the example. First, one can argue that, despite appearances, Alice must still have the disposition to be appropriately motivated. Second, one can argue that Alice cannot have thoughts involving the concept of fairness. I will consider each type of objection in turn.

Alice manifests no tendency to be motivated in favor of fair actions. Therefore, if she still has the relevant disposition, it must be the case that something is interfering with its manifestation.

The distinction between a system's not having a disposition to Φ and its having the disposition but its not being manifested presupposes roughly that the best explanation of Φing, when it occurs, involves appeal to different

subsystems — a discrete disposition to Φ together with the cooperation of a range of other subsystems or necessary conditions. The other subsystems can be damaged without damaging the mechanism that underwrites the disposition to Φ, so the existence of the disposition is consistent with its not being manifested. Since the other subsystems are necessary for Φing to occur, there has to be some reason why explanation is furthered by treating them as not necessary for the existence of the disposition to Φ. A typical kind of reason is that they are not specific to Φing but are necessary for a range of activities. For example, the ability to add integers may not be manifested because of a defect of short-term memory, attention, or communication. A different, though often overlapping, reason is that with respect to the question at stake, it makes sense to consider the functioning of the other subsystems a normal background condition of the system. For example, if we are interested in the question of whether someone has the ability to add integers, it makes sense to treat the presence of oxygen or the person's being conscious as a normal state of the system.

The general question of when a disposition is present but not manifested is complex, but, for present purposes, we can largely avoid the complexities. The best available candidate for an interfering factor is the fact that Alice no longer believes that fairness is a reason for action.[12] I am trying to show only that there is a possible case in which someone judges that an action is fair and yet is not disposed to be motivated. Therefore, in order for the present objection to succeed, the objector must give a principled reason for insisting that the best explanation of Alice's no longer being motivated to act fairly could not be that her belief that fairness is a reason for action was part of the mechanism underwriting the relevant motivational disposition or something on which that mechanism depended. It is difficult to see what this reason could be.

For example, it is not at all plausible that the loss of the belief that fairness is a reason for action must prevent Alice from forming the relevant motivation by interfering with some subsystem, not specific to the motivation to perform fair actions, but needed for the formation of motivations in general. In general, what we are disposed to do depends on our beliefs. This is particularly obvious in the case of the dispositions we are concerned with — dispositions to perform actions that are judged to fall under sophisticated abstract concepts, such as the concept of fairness. We would expect such dispositions to be developed in part because of, and to depend for their continued existence on, beliefs of precisely the sort in question.[13]

Similarly, it is difficult to see why the presence of the specific belief that fairness is a reason for action is a normal condition. What beliefs one has is precisely the kind of thing that can affect what dispositions one has to make mental transitions. Therefore, in general, in asking whether someone has a disposition to make mental transitions putatively required to have a concept, we cannot take the presence or absence of specific beliefs to be normal conditions. In the present case, it would obviously be ad hoc to treat the belief that fairness is a reason for action as an exception.

The objector cannot claim that Alice's state is abnormal merely because she lacks a true belief (or even because she has a false belief). It would not make sense to take her being omniscient or having no false beliefs as a normal condition; for, under such a condition, she would have very different dispositions and concepts than she in fact has.

A final possibility is that the objector could claim that Alice's state is abnormal because she is irrational. Some conceptions of rationality would make this suggestion a nonstarter. For example, given a conception of rationality as responsiveness to all applicable reasons, an ordinary person would have very different dispositions if she were rational. At the end of this section, I argue that Alice is not irrational on a conception of rationality on which irrationality could plausibly count as an abnormal condition or interfering factor.

I now turn to the objection that concedes that Alice lacks the relevant motivational disposition and claims that it is not possible that she has the concept of fairness. There is no doubt that, before she developed her theory, Alice had the concept. She not only was skilled in classifying actions as fair, she also had the appropriate disposition to be motivated (as well as dispositions to make other plausibly relevant mental transitions). Indeed, as a moral philosopher, she had a particularly sophisticated understanding of the concept. It would be strange for her to lose the ability to have thoughts involving the concept merely by acquiring a new theory and adjusting her motivations accordingly.

In the case of the amoralist, as noted above, there is a temptation to think that the amoralist is merely mimicking others' use of moral terms without really having thoughts involving moral concepts, or that he is best described as making judgments only about, for example, what other people call "kind" or what other people call "all-things-considered morally required," rather than about what *is* kind or morally required. Part of the reason for the temptation is surely that the amoralist's mental life or outlook is not described or explained. Add to this the fact that the amoralist is utterly unmoved by *any* moral considerations. Since we are given no explanation of what kind of mental life could lead a human being to be like this, we naturally imagine an alien psychology. Consequently, if one's theory has the consequence that the amoralist cannot have moral concepts, it is easy to defend one's theory by maintaining that the amoralist's thoughts involve different concepts from ours, perhaps ones that would be unintelligible to us.

By contrast, it can't be maintained that Alice must use all moral concepts only in "inverted commas." Alice has appropriate motivational dispositions with respect to many moral concepts, such as *cruel* and *wrong*, and there seems to be no reason to think that she does not have those concepts. We can even suppose that Alice is apparently an especially morally good person. Moreover, we understand Alice's belief system, which is of a relatively familiar type, even if we disagree with it, and we can see how her belief system leads her to have the motivations that she has.

Given all this, it is hard to deny that, by ordinary standards, Alice has many thoughts involving the concept of fairness. In order to resist the conclusion that

Alice makes specific judgments involving the concept of fairness, it will therefore be necessary to rely on theoretical grounds. In the present context, however, it is question-begging to appeal to the widespread presupposition mentioned in the introduction: if concepts are individuated by transitions between mental states, then for a thought to involve a particular concept is for the thinker to be disposed to make the concept's canonical transitions.

First, the point of the example is to challenge the dispositional thesis, but, given the internalist idea that a connection between moral judgments and motivational states is built into moral concepts, the dispositional thesis follows immediately from the presupposition.[14] More importantly, as we will see below, the examples involving moral concepts are instances of a broader phenomenon of thinkers who apparently have concepts yet fail to be disposed to make the mental transitions that are plausibly constitutive of those concepts. This phenomenon presents a serious challenge to the presupposition. Given the dialectical situation, an argument that Alice does not have the concept of fairness cannot be based on grounds that take for granted the presupposition. Other than such grounds, it is not clear what would be the theoretical basis for claiming that Alice does not have the concept of fairness.

I now want briefly to make a more positive case for thinking that Alice has the concept. Alice's view about fairness is similar to the view that many contemporary people have about chastity, temperance, patriotism, humility, or devoutness, each of which was once widely thought to be a moral virtue (and, to differing degrees, still is). There was undoubtedly a time when those who doubted whether, say, chastity was a moral virtue were a tiny minority. As such examples show, it is possible to doubt whether something that is widely thought to be a moral virtue really is one. Moreover, the possibility of doubting whether fairness is a moral virtue should not depend on whether it really is one. If it is conceded that it is possible to doubt that fairness is a moral virtue, it would be strange to maintain that one can do so only as long as one has the disposition to perform actions one believes to be fair. Once Alice comes to doubt that fairness is a virtue, she no longer has reason to be motivated to perform fair actions (*qua* fair actions), so being disposed to be so motivated while having that doubt is, in a clear sense, irrational.

It would be relatively straightforward to construct an example similar to that of Alice involving the concept of what is morally required (all things considered). For example, we could imagine a philosopher who comes to question, perhaps along the lines attributed by Plato to Thrasymachus, whether there is reason to do what morality requires (but does not doubt that the concept of what is morally required has application).

We could also develop examples involving thinkers who have run-of-the-mill incomplete (or incorrect) understanding of moral concepts on the model of familiar examples, such as Burge's (1979) person who believes he has arthritis in his thigh.[15] Just as an unusual theory can lead a sophisticated theorist to lose a motivational disposition, so incomplete understanding can lead an

agent to believe, for example, that fairness or rightness is not a reason for action and consequently not to acquire the relevant motivational disposition. I believe that such examples can be deployed effectively against standard interpretations of internalism, but I don't use them here because they raise additional complications.[16]

Alice's case and similar cases are problematic for any version of internalism that has the consequence that people who make moral judgments are necessarily motivated or disposed to be motivated. Such a version of internalism may seem the only available understanding of internalism, however. On the one hand, on non-cognitivist accounts of moral judgments, those judgments simply express an appropriate conative or affective attitude rather than belief, so one who makes a moral judgment necessarily has an appropriate motivational state. After all, one main aim of such accounts is to explain the presence of such a state.

On the other hand, assuming cognitivism about moral judgments, the attitude involved in a moral judgment must be the same as that involved in an ordinary nonnormative judgment. In general, the connection between ordinary judgments and motivation is contingent. Therefore, if there is a necessary connection between moral judgments and motivation, that connection cannot be explained by the nature of the *attitude* of judgment. It must be accounted for by the *content* of the judgments (given a standard picture of judgments as consisting of an attitude to a content). Specifically, the connection must be built into moral concepts since they are the distinctive component of the content of moral judgments.[17] In that case, concepts must be individuated at least in part by their role in transitions between mental states, as opposed to purely by the properties to which they refer. (Such a view of concepts can allow, for example, that the transition from judging that an action is right to an intention to take that action is partly individuative of the concept of rightness.) Now, as noted, it is widely presupposed that given such a view of concepts, *conceptual role semantics* — the view that what it is for a thought to involve a particular concept is for the thinker to be disposed to make the concept's canonical mental transitions[18] — follows immediately. Given the presupposition, it follows that a thinker cannot, for example, judge that an action is right without being disposed to make the concept's individuating transitions. We therefore have a straightforward line of argument from the abstract idea of internalism to the dispositional thesis.

Jackson and Pettit (2004) offer a conceptual role semantics account of moral beliefs and judgments in order to explain, among other things, the connection between moral judgments and motivation.[19] They distinguish between two different ways of believing that something is, for example, fair or right: a non-intellectual way and an intellectual way. To believe in the non-intellectual way that something is fair is roughly to have a state that plays the appropriate role in one's mental economy. Consequently, "there is no way of judging non-intellectually that something is fair without experiencing a suitable desire for the option in question. The idea of forming a fairness-belief in this non-intellectual way, and yet lacking the desire, will be . . . incoherent." (2004, 208) Jackson and Pettit

appeal to the intellectual way of believing that something is fair to explain the possibility of judging that something is fair without at the same time desiring it. To believe in the intellectual way that something is fair is, roughly, to recognize the correctness of the concept's canonical mental transitions. For example, to believe in this way that something is fair is, among other things, to believe that "I can justify myself in choosing the prospect in question, [and] that . . . I would desire the option were I ideally situated." (2004, 200–201) In sum, the only way in which one can judge that something is fair without being motivated accordingly is to recognize in a purely intellectual way "that one would desire it were one not paralyzed by some evaluative malaise." (2004, 208) But the Alice example shows that one can judge that something is fair without either being disposed to be motivated or recognizing that one should be motivated or would be motivated if one were ideally situated.

Alice's case also presents a challenge to the rationality thesis, though the issues here are delicate, and I lack space to explore them adequately. Remember that the rationality thesis holds that, necessarily, a person who makes a moral judgment is motivated to act accordingly, if he or she is rational. As others have pointed out, if the rationality thesis is going to be an interesting view, we need a non-question-begging conception of rationality. In addition, if the rationality thesis is going to be a form of internalism, we need a conception of rationality that is not so strong that it accounts for the link to motivation without the need for an internal connection between moral judgments and motivation.

For example, for some purposes, it may be useful to conceive of rationality as responsiveness to reasons. On this conception, a fully rational person is one who is always motivated to act (and does act) in accordance with the reasons that apply to him or her. In a context in which the issue is the connection between moral facts or judgments and motivation, however, it is beside the point to consider what someone would do if he were always motivated to act on all the relevant reasons. For example, on this conception of rationality, an externalist — one who denies an internal connection between moral judgments and motivation — can accept the rationality thesis. An externalist can accept that one who recognizes a moral fact (or makes a moral judgment) necessarily has a reason for action. But from that position, it follows trivially that one who recognizes a moral fact (or makes a moral judgment) necessarily is motivated, if he is responsive to the reasons that apply to him. Similarly, on this conception, we can't count irrationality as an interfering factor that prevents the manifestation of a disposition to be motivated — on pain of making vacuous the claim that someone has a disposition to be motivated in any case where the person has a reason for action.

On a familiar and intuitive conception of rationality, rationality is the absence of certain patterns of incoherence in one's attitudes and actions.[20] Examples include: believing a proposition and its negation; not intending to pursue what one believes are the means to one's ends; and intending to do something that one believes one cannot do. In the present context, this conception

of rationality yields the view that, necessarily, one who makes a moral judgment is motivated to act accordingly, unless a specific incoherence interferes. This view deserves to be considered a form of moral internalism, indeed a version of the dispositional thesis, because the role of the rationality condition is to provide for the possibility that some kind of incoherence may interfere with the manifestation of the agent's motivational dispositions.

On the face of it, it seems that one can deny what are intuitively conceptual connections without irrationality (understood as incoherence). For example, Burge's (1986, 263) character who judges that sofas are not pieces of furniture made or used for sitting on arguably makes a conceptual error, but need not be irrational. There need be no incoherence in his mental economy. Similarly, Alice, given her theoretical views about fairness, does not seem to be irrational in maintaining that an action is fair but having no motivation to take that action. Alice's combination of attitudes may involve conceptual error, but it is not intuitively incoherent. Her defect seems substantive, not formal. Similar points apply to parallel examples involving other moral concepts, including general moral concepts such as *right*.

On a traditional view of the analytic, it might seem promising to try to assimilate conceptual errors to cases of believing P and Not P. For example, on such a view, to judge that something is a sofa *is* to judge that it is a piece of furniture made for sitting on, etc. (assuming for the sake of argument that some such analysis of the concept of a sofa is correct). Since Burge's sofa theorist judges that a sofa is a sofa, he judges that a sofa is a piece of furniture made for sitting on. But he also judges that it is not the case that a sofa is a piece of furniture made for sitting on. Therefore, this view implies that Burge's sofa theorist is incoherent. An analogous argument could be constructed with respect to Alice.

The problem with this line of thought is that the traditional view of concepts on which it is premised is undermined by the type of example under discussion. If concepts have constitutive connections, we cannot assume that they are necessarily available to every thinker who has the concept.[21] It is possible to think that something is a sofa without thinking that it is a piece of furniture made for sitting on.

We can now see that some of the standard ways in which theorists have argued for the rationality thesis are undermined. For example, Michael Smith (1994) argues that the concept of an action's being right is in part the concept of there being a normative reason in favor of that action. He further analyzes the concept of a normative reason in terms of what one would desire to do if one were fully rational. From these two analyses, Smith draws the conclusion that one who judges that an action is right judges that he would desire to perform the action if he were fully rational. And, according to Smith, it is irrational to judge that one would desire to perform an action if one were fully rational and not to desire that action. But examples like Alice's show that even if Smith's conceptual analyses are correct, a person can judge that an action is right without judging

that he would desire to perform that action if he were fully rational. Therefore, one can judge that an action is right and not desire that action without falling into the state that Smith claims is irrational.

Of course, we could stipulate that making a conceptual mistake is constitutive of irrationality. If we use the term in this way, however, we should recognize that conceptual mistakes seem to involve a different kind of irrationality from the more familiar and intuitive one we have been considering. Given this different understanding of rationality, the rationality thesis is true (assuming a conceptual connection between moral judgments and motivation), but it says something very different from what it is usually taken to say. It would be more precise and informative to say that the connection between making a moral judgment and being motivated accordingly is conceptually required, but thinkers need not be disposed to satisfy that requirement.

Thinkers like Alice have false beliefs, but merely having false beliefs would not normally be considered irrational, and certainly not on the conception of rationality we are considering. Beliefs formed in certain ways may for that reason be irrational, for example beliefs formed because of wishful thinking or on the basis of certain kinds of fallacious reasoning. But there is no reason that thinkers like Alice must make errors of this kind. As philosophers well know, strange and mistaken conclusions can be reached without irrational thought processes.

I have developed a type of example that presents a serious challenge to the most natural (moral) internalist understanding of the rationality thesis. The more run-of-the-mill incomplete understanding cases mentioned above could be used in a similar way. Incomplete understanding of a concept need be no more irrational than a nonstandard theory and can have similar consequences. The discussion in section 4 of the nonmoral, normative concept of what one ought to do, all things considered, will further support the conclusion that Alice is not irrational on the incoherence conception of rationality. But I don't purport to have shown that no conception of rationality yields an interesting form of the rationality thesis that is immune to examples such as Alice's case.

3. A Solution in the Theory of Mental Content

The examples developed in the last section challenge internalism as it is standardly understood and therefore give us reason to revisit its interpretation. My proposal will be that one who recognizes a moral fact or makes a moral judgment is conceptually required to be appropriately motivated, but may fail even to have a disposition to satisfy that requirement. In this section, I want to show that this proposal — or, more precisely, a more general version of it — is independently motivated by considerations in the theory of content. I will begin by arguing that the pattern of difficulties encountered by attempts to capture a conceptual connection between morality and motivation is just a specific instance of a pattern to be found in the theory of mental content in general.

Before proceeding, I want to call attention to one complication in order to set it aside. In these discussions, it is common to move back and forth between the idea that concepts are individuated by connections between concepts and the idea that concepts are individuated by requirements on inferences or mental transitions more generally. The relationship between constitutive conceptual connections and conceptually required mental transitions is not straightforward, however. For one thing, as Gilbert Harman (1986, 11–19) has emphasized, relations of implication do not translate straightforwardly into requirements on inference. A different problem, which is of particular importance in the present context, arises once we extend the notion of conceptual role to include mental transitions that are not inferences, such as transitions to intentions or desires. Such transitions do not have a parallel in connections between concepts or propositions. If a connection between moral judgments and motivation is to be built into moral concepts, we therefore need to think in terms of the individuation of concepts by mental transitions. Since talk of constitutive conceptual connections is familiar and eases exposition, however, I will continue to use that way of talking where it is harmless. I will call the view that concepts are individuated by mental transitions the *network theory of concepts*.[22] My arguments will not depend on the specific content of requirements on mental transitions (or on any assumptions about particular correspondences between such requirements and constitutive conceptual connections).

It is by now a familiar claim that there are no inferences or judgments that a thinker who has a particular concept is necessarily disposed to make.[23] This *no-inference point* derives from an influential line of thought that goes back at least to Quine's argument that no beliefs are immune from revision.[24] Tyler Burge's well-known *sofa* example, which introduced the method that I relied on in developing the Alice example, is a good illustration. Burge (1986, 263) argues that a thinker who fully understands the term "sofa" and is fluent in the term's use could propose as a testable, empirical hypothesis, and could even come to believe, that sofas are not items of furniture made for sitting on, but religious artifacts or works of art that would not support a person's weight. Although Burge does not put the point in these terms, one who proposes the non-standard theory about sofas ipso facto has thoughts involving the concept of a sofa, though he is likely not disposed to make the judgments or inferences most plausibly individuating of the concept. Burge points out that similar examples can be developed for a very wide range of notions.[25]

As mentioned above, in contrast to examples of sophisticated thinkers who begin with full understanding and develop non-standard theories, other well-known examples involve more ordinary incomplete (or incorrect) understanding. It is often claimed that the explanation of such cases involves deference to other people who fully grasp the relevant concepts. As I argue elsewhere, however, in general, thinkers who defer to others do not thereby acquire the relevant inferential dispositions.[26] Therefore, even if deference accounted for all such cases of thoughts involving incompletely grasped concepts, it remains the case

that the thinkers in question have thoughts involving concepts without having dispositions to make the relevant concepts' canonical transitions. (In the non-standard theory examples, deference is not relevant because the theorists plainly do not defer to others; they know what others believe and have come to doubt it.)

The no-inference point has been extremely influential and is widely, though by no means universally, accepted. In my view, the point has been convincingly established, and I don't have space here to give arguments for it beyond the arguments involved in defending my examples. My goal in what follows is to draw out the implications of the point, rather than to defend it. It is worth noting, however, that the success of examples of the sort developed in the last section and the next one does not depend on the point.

If the no-inference point is correct, however, it provides strong support for the examples. Indeed, if it is correct, it would be surprising if a generalized version of the point did not hold for all mental transitions, as opposed merely to inferences. That is, it is difficult to see why transitions from beliefs to intentions or other motivational states would have a different status from transitions between beliefs.[27] If it is true that there are no transitions in thought that a thinker who has a particular concept is necessarily disposed to make, it follows that a thinker who has a particular moral concept need not have the appropriate motivational dispositions.

The no-inference point is widely believed to provide the basis for a devastating argument against conceptual role semantics and against the network theory of concepts. Jerry Fodor is perhaps the most influential advocate of this argument.[28] Fodor thinks that the no-inference point shows that there isn't a principled distinction between individuating and non-individuating inferences and therefore that concepts must not be individuated in terms of inferences or epistemic capacities more generally. He maintains that conceptual-role semantics collapses, and a covariation theory of content is needed.[29]

On the other hand, there are familiar and not-so-familiar reasons for thinking that concepts are individuated by constitutive connections.[30] I have space only to touch on a few. For example, there are the so-called analytic intuitions. As has often been pointed out, the standard, Quinean way of trying to explain them away, by claiming that intuitively analytic propositions are merely more "central" and therefore more difficult to give up than other propositions, is inadequate. The difference between what is involved in trying to give up, on the one hand, propositions such as *a person who descends the stairs goes down the stairs* or *squares have four sides* and, on the other hand, obvious ordinary propositions such as *people sometimes eat on Tuesdays* or *there have been black dogs* seems not to be that the former are harder to give up, indeed it seems not to be merely, if at all, a difference of degree.

Most importantly, without the resources of constitutive connections between concepts, it is difficult to account for so-called Frege cases — cases in which content apparently cuts more finely than the level of reference.[31] The problem is familiar, but, in light of the no-inference point, we can see that it is far more extensive than is typically understood.

Consider the pairs of concepts *vitamin C* and *ascorbic acid* and *water* and *H2O*. The concepts in each pair are, let us assume, necessarily co-referential. Yet a thinker can think, for example, that vitamin C prevents colds without thinking that ascorbic acid prevents colds. A theory of content that individuates concepts at the level of reference faces a dilemma in trying to account for this phenomenon.

If the concepts are atomic — not composed of constituent concepts — the theory cannot distinguish a thinker's having a thought involving the concept *vitamin C* from a thinker's having a thought involving the concept *ascorbic acid*.

The obvious solution is to hold that the concepts are complex; although the complex concepts are necessarily co-referential, they are different concepts because they are composed of different constituent concepts. For example, the concept *vitamin* is a constituent of the concept *vitamin C,* but not of *ascorbic acid*. In this way, it is often thought that theories that individuate concepts at the level of reference can allow for constitutive connections between concepts by maintaining that many or most concepts that are expressed by more than one word are complex.

It seems not to have been noticed, however, that the examples that are supposed to show that there are no constitutive connections between concepts work equally well against multi-word concepts. For example, a thinker could develop a theory according to which vitamin C is not a vitamin or a theory on which ascorbic acid is not an acid. The reply that vitamin C is, by definition, a vitamin is no better than the argument that sofas are, by definition, items of furniture made for sitting on. (After all, the morning star turned out not to be a star, and the naturalistic fallacy is not a fallacy.) If such arguments show that concepts like *sofa* or *fairness* lack constitutive connections, then they show that concepts like *ascorbic acid* and *H2O* lack constitutive connections and therefore are not complex.

Finally, although I cannot argue the point here, covariation theories — the most prominent and well-developed theories of content that attempt to do without constitutive connections between concepts — face a problem parallel to the problem that the no-inference point poses for conceptual role semantics. As a consequence of coming to believe a non-standard theory, a thinker can lose the disposition to apply a concept to the objects that fall under that concept.

In sum, there seem to be various phenomena that constitutive connections between concepts are well suited to account for. However, thinkers seem to be able to have concepts without having the dispositions to make those concepts' putatively canonical mental transitions. If we conclude that there are no constitutive connections between concepts, we deprive ourselves of the resources to account for the phenomena we began with.

The case of moral concepts presents a specific instance of this pattern. Theories of content that individuate concepts no more finely than the level of reference cannot distinguish having a moral concept from having a nonmoral, purely descriptive concept with the same reference. (Given the supervenience of

the moral on the natural, every moral concept will have some natural property, perhaps gerrymandered or disjunctive, as its reference.) If we appeal to canonical mental transitions to distinguish moral concepts from purely descriptive ones, for example to account for the internalist insight that there is a conceptual link between moral judgments and motivation, we run into the problem that thinkers seem to be able to have moral concepts without having dispositions to make the putatively canonical mental transitions. If we conclude that there are no canonical transitions, we deprive ourselves of the resources to explain the ways in which moral concepts differ from descriptive ones.

I have elsewhere argued that the right response to the no-inference point is to give up conceptual role semantics but to retain the network theory of concepts.[32] In order to make this option visible, we need to disentangle the two views. The network theory is a view about the nature of concepts, the components of the contents of thoughts. By contrast, conceptual role semantics is a view about what it is for a thought to involve a particular concept (and the closely related question of what it is to have a concept) that implies or presupposes the network theory of concepts. (I mean to remain neutral here on whether an account of the nature of concepts or an account of what it is to have a concept is prior in the order of explanation.) As I have mentioned, it is apparently widely taken for granted that, if concepts are individuated by requirements on mental transitions, then to have a concept is to be disposed to make those mental transitions. We can formulate this presupposition in the terminology we have introduced:

LINK: If the network theory of content is true, then conceptual role semantics is true.

If we accept the no-inference point, then conceptual role semantics must be abandoned as either false or based on a false presupposition. If there are no non-trivial inferences that a thinker must make in order to have a concept, it can't be the case that having a concept is making the concept's canonical mental transitions.

Given LINK, the no-inference point also implies that the network theory is false. But in light of the no-inference point, LINK is at best vacuously true. Why not give up LINK — in which case the no-inference point is consistent with the network theory?

To elaborate, my suggestion is that concepts are individuated by requirements on mental transitions. A thinker who has a thought involving a given concept is *subject* to a requirement that he make its canonical mental transitions, but need not be *disposed* to satisfy that requirement. By itself, the suggestion that a thinker can be *subject to* a requirement without being *disposed to* satisfy it should not be surprising. In general, it is plausible that one must have certain minimum capacities in order to be subject to a requirement, but one need not have a disposition to satisfy the requirement. For example, people who lack a disposition to be honest are nevertheless subject to moral standards of honesty.

Similarly, it has become a commonplace that certain dispositions to irrational behavior are widespread. As this use of the term "irrational" implies, it does not follow that one who has such a disposition is not subject to the corresponding standard of rationality.[33]

But, it may be objected, what could make it the case that a thinker's thought involves a particular concept whose canonical mental transitions the thinker is not disposed to make — as opposed, for example, to a different concept whose canonical mental transitions the thinker is disposed to make? In other words, the objection is that if we reject conceptual role semantics, but retain the network theory, we are left without an account of what determines content. Even if we grant that a *necessary condition* for having a concept is being subject to certain requirements, we still need an account of *what makes it the case* that a thought involves a particular concept.

First, dispositions are not the only candidates for determinants of content. For example, social environment, evolutionary selection for a particular function, and the structure of the world have been suggested as determinants of content. And we cannot assume that the determinants of content, whether dispositions or not, determine content in the straightforward way envisaged by conceptual role theories of content. The function from determinants of content to content may not be intuitive or transparent.[34]

Second, one possibility is that what determines that a given mental representation expresses a particular concept is that a specification of the concept's constitutive connections, encoded at a level inaccessible to the thinker, plays some kind of fundamental role in explaining the thinker's overall deployment of the mental representation. The explanatory relation need not be so simple or direct that it guarantees a disposition to make the concept's canonical transitions. Of course, this speculative suggestion helps itself to the notion of representation; it is not meant to address how content is ultimately determined. The point is just that the basis for a principled distinction between individuating and non-individuating mental transitions may be located at a less superficial level than LINK would have it.[35]

Third, a very different possibility is that what makes it the case that a thinker's thought involves a particular concept is in part that the thinker is subject to a requirement that he make certain mental transitions—those that are constitutive of the concept. On this account, a thinker's being subject to certain requirements is part of the explanation of content, rather than a consequence of it. In order to develop such a view, more needs to be said about what it is in virtue of which a thinker is subject to the relevant requirements. The thinker's dispositions will no doubt be part of this story.

Finally, for present purposes, it is possible to resist the demand for a complete reductive account of what it is to have a particular concept. It may be too much to expect such an account.[36] At any rate, as discussed above, it is not as if retaining LINK (and therefore rejecting the network theory of concepts) opens the way to an unproblematic account of content.

The crucial point is that in order for there to be a principled distinction between individuating and non-individuating mental transitions, it needs to be the case that there are facts that determine that a particular concept, individuated by certain mental transitions, figures in the content of a thought. It does not need to be the case that a thinker who has the concept must be disposed to make (or recognize as correct) the individuating mental transitions.

In sum, the no-inference point supports rejecting conceptual role semantics but not rejecting the network theory of concepts. We can retain the resources of the network theory, without the dubious claim that thinkers must be disposed to make their concepts' canonical transitions in thought.

Suppose that I am right about the lessons to be drawn in the theory of content. That is, in order to have a thought involving a concept, there are no non-trivial mental transitions that a thinker must be disposed to make. The thinker must, however, be subject to requirements that are constitutive of the concept. What are the implications for the case of moral concepts?

First, all versions of internalism according to which those who make moral judgments must be disposed to be appropriately motivated must be false. Since being disposed to move from a moral judgment to an intention or other motivational state is a disposition to move between mental states, such versions of internalism are inconsistent with the proposition that there are no mental transitions that a thinker must be disposed to make.

Second, on the other hand, the implausibility of the claim that moral facts necessarily motivate provides no argument against there being a conceptual connection between morality and motivation. Since there can be conceptual connections without thinkers having the corresponding dispositions, one cannot argue against conceptual connections by pointing to the absence of the relevant dispositions. Thus, a common form of argument for externalist views is undermined. Conversely, one cannot argue from a conceptual connection to motivation to the conclusion that, for example, there cannot be an amoralist who uses moral concepts.

The proposal — that, in order to have a thought involving a concept, a thinker must be subject to requirements that are constitutive of the concept — seems custom-made to reconcile the insights of internalists and externalists. It explains how it can be that the connection between morality and motivation is built into moral concepts, yet people can recognize moral facts and make moral judgments without even being disposed to be motivated. The canonical transitions for moral concepts include transitions to motivational states. One who recognizes a moral fact or makes a moral judgment is therefore subject to a conceptual requirement to be appropriately motivated. But not only does she not need to be actually motivated in a particular case, she does not need even to be disposed to be motivated. A view that is independently motivated by considerations in the theory of content yields a tidy account of the relation between moral judgments and motivation.

Moreover, the present proposal is not subject to some prominent objections. Mackie (1977, 38–42) famously argued that moral facts that necessarily motivate would be "queer." On my proposal, it is not the case that moral facts must motivate anyone who recognizes them. Are there other consequences of the proposal that are metaphysically queer? Moral facts are such that anyone who recognizes them is subject to a conceptual requirement to be motivated accordingly. But one who recognizes a moral fact or makes a moral judgment is using a moral concept, so it is not strange that he is subject to conceptual requirements. And we can understand, without postulating any strange metaphysics, that a conceptual requirement could require a thinker to move to a motivational state rather than a belief.

The claim that moral facts necessarily motivate has also been criticized on the ground that it is inconsistent with the so-called Humean theory of motivation, according to which only contingent desires motivate, and what desires one has is independent of what beliefs one has. Is the present proposal inconsistent with the Humean theory of motivation? The proposal does not have the consequence that beliefs must motivate regardless of one's contingent desires.

It is true that, as I am using the term "mastery of a concept," a thinker who makes a moral judgment will necessarily have a disposition to be motivated to act accordingly, if he has mastery of the relevant concept or concepts. (As noted in note 8, for want of a better term, I say that a thinker has "mastery of a concept" just in case the thinker has the concept *and* is disposed to make its canonical mental transitions.)

Since having mastery of a moral concept consists in part of having certain dispositions to be motivated — roughly, dispositions to have desires — and since it is a contingent matter whether a thinker who makes a moral judgment has mastery of the relevant concepts, there is no claim here that a belief or judgment must motivate the thinker to action independently of the thinker's desires. And to see that the Humean theory does not rule out the possibility that one could have a disposition to move from certain beliefs or judgments to certain desires, it is sufficient to see that the disposition (or the mechanism that underwrites it) might itself be constituted in part by a desire. Similarly, the present proposal need not claim that one's beliefs, for example about what is a reason for action, can by themselves create a disposition to be motivated to act.

4. Non-moral Normative Concepts

In this section, I will briefly sketch an example involving the normative but non-moral concept of what one ought, all things considered, to do. The example illustrates that the arguments of this paper can be generalized to attack the view that, necessarily, one who judges that he has reason to take a particular action or that he ought to take a particular action is disposed to be motivated to do so. The points generally apply *a fortiori* to moral internalism.

Bernard is a moral philosopher who has recently come to hold a strange combination of views. He has always believed that it is possible and rational to make judgments about what one ought, all things considered, to do. Moreover, he has regularly made such judgments and has had a normal tendency to be motivated to act on them. He continues to believe that it is possible and rational to make judgments about what one ought to do. He hypothesizes, however, that it is irrational to act on all-things-considered judgments about what one ought to do that are based on considerations that cannot be reduced to a single type of value.

Bernard takes his hypothesis seriously and, as a result, loses his general disposition to act on judgments about what he ought, all things considered, to do. This loss does not leave Bernard paralyzed, however. He continues to act on certain all-things-considered judgments, and he continues to be disposed to act on judgments about what he has more specific reason to do, for example judgments about what would be generous, or prudent, or morally good. He also goes on making judgments about what he ought, all things considered, to do, though he treats some of them as having purely theoretical interest.

We can flesh out Bernard's view somewhat. He has an elaborate theory about the differences between the principles of rationality that apply to beliefs and those that apply to intentions and other motivational states. This theory has the consequence that, although one can coherently make theoretical judgments that require comparisons between ineliminably plural values, rational action must be based on a unified system of values — a system that can be reduced to a single fundamental value. Therefore, it is irrational to base intentions and other motivational states on those all-things-considered judgments that require weighing ineliminably plural values. This consequence of the theory gives Bernard serious pause, but he believes that his theory is strongly supported on other grounds. While he is working out whether he ought to revise his theory or accept the consequence, he suspends judgment on whether it is rational to act on judgments about what one ought, all things considered, to do that are based on plural values — with the consequence that he loses the corresponding disposition. It is certainly possible for people to hold views as strange as Bernard's, which after all are no stranger than many other positions that philosophers have held.

Before developing his theory, Bernard clearly made judgments involving the concept of what one ought to do, all things considered. By our usual standards, we would continue to attribute such judgments to him even after he entertained his hypothesis and his dispositions changed accordingly.

The possible objections to this example are similar to those considered above with respect to Alice's case, so I won't repeat that discussion. We have also seen that there is a much more general case for the conclusion that thinkers in the type of example under discussion can continue to have the concepts in question.

With respect to rationality, Bernard looks very different from Alice or from someone who thinks that sofas are not pieces of furniture made for sitting on. Bernard judges that an action is all-things-considered required but he has no

motivation to perform that action, not even other things being equal. This looks like classic *akrasia*, which is often taken to be a paradigm of irrationality.[37] Wedgwood (2004, 408) explicitly argues that one who judges that an action is all-things-considered required, and yet does not intend to take that action is irrational on the ground that he has "an incoherent combination of mental states: one's judgments and one's will are in conflict with each other; one's will refuses to pursue a course of action that one judges that one ought to pursue." Bernard's case thus illustrates that conceptual errors can lead to patterns of attitudes that are typically taken to be irrational.

I'm going to suggest that, on closer inspection, there is a good case that Bernard should not be counted as irrational, though I ultimately do not think that the conception of rationality under consideration is rich enough to yield a determinate verdict. The discussion should reinforce the already strong argument with respect to the easier case of moral concepts.

Unlike the standard cases of *akrasia* discussed in the literature, there is nothing weak or defective about Bernard's will. For example, it is not that his will is overcome by a conflicting force, such as a powerful desire, or sapped by exhaustion, intoxication, depression, or other illness. Bernard lacks a disposition to act precisely because he *has* succeeded in bringing his motivations into line with his unorthodox views. Thus, Wedgwood's explanation of why the combination of attitudes standardly taken to constitute *akrasia* is irrational does not apply.

Although Bernard does not suffer from weakness of will, and in fact succeeds in bringing his motivations into line with his theoretical views, it cannot be denied that the result is a combination of attitudes that we would ordinarily consider to be incoherent. Bernard's case thus raises the question whether, in certain cases, paradigmatically irrational patterns of attitudes may not be irrational in light of the way in which they come about.

Examples developed (for a different purpose) by Timothy Williamson are illustrative. The examples involve thinkers who, because of unusual theoretical views, deny propositions or are not disposed to make inferences that are plausibly thought to be constitutive of the meanings of logical terms or concepts.[38] Because the examples involve logical concepts, the resulting attitudes are ones that would standardly be considered paradigmatically irrational. One example involves two theorists who, for different reasons, deny that every vixen is a vixen. On first learning that someone denies that every vixen is a vixen, one might well consider that person *ipso facto* convicted of irrationality. But Williamson's theorists have well-thought-out theoretical views that lead them to their bizarre-sounding conclusion. For example, one theorist holds the view that universal quantification is existentially committing. He therefore believes that a necessary condition for the truth of the proposition is that there be at least one vixen. Since misleading evidence has led him to the empirical belief that there are no foxes, he believes that it is not the case that every vixen is a vixen. Once we understand the reasoning, it is not at all obvious that there is a non-question-begging sense in which the theorist is irrational.

A preliminary point is that it cannot be *obvious* that the intuitively irrational patterns of attitudes are in fact *per se* irrational. It certainly is not obvious that the unusual theorists are wrong. But if one of the unusual theorists turned out to be correct, we would have to conclude that the pattern of attitudes that he exemplifies is not even incoherent, let alone irrational. In that case, *no one* would be irrational in virtue of exemplifying that pattern of attitudes.

More importantly, the examples help to bring out several related ways in which the non-standard theorists differ from typical cases of people who are irrational in virtue of the corresponding patterns of attitudes. First, the theorists are not incoherent or irrational by their own lights, while in typical cases, people have at least an implicit understanding that would condemn the general pattern of attitudes that they exemplify. Second, the theorists deliberately achieve the combination of attitudes in question on the basis of reasons; they follow where reason leads them. Third, if the relevant combinations of attitudes are incoherent (and if it is incoherent to have a disposition to produce an incoherent combination of attitudes), then, given their theoretical views, the theorists are not able to avoid incoherence. And they have achieved the least bad of the possible incoherences (assuming that it is worse to fail to bring one's dispositions with respect to first-order attitudes into coherence with one's second-order beliefs about those attitudes than to have dispositions to produce incoherent combinations of first-order attitudes). These points apply *a fortiori* to Alice and other cases involving moral concepts.

These points make a case that, whether or not we consider him incoherent, Bernard is importantly different from the examples with which I introduced the conception of irrationality that we are considering. We would need a richer theoretical understanding of the relevant conception, however, in order to reach a determinate verdict on whether Bernard is irrational. The examples at very least suggest that, in order to diagnose irrationality, we may need to look not just at a pattern of attitudes but at the explanation of that pattern.

5. Conclusion

I first argued that standard versions of moral internalism are untenable in light of a type of example that has not, to my knowledge, previously been considered in metaethical discussions. As a consequence of having unusual views, a good-willed thinker could make moral judgments and recognize moral facts without having a disposition to be motivated to act accordingly (and without believing himself to have reasons for action). Moreover, on a familiar conception of irrationality as incoherence, the thinker in question need not be irrational, and it is not clear that there is a relevant conception of rationality on which he must be irrational. The first part of the paper therefore strengthens some of the most important arguments that have been taken to undermine internalism and support externalism.

Next, however, I argued that these arguments seem to support externalism over internalism only because it is widely taken for granted that, if concepts are individuated in part by mental transitions, then a thinker cannot have a particular concept without having a disposition to make the concept's individuating transitions. Considerations in the philosophy of mind, including the general type of example deployed in the first part of the paper, support rejecting that presupposition. Rejecting the presupposition makes visible a different understanding of internalism, one that is not vulnerable to some of the most influential arguments that have been taken to support externalism. According to the proposed position, a thinker who makes a moral judgment or recognizes a moral fact is subject to a conceptual requirement that he be motivated to act accordingly, but need not be disposed to satisfy that requirement. The second part of the paper thus argues that considerations independent of metaethics yield a position that neatly reconciles important internalist and externalist insights. Finally, in the last part of the paper, I began the project of extending the arguments to internalism about reasons for action. I gave reasons for thinking that patterns of attitudes that are typically taken to be paradigmatic of irrationality — for example, believing that one ought, all things considered, to take a particular action but not being motivated to do so — may not be irrational if they are the product of unusual views.

Notes

1. This article is a descendant of a paper titled "Ethics and Concepts," originally written in 1992–93 as an Oxford D.Phil. thesis chapter, but ultimately not included in the thesis. For comments on the original paper, I owe thanks to Roger Crisp, Martin Davies, Ronald Dworkin, Noah Feldman, Kinch Hoekstra, Philip Pettit, Michael Smith, Nicos Stavropoulos, Galen Strawson, and Bernard Williams. For comments on this version, I am grateful to Gil Harman, Barbara Herman, Ram Neta, and Georges Rey. I'm especially indebted to my colleagues Pamela Hieronymi, A. J. Julius, and Seana Shiffrin for their close reading and thoughtful feedback.
2. See, for example, Smith (1994, 6–7, 60). Internalism about moral facts is standardly distinguished from internalism about moral judgments. Most of the arguments in this paper, because they concern moral concepts, apply to both forms of internalism. In making a moral judgment or recognizing a moral fact, a thinker must use moral concepts. Therefore, arguments that apply to agents in virtue of their using moral concepts apply both to those who make moral judgments and those who recognize moral facts. To save words, I will often write simply "moral facts" or "moral judgments," but unless I specify otherwise, I mean my remarks to apply to both. On why moral internalism requires that the connection between morality and motivation be conceptual, see page 145. The term "motivation" is sometimes used to refer to normative reasons for action — considerations that support taking or not taking a particular course of action. This is not how I will use the term. In this paper, motivational states

are psychological states that are involved in the production of action. Typical examples include conative states, such as intentions and desires, and also affective states, such as fear or disgust.

3. By "motivated accordingly," I mean motivated in the appropriate direction, i.e., in favor of or against the action in question, depending on whether the relevant concept is a positive or negative moral concept.
4. See, for example Brink (1986, 29–31).
5. Concepts are components of the contents of thoughts. We can avoid delicate issues concerning what notion of ability is apt here since the issue of whether a thinker who never has thoughts involving a concept nevertheless has the concept will not be relevant. For our purposes, it will be sufficient that, if one has thoughts involving a concept, one has the concept.
6. Smith (1994, 76). For other examples, see McDowell (1978, 22–23), Brink (1986, 29–31).
7. This paper focuses on cognitivist accounts of moral judgments. When we make moral claims, we seem to be making claims that can be straightforwardly true and false. Ultimately, it may be that the pre-theoretical appearances cannot be preserved, but, in the first instance, we should aim for a cognitivist account.
8. I will say that someone has mastery of a particular concept just in case he has the concept and has a disposition to make the concept's canonical mental transitions. This usage may be misleading because one possibility is that one who lacks such personal-level mastery of a concept may nonetheless have some kind of sub-personal representation of the concept's canonical mental transitions that is in no way deficient. See page 153. I use the term in part for lack of a better one. Also, I will mostly ignore the possibility of a more intellectual kind of mastery that involves the ability to *recognize* that the relevant mental transitions are correct, as it will not affect my arguments. See page 145–146 for brief discussion.
9. The qualification "if the question of the thinker's taking that action arises" is needed because, for example, one can make moral judgments about actions that one is not in a position to take. One might judge that a particular action of Napoleon on the battlefield was courageous. I will usually omit the qualification. This kind of problem is sometimes dealt with by restricting the requirement that a thinker be motivated to the thinker's judgment that *her* taking a certain action would have a certain moral property. In my view, this is unduly restrictive. Whatever motivational requirement applies to one who judges that her taking a particular action would be wrong applies also to one who judges that another person's taking a particular action would be wrong — at least if the question of the thinker's taking that action arises. (Indeed, I think that the restriction to judgments about actions is too narrow. Motivational requirements plausibly apply mutatis mutandis to judgments about states of affairs, character traits, and so on.) I will often talk simply of moral judgments, by which I mean specific judgments that an action falls under a moral concept. But nothing in this paper will turn on these issues. Those who think that the relevant motivational requirements plausibly apply only to the thinker's judgment that her taking a particular action would be, for example, wrong, can simply read my arguments as restricted accordingly. Also, in order to avoid having to write of taking an action under certain circumstances, I am using "an action" to mean an action type, where an action type is individuated in part by the circumstances.

10. The term "internalism" is often used for the strict thesis, the dispositional thesis, or the rationality thesis. As I will explain, there is a way of understanding the idea that moral judgments are internally or necessarily connected to motivation that does not entail any of these theses. I therefore prefer to use "internalism" for this more abstract idea, but of course nothing will turn on this terminological choice.
11. I draw on the method of constructing examples introduced by Burge (1986).
12. Points similar to the ones I make here could be made with respect to Alice's affirmative doubt that fairness is a reason for action or with respect to her suspension of the belief that it is.
13. Now it may be that once the disposition to perform fair actions is in place, the causal chain from the judgment that an action is fair to the motivation to take that action does not go through the belief that fairness is a reason for action. The mechanism may be more automatic and less intellectual than that. We know, however, that Alice's loss of the belief resulted in the appropriate motivation not being formed, not merely its not being effective in producing action. The question, therefore, is whether the loss of the belief could have this effect by coming, over time, to damage the mechanism that underwrites the disposition or whether it must do so by removing some other necessary condition for the motivation to be produced. There is no reason to think that the mechanism of the disposition in question must be encapsulated against the effects of changes in belief.
14. I spell this out in the text below, page 145.
15. In my view, this example does not work because it is not even true, let alone constitutive of the concept, that arthritis can only be in the joints. I use the example because its prominence makes it an effective way of bringing to mind a large class of cases.
16. See pages 149–150 for a brief discussion of the most important complication.
17. For a similar argument, see Wedgwood (2004, 414).
18. To flesh this out slightly, the idea is that a concept is individuated by its role in reasoning. A disposition to make a concept's canonical transitions is, roughly, a disposition that constitutes an implicit recognition of the concept's individuating role. See also the second paragraph of section 3.
19. Ralph Wedgwood offers a theory that he labels a conceptual role theory of moral terms. It is unclear to me, however, whether the theory is a conceptual role theory as I am using the term. Specifically, I'm not sure whether, according to the theory, what makes it the case that a thought involves a moral concept is that the thinker is disposed to make (or recognize as correct) the concept's canonical transitions. On the one hand, he holds that the meaning of a term is given by the basic rules "governing its use" — in my terms, canonical mental transitions. (2001, 7–8) He holds that one "follows" these rules only if, roughly speaking, one is disposed to make the term's canonical mental transitions. (2001, 7) If what makes it the case that a rule governs the use of a term is that one follows the rule in one's use of the term, then the account is a conceptual role theory as I have defined it, and is vulnerable to the type of example I have developed. On the other hand, Wedgwood holds that to understand a term, "in effect . . . one must master [the basic rules governing its use] (that is, one must have the ability to follow these rules)." (2001, 8–9) In a footnote, he adds

that he "strongly suspect[s] that the notion of "mastering a rule" can only be explained in partly normative terms." (2001, 9, n. 17) I am uncertain, however, what role the notion of understanding a term — and therefore the notion of mastering a rule in terms of which the former notion is explained — plays in Wedgwood's account of what determines meaning or content. It is possible that Wedgwood has in mind the kind of account I describe toward the end of section 3, according to which normative facts play a role in making it the case that a thought involves a particular concept. The crucial point for present purposes is that even if Wedgwood's theory holds that normative facts are part of what makes it the case that a thought involves a particular concept, the theory does not escape the present critique if a necessary condition for a thinker's having a thought involving a particular concept is the thinker's having the concept's canonical dispositions.

20. In a recent paper, Wedgwood (2004) appeals to this conception of rationality to defend a form of internalism about reasons for action, as opposed to moral internalism. As Wedgwood says, in addition to the "synchronic" requirements that one avoid such incoherent patterns of attitudes, there are also "diachronic" requirements according to which "one should follow appropriate rules or procedures . . . in the process of forming and revising one's beliefs and intentions." (2004, 407) These latter requirements will be less relevant to our discussion than the synchronic ones because the strongest case that thinkers such as Alice are irrational is that they have incoherent patterns of attitudes. As I say in the text, one need not be irrational in one's reasoning in order to reach an unusual and false theory such as Alice's.
21. See section 3 for elaboration of this point.
22. The view would more naturally be called the "conceptual role view of concepts," but that usage would be extremely confusing because it will be important to my argument to distinguish the view from conceptual role semantics.
23. In order to avoid issues raised by rather trivial inferences, such as the inference from a's being an X to something's being an X, we could give a more cautious statement of the position, which is sufficiently strong for our purposes: in order to have a given concept, a thinker need not be disposed to make any of the potentially individuating inferences. For convenience, I will stick with the simpler formulation of the position in the text.
24. See, for example, Quine (1953); Mates (1952); Burge (1979; 1986); Fodor (1998a; 1998b). For a recent discussion, see Williamson (2007, chapter 4). I discuss some of Williamson's examples briefly in section 4.
25. Burge (1986, 709).
26. Greenberg (forthcoming). I also argue that deference to other people cannot bear the explanatory burden placed on it.
27. I've elsewhere argued that an analogous point applies to dispositions to apply concepts to perceptibly presented objects.
28. Fodor and Lepore (1992, chapter 6); Fodor (1998a; 1998b; 2000).
29. Of course, there is much more to Fodor's argument. For a brief summary and reply, see Greenberg and Harman (2006, section 5.2). For his own version of a covariation theory, see Fodor (1990). Like Fodor, Williamson (2007, chapter 4) draws the conclusion that understanding a word or concept does not require being disposed to make any particular judgments or inferences. He makes this

argument in the context of attacking epistemic accounts of analyticity. He seems also to conclude that the identities of word meanings and concepts are not constituted by constitutive connections, for he denies that there is a kind of understanding of a word or concept that consists in recognizing, implicitly or explicitly, its constitutive connections (2007, 123–133).

30. For a brief summary, see Greenberg and Harman (2006, section 5.1). For more detail, see Rey, (1993; 2005).
31. Similar points apply with respect to necessarily empty concepts and concepts of properties that no one is reliably able to track.
32. Greenberg (2001; 2005; forthcoming).
33. It might be thought that the maxim that "ought implies can" supplies the basis for an objection. But, as the examples in the text show, the sense in which one who lacks a disposition to do something cannot do it is not the sense of "can" in which the maxim is plausible.
34. For discussion, see Greenberg (2001); Horwich (2005, 65–78).
35. I am grateful to Georges Rey for many rich discussions of this topic over the years. See Rey (2009).
36. See, e.g., Burge (1986, 718–720).
37. As Seana Shiffrin pointed out to me, for the defender of rational internalism to take *akrasia* to be irrational may beg the question against some Humean opponents. But in order to make the objection as strong as possible, let us grant the premise.
38. Williamson (2007, chapter 4).

References

Brink, D. 1986: "Externalist Moral Realism," *Southern Journal of Philosophy* Supplement: 23–42.

Burge, T. 1979: "Individualism and the Mental," *Midwest Studies in Philosophy* 4. Minneapolis: University of Minnesota Press, pp. 73–121.

——— 1986: "Intellectual Norms and Foundations of Mind," *Journal of Philosophy* 83: 697–720.

Fodor, J. 1990: "A Theory of Content, II: The Theory," in J. Fodor, *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

——— 1998a: *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

——— 1998b: "Review of Christopher Peacocke's *A Study of Concepts*," reprinted in *In Critical Condition*, Cambridge, MA: MIT Press.

——— 2000: *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, MA: MIT Press.

Fodor, J. and Lepore, E. 1992: *Holism: A Shopper's Guide.* Cambridge, MA: Blackwell.

Greenberg, M. 2001: *Thoughts without Masters: Incomplete Understanding and the Content of Mind*. University of Oxford, D.Phil. Dissertation.

——— 2005: "A New Map of Theories of Mental Content: Constitutive Accounts and Normative Theories," *Philosophical Issues* 15: 299–320.

——— forthcoming: "Incomplete Understanding: Deference and the Content of Thought," available at http://ssrn.com/author=336071.

Greenberg, M. and Harman, G. 2006: "Conceptual Role Semantics," in E. Lepore and B. Smith, eds., *Oxford Handbook of Philosophy of Language*, Oxford: Oxford University Press.

Harman, G. 1986: *Change In View: Principles of Reasoning*. Cambridge, MA: MIT Press.
Horwich, P. 2005: *Reflections on Meaning*. Oxford: Oxford University Press.
Jackson, F. and Pettit, P. 1995: "Moral Functionalism and Moral Motivation," Philosophical Quarterly 45: 20–40. Reprinted in Jackson, F., Pettit, P., and Smith, M., eds., *Mind, Morality, and Explanation*. Oxford: Oxford University Press (2004, pp. 190–210).
Mackie, J. L. 1977: *Ethics: Inventing Right and Wrong*. London: Pelican. Reprinted, London: Penguin (1990).
Mates, Benson. 1952: "Synonymity" in L. Linksy, ed., *Semantics and the Philosophy of Language*. Urbana, IL: University of Illinois Press.
McDowell, J. 1978: "Are Moral Requirements Hypothetical Imperatives?" *Proceedings of the Aristotelian Society* Supplementary Volume: 13–29.
Quine, W. V. 1953: "Two Dogmas of Empricism," in W. V. Quine, *From a Logical Point of View*. NY: Harper & Row (1963).
Rey, G. 1993: "The Unavailability of What We Mean," *Grazer Philosophische Studien* 46: 61–101. Reprinted in Fodor, J., ed., *Holism: A Consumer Update*. Amsterdam: Rodopi (1993).
——— 2005: "Philosophical Analysis as Cognitive Psychology: the Case of Empty Concepts," in H. Cohen and C. Lefebvre, eds., *Handbook of Categorization in Cognitive Science*. Oxford: Elsevier, pp. 71–89.
——— 2009: "Concepts, Defaults, and Internal Asymmetric Dependencies: Distillations of Fodor and Horwich," in N. Kompa, C. Nimtz, and C. Suhm, eds., *The A Priori and Its Role in Philosophy*. Paderborn: Mentis
Smith, M. 1994: *The Moral Problem*. Oxford: Blackwell.
Wedgwood, R. 2001: "Conceptual Role Semantics for Moral Terms," *The Philosophical Review* 110.1: 1–30.
——— 2004: "The Metaethicists' Mistake," *Philosophical Perspectives*, 18, *Ethics*.
Williamson, T. 2007: *The Philosophy of Philosophy*. Oxford: Blackwell.