TRANSPARENCY IN METAETHICS

by

Sarah Raskoff

Copyright © Sarah Raskoff 2018

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PHILOSOPHY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2018
As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Sarah Raskoff, titled Transparency in Metaethics and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Juan Comesaña

Date: (11/30/2018)

Stewart Cohen

Date: (11/30/2018)

Mark Timmons

Date: (11/30/2018)

Final approval and acceptance of this dissertation is contingent upon the candidate’s submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Juan Comesaña
Professor
Philosophy

Date: (11/30/2018)
STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Sarah Raskoff
ACKNOWLEDGMENTS

There are so many people whose support made this dissertation possible. First, I wish to thank my advisor Juan Comesaña, for the encouragement and inspiration to write these papers. Not everyone has a dissertation chair that is willing to meet every week and is happy to read a continual supply of very rough drafts. You’re the best! Thanks also to the other two wonderful members of my committee, Stew Cohen and Mark Timmons, and to Jay Odenbaugh, Connie Rosati, Dave Schmidt, and Jason Turner for encouragement and helpful feedback on previous drafts of these papers. Thanks to my mom, Carol Tabak, and dad, William Raskoff, for supporting and believing in me, and for always being interested in my papers and projects. I am endlessly thankful for you two and feel lucky to have such amazing parents. And finally, thank you to Jacob, who believed in me when I stopped believing in myself. I could not have completed this project without his endless support, intelligence, kindness, and patience.
CONTENTS

1. Abstract 6
2. Introduction 8
3. The Transparency Fallacy in Metaethics 14
4. Getting Expressivism Out of the Woods 39
5. Does Internalism Support Noncognitivism? 76
6. Complete Works Cited 105
ABSTRACT

Many of us pride ourselves on our moral credentials. But, while we may be confident in our views about what is right or wrong, when we take a step back to consider our moral views more generally or abstractly, we recognize that there are many features of our moral thought and talk that moral competence does not shed light on. Being able to answer questions about what kind of person we ought to strive to be, or what actions we are forbidden from performing does not mean that we are able to answer metaethical questions about the metaphysical, epistemological, semantic, and psychological views, assumptions, and commitments that underlie these practices. Moral competence does not guarantee true beliefs about metaethical issues, or competence to engage with such issues at the level of abstraction this kind of engagement actually requires. Answers to metaethical questions are often opaque to us rather than transparent: we can participate competently in our moral practice without having true beliefs about its metaethical underpinnings; indeed, we can participate without having any such beliefs at all.

Metaethicists have lots of beliefs about the metaethical underpinnings of our moral practices. But for all their successes, metaethicists often fail to pay attention to the ways that failures of transparency matter to metaethical theorizing. Often, in the course of defending or arguing against a particular metaethical theory, metaethicists assume that the claims made by that theory would have to be transparent to competent moral speakers if they were true: that if the theory is right that something is true of our moral thought and talk, then the theory is also committed to those who are competent with moral thought and talk believing that that something is true. I call this mistake “the transparency fallacy.” In the first paper of my dissertation, I
provide a clear, general formulation of the transparency fallacy, which is assuming, without further argument, that if p, then competent moral speakers and thinkers believe that p, where p is some putative metaethical fact. In the second paper of my dissertation, I consider a particular commission of the transparency fallacy made by Jack Woods in his recent argument against expressivism. And in the third paper of my dissertation, I consider a case where a failure of transparency leads to a dialectical impasse between cognitivists and expressivists. My dissertation represents only a first attempt to tackle some very large questions in metaethics and metametaethics (to borrow Jamie Dreier’s unfortunate but apt term). I gesture at how these questions might be answered but leave the details to be worked out in later research.
INTRODUCTION

Many of us pride ourselves on our moral credentials. We have views about what is right and wrong, and we take ourselves to be justified not only in maintaining those views but also in acting in accordance with them and convincing others to do so as well. Often, we are willing to argue for our views with friends, family members, colleagues; sometimes we even get into it with a stranger on the street holding an especially offensive sign. Although we recognize that we might be wrong about particular issues, we still take ourselves to be competent moral speakers and thinkers. We have a general sense of what is right and what is wrong, what is good and what is bad. We have a general sense of what sorts of actions we are forbidden from performing, what sort of person we ought to strive to be, and how to reason and argue about these things with people whose views might differ quite substantially from our own. Sure, we don’t know every detail of the moral landscape, but we have a sense for the general “lay of the land,” and this general sense is enough to get around.

But, while we may be confident in our views about what is right or wrong, when we take a step back to consider our moral views more generally or abstractly, we recognize that there are many features of our moral thought and talk that moral competence does not shed light on. Being able to answer questions about what kind of person we ought to strive to be, or what actions we are forbidden from performing does not mean that we are able to answer metaethical questions about what exactly it means to call something ‘wrong’ or what the word ‘wrong’ refers to, or whether its meaning or reference is stable or varies from culture to culture, or whether judging that something is wrong is more like judging that something is a saguaro cactus or more like
judging that something is tasty. These are metaethical questions: they are questions not about the particular details of substantive debates within our moral practices, but about the metaphysical, epistemological, semantic, and psychological views, assumptions, and commitments that underlie these practices. The important point here is that moral competence does not guarantee true beliefs about metaethical issues, or competence to engage with such issues at the level of abstraction this kind of engagement actually requires. Answers to metaethical questions are often opaque to us rather than transparent: we can participate competently in our moral practice without having true beliefs about its metaethical underpinnings; indeed, we can participate without having any such beliefs at all.

Metaethicists have lots of beliefs about the metaethical underpinnings of our moral practices. They are trained to have such beliefs, and to map out the details of the metaethical terrain by studying the abstract features or moral thought and talk. But for all their successes, metaethicists often fail to pay attention to the ways that failures of transparency matter to metaethical theorizing. Often, in the course of defending or arguing against a particular metaethical theory, metaethicists assume that the claims made by that theory would have to be transparent to competent moral speakers if they were true: that if the theory is right that something is true of our moral thought and talk, then the theory is also committed to those who are competent with moral thought and talk believing that that something is true. For example, consider non-naturalist realism. Non-naturalist realists hold that moral properties are non-natural properties, and that these non-natural properties are ultimately the truth-makers of our moral judgments. This has led some to argue against non-naturalist realism by claiming that if our moral judgments really are made true by non-natural properties, then it must be transparent to
competent moral speakers and thinkers that this is so—even though one can clearly engage in
moral thought and talk without thinking it has anything to do with non-natural properties, or
without having ever considered the distinction between natural and non-natural properties,
whatever that distinction is supposed to amount to anyway.

I call this mistake “the transparency fallacy.” In the first paper of my dissertation, I
provide a clear, general formulation of the transparency fallacy, which is assuming, without
further argument, that if \( p \), then competent moral speakers and thinkers believe that \( p \), where \( p \) is
some putative metaethical fact. I then go on to consider various instances of this mistake as they
appear in contemporary metaethics, beginning, most famously, with Moore’s open question
argument against naturalism, but also considering more recent arguments against non-naturalism
and expressivism that make the same error. I argue that in the course of objecting to competing
views, metaethicists problematically assume that the meaning, reference, truth-makers, and
attitude-types of our moral judgments are transparent to competent moral speakers and thinkers.
Sometimes this assumption is explicit, but more often, it is implicit. Either way, it is a mistake.

There are, I think, many positives to be gained from identifying failures of transparency.
Besides allowing us to diagnose the problems with the arguments I mentioned, identifying
failures of transparency also allows us to explain why the dialectic in certain metaethical debates
proceeds as it does. I demonstrate this in the remaining two papers in the dissertation, the second
of which focuses on a particular commission of the transparency fallacy, and the third on a case
where failure of transparency leads to a dialectical impasse.

In the second paper of my dissertation, I expand on one of the failures of transparency
that I only touch on in the first: that’s the failure of transparency of attitude-types. In that paper, I
take on Jack Woods’ recent argument against a simple version of expressivism according to which to assert a moral judgment is to express a noncognitive attitude of disapproval. Woods’ argument is based on Moore’s paradox. Most of us here are familiar with Moore’s paradox, but just in case, the basic idea of Moore’s paradox is this: Moore’s paradox arises for propositions of the form “p, but I don’t believe that p” or “it’s raining, but I don’t believe that it’s raining.” Moore observed that assertions of these propositions seem “absurd” or self-defeating in much the same way as assertions of contradictions like “it’s raining and it’s not raining” even though these propositions are not contradictory. It is easy to imagine a scenario in which it’s raining, but I fail to believe that it’s raining. For Moore, it is a “paradox” that it is absurd to assert such a possible truth, and the puzzle of Moore’s paradox is to explain this absurdity.

Woods argues that if expressivism is true, we should expect moral versions of Moore’s paradox. That is, we should expect a sincere assertion of “murder is wrong, but I don’t disapprove of murdering” to produce exactly the same absurdity as a sincere assertion of a standard Moorean proposition such as “it’s raining, but I don’t believe that it’s raining.” But, Woods argues, it is not absurd to assert sincerely, “murder is wrong, but I don’t disapprove of murdering,” and so expressivism must be false. The basic problem with Wood’s argument is that it implicitly assumes that, if expressivism were true, it would be transparent to those who are morally competent. More specifically, Moore’s paradox would be a good test of expressivism only if the truth of expressivism guaranteed that competent moral speakers and thinkers recognized that what it is to believe that murder is wrong just is to disapprove of murdering. But it need not be transparent to competent moral speakers and thinkers that what it is to believe that murder is wrong is to disapprove of it, such that whenever they believe something is wrong, they
believe they disapprove of it. The attitude types of our moral judgments need not be transparent, and in fact, expressivists explicitly deny that expressivism is transparent in this way. And so Woods commits the transparency fallacy in his argument against expressivism.

The third paper of my dissertation concerns a case where a failure of transparency leads to a dialectical impasse between cognitivists and expressivists. The debate turns on the phenomenon of moral motivation, and in particular whether internalism about moral motivation lends supports expressivism. Internalists believe that moral judgment is necessarily connected to motivation, while externalists hold that the connection is merely contingent. Expressivists often point to the ease with which their view accounts for the connection between moral judgment and motivation, and claim this to be a distinctive explanatory advantage of their view—after all, if moral judgments just are noncognitive, desire-like states as expressivists claim, then it’s no surprise that they bear this especially intimate connection to motivation. But, I argue that expressivism does not have this explanatory advantage after all.

While it is transparent to competent moral thinkers and speakers, and thus common ground among metaethicists, that moral judgments are in practical or action-guiding in the sense that we are at least typically motivated by our moral judgments, it is not transparent whether the connection between moral judgment and motivation is necessary or contingent connection, nor is it transparent under what conditions moral judgments motivate the individuals who make them. The problem is that expressivists still haven’t provided a plausible characterization of the phenomenon of moral motivation that allows them to retain the explanatory advantage that they claim. They either posit a strong and controversial connection between moral judgment and motivation that cognitivists have no trouble rejecting, or they posit a weaker connection that
cognitivists can explain just as well. Thus, precisely because facts about the connection between moral judgment and motivation are not transparent, neither side can make much headway in convincing the other—the premises they would seem to need to rely on are themselves subject to metaethical debate, not available to us simply in virtue of our moral competence. And if expressivists claim otherwise, they are implicitly sliding from the transparent claim that moral judgments typically motivate to the opaque claim that they necessarily motivate.

My dissertation represents only a first attempt to tackle some very large questions in metaethics and metametaethics (to borrow Jamie Dreier’s unfortunate but apt term). Metaethicists must be careful not to assume the transparency of metaethical facts, if they are to engage fruitfully with one another and make progress in metaethical debates. And focusing on transparency also opens up a new domain of metaethical inquiry, since metaethicists are well-placed to think hard about the conditions under which we should and should not expect certain facts to be transparent, and to construct independent arguments that at least some of these facts may indeed be transparent. Here, I gesture at how these arguments might proceed but leave the details to be worked out in future research.
1. The Transparency Fallacy

In this paper I argue that many contemporary metaethicists commit a fallacy that has not yet received a clear and unified description. I develop a characterization of this “transparency fallacy” by illustrations drawn from the metaethical writings of both historical and contemporary figures. Though the precise contours of this fallacy will become clearer as we proceed, the basic idea is that metaethicists often illicitly assume that if some fact about our moral thought and talk is true, then this fact is transparent to competent moral speakers and thinkers, in the sense that they believe it. In other words, metaethicists often assume without argument that we may rely on the following conditional, where \( p \) is some metaethical fact:

\[
\text{TRANSPARENCY: If } p, \text{ then competent moral speakers and thinkers believe that } p.
\]

Perhaps no one would openly endorse this assumption when stated so explicitly and without qualification, but we will see that metaethicists often implicitly rely on particular instantiations of it in the course of defending particular metaethical positions.\(^1\) I begin with an examination of perhaps the most notorious example of the transparency fallacy—G.E. Moore’s Open Question Argument against naturalism.\(^2\) This will help us to see two different ways that the transparency

\(^1\) The sort of qualifications I have in mind concern the conditions under which competent moral speakers and thinkers form these believes. For example, some versions of the transparency fallacy assume that individuals believe these metaethical facts only \( on reflection \). I will generally avoid such qualifications in what follows since they should be obvious from the context.

\(^2\) Moore (1903)
fallacy may occur: one may assume that the meaning or the reference of moral terms is transparent to competent moral speakers and thinkers. I then turn to some more recent examples of the transparency fallacy in order to illuminate two other ways the transparency fallacy may occur. One may assume that the truth-makers of moral propositions are transparent or that the attitude-types of our moral judgments are transparent. I then discuss a fifth way that one might potentially commit the transparency fallacy, by assuming that the psychological role of our moral judgments is transparent. Although metaethicists have not explicitly committed this version of the transparency fallacy, it is worth considering not only because it demonstrates a distinct failure of transparency, but also because it forces us to consider more carefully which sorts of metaethical facts we can assume are transparent, as well as how we might uncover such facts when they are not. I discuss these issues further in the conclusion, in which I also provide a final statement of the transparency fallacy.

2. Transparency of Meaning and Reference

Moore’s Open Question Argument attacks metaethical naturalism, which we may understand for our purposes here as the view that moral properties are identical to natural properties. Against naturalism, Moore argues that for any proposed identity of a moral property M with a natural property N, it will always be possible for a competent moral speaker, without conceptual confusion, to grant that something is N but still wonder whether it is really M. In other words, according to Moore, it is always an open question whether M is co-instantiated with N. So, for example, consider the simple version of naturalism that Moore attributes to Mill: that

3 This is of course a very rough characterization of a view that takes many shapes. For other characterizations of naturalism, see Boyd (1988), Brink (1989), and Foot (2001).
for something to be good is for it to be desired. The trouble with this proposed identity is that we
can meaningfully ask of anything that is desired, say, fame, whether it is good. The fact that this
is an open question, according to Moore, is evidence of its falsity. If ‘good’ was identical with
‘desired’, then the question, “I know fame is desired, but is it good?” would not be open in this
way for a competent speaker. Rather, the question would be closed: it would look silly and
trivial, like asking if ‘good’ really is ‘good’ or if ‘something people desire’ is really ‘desired’.

In this argument, Moore implicitly assumes something like the following version of
transparency:

**TRANSPARENCY-MOORE**: If moral property M is identical to natural property N,
then competent moral speakers and thinkers will believe that moral property M is
identical with natural property N.⁴

And, sure enough, the two most common responses to Moore’s Open Question Argument have
put pressure on exactly this claim. The first response aims to show that Moore’s argument would
only apply to analytic naturalism: a particular sect of naturalism according to which the identity
between moral and natural properties holds by virtue of the meaning of moral term ‘M’ and
natural term ‘N’.⁵ On this view, the identity between M and N is like the identity between
‘bachelor’ and ‘unmarried male’: ‘bachelor’ is synonymous with ‘unmarried male’ and thus the

⁴ Strictly speaking, Moore only claims that people will believe this when they reflect on the
question, “is M N?” But as I mentioned in fn. 1, glide over such details in this paper.

⁵ At least, this is how to characterize the form of analytic naturalism that Moore takes aim at. In
recent days, some philosophers have revived analytic naturalism, offering fresh accounts that
they claim are in vulnerable to the Open Question Argument. More on that in what follows.
identity between the two analytic and *a priori.* The alternative view, endorsed by *synthetic* naturalists, is that the identity between ‘M’ and ‘N’ is more like the identity between ‘water’ and ‘H₂O’. Although ‘water’ and ‘H₂O’ refer to the very same clear, odorless, flavorless liquid, this identity is *a posteriori,* since there is nothing about the meaning of ‘water’ and ‘H₂O’ that implies that they are identical. So, while they do not put it in these terms, such synthetic naturalists effectively distinguish between two versions of Moore’s transparency assumption:

**TRANSPARENCY-MEANING:** If moral term ‘M’ is synonymous with term ‘T’, then competent moral speakers and thinkers will believe that ‘M’ is synonymous with ‘T.’

**TRANSPARENCY-REFERENCE:** If moral term ‘M’ refers to property P, then competent moral speakers and thinkers will believe that ‘M’ refers to P.

They then argue that while the former may be true, the latter is not. It may be plausible to think that, if the identity between, say, ‘good’ and ‘desired’ is analytic and *a priori,* this would be transparent to competent moral speakers and thinkers. But it is not plausible to think synthetic and *a posteriori* identities are transparent in this way. Just as a speaker can competently use the word ‘water’ without believing that it refers to H₂O, a speaker can competently use the word ‘good’ (or some other moral term) without believing that it has its actual reference. Thus, it is no objection to synthetic naturalism that it appears to be an open question whether moral terms are

6 Of course, one might—following Quine—reasonably doubt that the distinction between analytic and synthetic truths can be sustained. Although Quine’s arguments in “Two Dogmas of Empiricism” are no doubt important, I will set them aside in what follows: regardless of Quine’s criticisms, the analytic-synthetic distinction is alive and well among contemporary metaethical naturalists.
identical with any particular natural properties, since we should not expect mere competency with moral terms to guarantee the transparency of their reference.

The second response comes from analytic naturalists, who instead deny the transparency of meaning. According to these analytic naturalists, “analyticity can be opaque or unobvious,” which means the analogy with the identity of bachelors and unmarried males is misleading.⁷ Though the identity between bachelors and unmarried men is like the identity that analytic naturalists claim holds between moral and natural terms in one respect—namely, that it is true by virtue of meaning—it is unlike the identity in another—namely, that the meaning of ‘bachelor’ is simple and therefore transparent to competent speakers while the meaning of a term like ‘good’ might be highly complex and therefore opaque even to competent speakers. Just as someone might competently and correctly apply the term ‘triangle’ without believing that a triangle is a three-sided polygon whose interior angles add up to 180-degrees, someone might competently and correctly apply the term ‘good’ without believing, say, that goodness is a large conjunction of various platitudes surrounding the application of the term that determine its meaning.⁸ In that case, the Open Question Argument does not even pose a problem for analytic naturalism.

⁷ Finlay (2014, p. 9), but see also Smith (1994) and Jackson (1998).

⁸ If one doesn’t like this example, just think of any analytic proof which shows that a certain property is uniquely satisfied by something that satisfies some other property (or set of properties). For example, one can understand what Kenneth Arrow means by a “dictatorship,” and understand the other axioms in his famous proof, without recognizing that dictatorship is analytically equivalent to a preference-aggregation mechanism that satisfies these other axioms. See Arrow (1951/1963).
Defenders of the view can reply that the reason the question remains “open” is that the meanings of our moral terms are not transparent to competent moral speakers and thinkers, in much the same way that synthetic naturalists argue that their referents are not.

Summing up our discussion to this point, Moore’s Open Question Argument fails as a critique of both analytic and synthetic naturalism. The argument assumes either that meaning or that reference is transparent to competent moral speakers and thinkers, even though both are substantive metaethical claims that one cannot help oneself to without further argument. Thus, Moore is guilty of committing the transparency fallacy—of assuming, without argument, that certain features of moral thought or talk are transparent to competent speakers—and his objections to naturalism is unsuccessful as a result.

3. Transparency of Truthmakers

Whereas Moore illicitly relies on transparency in an argument against naturalism, a more recent argument against Moore’s favored view, non-naturalism, makes a similar error. Like the sort of naturalists Moore targets, non-naturalists believe that moral judgments are ordinary beliefs that are capable of being true or false. But unlike naturalists, non-naturalists deny that the truth-makers of such moral judgments can be entirely natural. Thus, whereas a naturalist might claim that murder is wrong because wrongness really refers to some perfectly natural property that murder happens to instantiate, non-naturalists insist that the wrongness of murder is instead explained by something non-natural.9

9 Again, ‘non-naturalism’ picks out a range of views, but see Moore (1903) for the seminal version of the view. For contemporary versions, see Enoch (2011), Fitzpatrick (2008), Shafer-Landau (2003).
In a recent paper, Matt Bedke argues against a particular version of non-naturalism that claims that properties such as wrongness are non-natural, and that our moral judgments are beliefs about such properties. He then argues that non-naturalism must be false because our moral judgments do not seem to respond to evidence in the way we would expect them to if non-naturalism were true. If moral judgments were beliefs about non-natural properties, he argues, then we should expect people to change or abandon their first-order moral judgments if they come to accept that there is conclusive evidence against the existence of non-natural properties.

That is just how beliefs work. But, he argues at length, this is not what we find: our first-order moral judgments are remarkably stable in the face of (what we take to be) compelling evidence against the existence of non-natural properties. So non-naturalism must be false: as Bedke puts it, “Normative judgments might be beliefs, and there might be non-natural properties, but normative judgments are not beliefs about non-natural properties” (199).

Bedke’s commission of the transparency fallacy becomes apparent when we distinguish two senses in which our moral judgments might be beliefs about non-natural properties. On the one hand, a judgment that murder is wrong might be a belief with the explicit content that murdering instantiates the non-natural property of wrongness. In this case, we should expect people to change or abandon their moral judgments in response to conclusive evidence against the existence of non-natural properties in precisely the way Bedke thinks: people take themselves to be referring to such non-natural properties when they make such judgments. But on the other hand, a judgment that murder is wrong might be a judgment with the less specific content that murdering instantiates the property of wrongness—where the content of this judgment makes no

---

10 Bedke (2014)
explicit reference to the non-naturalness of the property. In this case, we need not expect people to change or abandon their moral judgments in response to conclusive evidence against the existence of non-natural properties: it need not be transparent to them that the property of wrongness their moral judgments refer to is non-natural, or indeed that their moral judgments have anything to do with non-natural properties at all.

Consider an analogy. Suppose I believe that Clark Kent is in the phone booth, but then receive conclusive evidence that Superman is not in the phone booth when I see him flying by. If it were transparent to me that Superman is Clark Kent, then upon receiving this evidence, I would cease to believe that Clark Kent is in the phone booth. But if I have no idea that Superman is Clark Kent, then we should not expect me to revise my belief in this way. After all, by my lights, the fact that Superman is not in the phone booth bears not at all on whether Clark Kent is in the phone booth—it would never occur to me to think they are the same person. And in exactly the same way, if I believe that murdering is wrong, but then receive conclusive evidence that there are no non-natural properties, we should not expect me to revise my belief in murder’s wrongness unless it is transparent to me that wrongness is a non-natural property. But non-naturalists are not committed to the non-naturalness of moral properties being transparent to us in this way. It need not be transparent to us that the truth of our moral judgments is explained by the existence of non-natural properties, even if this is ultimately the right explanation of their truth.12

11 See William Fitzpatrick (2016, 2018)

12 For a more thorough defense of non-naturalism against this kind of objection, see Fitzpatrick (2018).
This point generalizes beyond the debate over non-naturalism. *Whatever* it is that makes our moral judgments true, this need not be transparent to us. For example, take a naturalistic utilitarian view on which a moral judgment, say, that murdering is wrong, is true because murdering realizes some natural property, say, it violates rules whose widespread adherence would maximize pleasure. Or consider a simple sort of subjectivism on which the moral judgment that murdering is wrong is true because murdering realizes the property of being disapproved of by some person or group. Proponents of such views are not commits to these truth-makers being transparent. Utilitarianism or subjectivism might be true, even if no one (other than a few philosophers) believe them. Thus, we have identified another instance of transparency that metaethicists ought not to assume:

**TRANSPARENCY-TRUTHMAKERS:** If moral judgment M is true in virtue of F, then competent moral speakers and thinkers will believe that M is true in virtue of F.

Bedke’s argument fails because he assumes that if moral judgments are true in virtue of the existence of non-natural properties, then this will be transparent, such that a belief that there are no non-natural properties would extinguish their first-order moral beliefs. But this need not be transparent to competent moral speakers and thinkers. And so, much like Moore, Bedke commits the transparency fallacy, and his argument fails.

4. Transparency of Attitude Types

So far we have seen the transparency fallacy committed in arguments against naturalist and non-naturalist versions of moral realism, but a similar mistake makes a frequent appearance in arguments against non-cognitivist views such as expressivism. Expressivists deny that moral judgments are cognitive, belief-like states that are capable of being true or false in a robust, non-
deflationary way, and maintain instead that moral judgments are non-cognitive, desire-like states such as approval or disapproval. According to expressivists, assertions of moral judgments express non-cognitive states: a sincere utterance of “murder is wrong” expresses the speaker’s (say) disapproval of murder in precisely the same way a sincere utterance of “Tucson is in Arizona” expresses the speaker’s belief that Tucson is in Arizona.

In a recent paper, Jack Woods provides an argument against expressivism based on Moore’s paradox. The standard case of Moore’s paradox arises for judgments of the form ‘p, but I don’t believe that p’. Moore observed that assertions of such Moorean judgments seem “absurd” or self-defeating in much the same way as assertions of contradictions, even though Moorean judgments are not contradictory. For instance, it’s easy to imagine a scenario in which it’s raining but I fail to believe that it’s raining. But my reporting this belief by asserting it out loud seems absurd. Woods thinks that the presence or absence of the absurdity characteristic of standard cases of Moore’s paradox is an adequate test of the truth of expressivism, since, after all, expressivists claim that moral assertions express disapproval (to use Woods’ stand-in non-

---


14 Schroeder calls this the parity thesis in (2008).

15 Moore (1944)

16 More carefully, Moorean propositions are paradoxical because even though they can be true, assertions of them are absurd in a way that cannot be fully explained in terms of a semantic contradiction generated by the words of the proposition.
cognitive attitude) in precisely the same way as ordinary, descriptive assertions express beliefs. He claims that if expressivism is true, a sincere assertion of ‘murder is wrong, but I don’t disapprove of murder’ should produce the same absurdity as a sincere assertion of ‘it’s raining, but I don’t believe that it’s raining’. But, he argues, it is not absurd to assert ‘murder is wrong, but I don’t disapprove of murder’, so expressivism is false.

Standard accounts of Moore’s paradox diagnose the presence of absurdity by attempting to show that people who express Moorean judgments have (or are in some relevant sense committed to) a belief with self-contradictory content.\(^{17}\) Without going into details, the basic idea is that when someone asserts a Moorean judgment such as ‘it’s raining, but I don’t believe it’, we can show (given suitably qualified principles governing belief) that they hold the belief ‘I believe that it’s raining and I don’t believe that it is raining’, which is self-contradictory and therefore absurd.\(^{18}\) The key move, in other words, is to show that when someone (sincerely) asserts, ‘p, but I don’t believe that p,’ we can attribute to that individual a belief in the first conjunct, and can combine this with the second conjunct to show that they must believe ‘I believe that p and I don’t believe that p.’ The problem with Wood’s argument is that when we perform this operation on an assertion like, ‘murder is wrong, but I don’t disapprove it,’ we can only show that a person believes ‘I believe that murder is wrong, but I don’t disprove of murder’—but this is not self-contradictory, and therefore not absurd. It is thus only if we assume

---


\(^{18}\) For the details, see my (2018) where I take on Woods’ argument at greater length. I draw on this in what follows.
that it is transparent to ordinary speakers that what it is to believe that murder is wrong is to disapprove of it, such that whenever they believe that something is wrong they believe that they disapprove of it, that we can attribute to them a self-contradictory belief: ‘I disapprove of murder but I don’t disapprove of murder.’ Absent this transparency, there is no belief with self-contradictory content, and we should not expect any absurdity. So we arrive at a third instance of problematically assuming transparency:

**TRANSPARENCY-ATTITUDE:** If moral judgments are of a particular attitude type (e.g., belief or disapproval), then competent moral speakers and thinkers will believe that moral judgments are of that particular attitude type.

Woods’ argument fails because he assumes that if what it is to judge that murder is wrong just is to disapprove of murder, then this will be transparent to competent moral speakers, such that someone who believes ‘murder is wrong’, would believe ‘I disapprove of murder’, which would combine with their belief that ‘I don’t disapprove of it’ to yield a belief with self-contradictory content: ‘I disapprove of murder and I don’t believe in it’. But the attitude types of our moral judgments need not be transparent, and in fact, expressivists explicitly deny that expressivism is transparent in this way. So, like Moore and Bedke, Woods commits the transparency fallacy, and his argument fails.

**5. Transparency of Psychological Role**

Though expressivists deny that our attitude types are transparent to us, they do seem to take for granted that other facts about our moral thought and talk are transparent to us, and to rely on this transparency in defending their views. For example, they generally assume that it is

---

19 See Blackburn (1992, pp. 172–173; 1996)
transparent to competent speakers and thinkers that moral judgments are in some way practical or action-guiding: that, at the very least, we tend to be motivated by our moral judgments. This, however, does not appear to be fallacious. Whereas there is room for disagreement about the meaning, reference, truth-makers, and attitude types of moral judgments, its practicality is non-negotiable among competent moral speakers and thinkers. Someone who does not recognize that morality is practical, and that moral judgments are therefore at least typically connected to motivation, does not get what morality is about. They do not qualify as competent moral speakers or thinkers. 20

One oft-touted advantage of expressivism concerns the ease with which the view accounts for this practicality of moral judgment. The basic idea is that expressivists can provide the best explanation of the practicality of moral judgment, and that this lends credence to their view. One common way this argument goes takes as a premise the Humean theory of motivation: that, taken by themselves, beliefs neither motivate nor generate motivationally efficacious states. The argument then proceeds to assert the internalist premise that, necessarily, if you judge that you ought to do something, then you will be motivated to do it. And it combines these two premises to arrive at the conclusion that moral judgments are (or are at least partly constituted by) noncognitive states, since cognitive states have no internal or necessary connection with motivation.

Although some cognitivists rebut this argument by rejecting the Humean theory of motivation, for our purposes, it is the internalist thesis and the debate surrounding it that is more interesting. The dialectic proceeds roughly as follows. Cognitivists attempt to block the argument

20 As Jackson puts it, I defend this claim further shortly.
for expressivism by denying internalism and pointing to cases where we act against our moral judgments—for example, cases of *akrasia*, or weakness of will.\(^{21}\) Expressivists, in turn, claim that even in these cases, we remain *to some extent* motivated: that our moral judgments always motivate us, even though this motivation is sometimes outweighed by competing desires, and this can only be explained by expressivism.\(^{22}\) From here, the cognitivist turns to the more extreme case of the amoralist, or perhaps the psychopath, who makes moral judgments without being motivated by them at all. She argues that the possibility of such characters demonstrates the falsity of internalism, again undermining the argument for expressivism.\(^{23}\) Finally, the expressivist replies that such characters are indeed (conceptually or metaphysically) impossible: that though an amoralist or psychopath might use moral language, they do so in an ‘inverted commas’ sense with which one might assert a moral judgment that one does not actually endorse (but which, say, one will be rewarded for asserting).\(^{24}\) The cognitivist digs in her heels, and claims that such people are making genuine moral judgments despite their lack of motivation; the expressivist remains equally intransigent.

Regardless of whether cognitivism or expressivism is the correct metaethical theory, this dialectic reveals another version of transparency that metaethics must avoid assuming:

---

\(^{21}\) See Brink (1989: Chapter 3); Mele (1996); Miller (2008); Roskies (2003); Stocker (1979); Svavarsdóttir (1999, sec.1).

\(^{22}\) See Gibbard (2003: 153)

\(^{23}\) See Nichols (2004: 65-82)

\(^{24}\) See Hare (1952: 124–126, 163–165)
TRANSPARENCY-PSYCHOLOGICAL ROLE: If moral judgments play a causal role in our psychology, then competent moral speakers and thinkers will believe that moral judgments play this role.

This version of transparency is clearly false, since the mere fact that expressivists claim that we are always to some extent motivated by our moral judgments, whereas cognitivists deny this and claim that such judgments only typically motivate (since they are typically accompanied by motivational states), shows that the precise motivational role of our moral judgments is not transparent. Both expressivists and cognitivists are competent moral speakers and thinkers, and yet they disagree about the causal role of moral judgments. And the same failure of transparency arises in debates over the amoralist. Though expressivists can maintain the consistency of their view by insisting that the amoralist does not make genuine moral judgments, they cannot argue from this premise to the conclusion that moral judgments necessarily motivate and are therefore, by the Humean theory of motivation, non-cognitive states. After all, precisely because it is not transparent to us competent moral speakers and thinkers whether a condition for competently and sincerely making a moral judgment is being motivated by them, we cannot take it for granted that those who are not motivated by their judgments are either incompetent or insincere.

To be clear, I am not suggesting that metaethicists on either side of this debate commit the transparency fallacy. Rather, it is precisely because metaethicists do not assume that psychological role is transparent that cognitivist and expressivists find themselves in an impasse over internalism. After all, if psychological role were transparent, and expressivists, say, were right about it, then expressivists could give the following very simple argument: (1) I am a competent moral speaker and thinker who believes that moral judgments necessarily motivate,
(2) the psychological role of moral judgments (and, in particular, whether or not they necessarily motivate) is transparent to competent moral speakers and thinkers, (3) so, moral judgments necessarily motivate. And if cognitivists were right, they could give a parallel argument for the opposite conclusion. What is interesting about this case, then, is that even though the weak claim that moral judgments typically motivate is plausibly transparent, the strong claim that moral judgments necessarily motivate is not transparent, regardless of whether it is true. While all competent moral speakers and thinkers accept the weak claim, many reject or are agnostic about the strong one, and so one would be committing a fallacy if one relied on it without independent argument.

This raises the important question of why some claims are transparent to competent moral speakers and thinkers while others are not. Although this is a very large question, at least part of the answer seems to rest on the distinction between two sorts of metaethical claims. First, there are the “platitudes” about morality that form a “network of moral opinions, intuitions, principles and concepts whose mastery is part and parcel of having a sense of what is right and wrong, and of being able to engage in meaningful debate about what ought to be done.” Such platitudes include first-order moral claims such as “if an act is an intentional killing, then normally it is wrong; pain is bad; ‘I cut, you choose’ is a fair procedure” as well as metaethical claims such as that “the realization that an act would be dishonest typically dissuades an agent

---

25 Jackson (1994, p. 130). Note that in giving this explanation, I do not commit myself to Jackson’s analytic functionalism. Jackson believes that the meanings of moral terms are determined by the sorts of platitudes I refer to, but I am not making a claim about meaning: I am merely suggesting that competence with moral language requires one to accept certain platitudes.
from performing it; properties that make something good are the properties we typically have some kind of pro-attitude towards” and, to take the example we have already relied on, that “the judgement that an act is right is normally accompanied by at least some desire to perform the act in question.” 26 Although it is controversial precisely which platitudes belong in this set, it seems fair to assume that the acceptance of at least some of these sort of claims is indeed a necessary condition of moral competence. Individuals who do not share some set of platitudes about morality cannot engage in meaningful debate about morality because they cannot see each other as engaging in a competent discussion of the same subject matter. We can therefore explain why certain facts are transparent by pointing out that, if someone did not accept them, we would for this reason cease to classify them as a competent moral speaker or thinker. 27

On the other hand, there are many facts about our moral practice that are not transparent to competent speakers, because individuals may disagree about them while still recognizing each other as competently engaged in a discussion about the same subject matter: morality. These include basically all of the controversial claims that normative ethicists and metaethicists defend; one can disagree with these claims without disqualifying oneself as incompetent. Thus, precisely because moral speakers can competently disagree and engage in debate about the meanings, reference, truth-makers, attitude-types, and psychological role of moral terms or judgments, we

26 Jackson (1994 p. 131).

27 Perhaps this is not the only reason facts might be transparent to competent moral speakers and thinkers: for example, we might have direct introspective access to certain facts about our psychology that are metaethically relevant, such as that we are averse to pain. I return to this point in the conclusion.
cannot take for granted that such facts are transparent to competent moral speakers and thinkers. This might seem an obvious point, but we have seen that many metaethicists at least implicitly fail to recognize it, and that is why it is worth calling attention to the transparency fallacy.

6. Avoiding the Transparency Fallacy

The transparency fallacy in metaethics can now be more carefully stated as follows:

**TRANSPARENCY FALLACY**: Assuming, as an undefended premise in metaethical argument, that competent moral speakers and thinkers have true beliefs about the meaning, reference, truth-makers, attitude-types, and psychological roles of their moral terms or judgments.

This fallacy, we have seen, is often committed in the course of providing negative arguments against metaethical views though it could also at least potentially be used to provide positive arguments as well. In particular, we have seen that the assumed transparency of meaning and reference has been used in arguments against naturalism, that the assumed transparency of truth-makers has been used in arguments against non-naturalism, and that the assumed transparency of attitude-types and psychological roles have been used in arguments against and can be used in arguments for expressivism. In each case, I have suggested, the problem lies not in the particular claims that these theorists are making about meaning, reference, truth-makers, attitude-types, or psychological roles. It lies rather in their assumption that we can recognize the truth about such things simply in virtue of our moral competence.

If metaethicists cannot appeal to the transparency of these various metaethical facts, how are they to proceed? Is metaethics a hopeless pursuit, one that cannot go beyond the platitudes all competent moral speakers and thinkers already accept? No. Nothing as strong follows: as on
many other issues, we can make progress in metaethics by engaging in philosophical
argumentation or empirical research. In the case of the meaning and reference of moral terms, for
example, the debate could perhaps be resolved through appeal to independent theories of
meaning and reference-determination.\(^{28}\) In the case of truth-makers, there is already a massive
literature on the metaphysical question of whether there could even be non-natural properties in
the first place, and the epistemic question of how we could ever come to be acquainted with
them.\(^{29}\) In the case of attitude-types, there is the well-known Frege-Geach problem, which
purports to show that unless our moral judgments are standard cognitive attitudes, moral
language could not have the logical properties that it apparently does.\(^{30}\) And though there may be
a philosophical impasse over the psychological role of moral judgments, recent empirical work
on moral judgment has the potential to shed important light on this subject. For example, there
has recently been an influx of fMRI studies involving the different patterns of activation
involved in making moral judgments versus ordinary descriptive judgments, whose goal is to
discover whether psychopathic moral judgments more closely resemble the former or the latter.\(^{31}\)

In addition to these more standard metaethical arguments, our discussion also points to a
fruitful new avenue of metaethical research. Although I have argued that we cannot assume that
various metaethical facts are transparent to competent moral speakers and thinkers, this does not

\(^{28}\) See Eklund (2017) and Finlay (2014)

Scanlon (2014)

\(^{30}\) Blackburn (1973, 1988), Geach (1965), Hare (1952), Searle (1962), van Roojen (1996)

\(^{31}\) See, for example, Karenski et al. (2010)
mean that we cannot construct independent arguments that at least some of these facts are indeed transparent, thus bolstering the sorts of arguments that I have shown are at best incomplete. In light of our discussion in the last section, the most straightforward way of doing this is to defend particular conditions of moral competence. For example, one might defend a theory of meaning, reference, or truth-makers on which competence with terms does require individuals to believe the truths about such meanings, references, and truth-makers. Or perhaps more plausibly, one might argue that competence in a domain of discourse at least requires one to recognize what attitudes one is expressing: just as we would think that one would be an incompetent sports fan if one yelled “Hooray for the Golden State Warriors!” without recognizing that one was expressing a positive, non-cognitive attitude towards the Golden State Warriors, one might argue that expressivists are committed to thinking that competence with moral talk and thought does indeed require the recognition that one is expressing a particular negative, non-cognitive attitude towards murder when one states that “murder is wrong!” Alternatively, however, one might also defend a particular instance of transparency by arguing that something is transparent not only to competent moral speakers and thinkers, but to a broader class that of individuals that includes them. For example, one might defend a general view about the conditions under which people’s psychological states are transparent to them through introspection, which one might use to get some purchase on the internalism debate—say, a view on which introspection is prone to lead to false positives but not false negatives about whether we are to some extent motivated by our moral judgments.

These are just some off the cuff examples, designed to demonstrate the type of argument that might resolve disputes over whether a particular range of facts is transparent. Of course, to
engage in metaethical argument, one need not always rely on the claim that any range of facts is transparent: some arguments do not proceed by reference to what ordinary moral speakers believe. But if one does one wish to appeal to transparency, one had better back it up. One cannot take for granted that facts about our moral thought and talk are transparent to competent moral speakers and thinkers. To do so is to commit the transparency fallacy.
Works Cited


1. Introduction

In a recent paper, Jack Woods (2014) advances an intriguing argument against expressivism based on Moore’s paradox. Woods argues that a central tenet of expressivism—which he, following Mark Schroeder (2008a), calls the parity thesis—is false. The parity thesis is the thesis that moral assertions express noncognitive, desire-like attitudes like disapproval in exactly the same way that ordinary, descriptive assertions express cognitive, belief-like attitudes.¹ Most contemporary defenders of expressivism seem not only to accept the parity thesis but also to rely on it to distinguish their view from subjectivism, so Woods’s argument against it poses a serious challenge to the view. In this paper, I argue that Woods’s argument is unsuccessful, but show that diagnosing precisely where it goes wrong raises interesting questions.

¹ Woods actually characterizes the parity thesis as “the thesis that moral assertions express noncognitive attitudes like disapproval in exactly the same way that non-moral assertions express cognitive attitudes like belief” (2014: 3). In this paper I follow Mark Schroeder and use ‘descriptive’ rather than ‘non-moral’ for the simple reason that doing so makes both the parity thesis, as well as Woods’s argument against it, easier to understand and articulate. Nothing about my objection to Woods depends on this switch. It is nevertheless important to be clear that Woods explicitly states that his target in the paper is expressivism about the moral and not expressivism about, for instance, epistemic modality (‘It might be raining outside.’) or about the normative in general (‘All things considered, I really ought to stay in tonight to grade these logic exams.’). For representative examples of defenders of metaethical expressivism committed to the parity thesis, see Gibbard (2003: 64) and Blackburn (1984: 169).
for expressivists—and metaethicists more generally—about the transparency of our moral attitudes.²

Here is how the paper goes. In Section 2, I introduce expressivism and the parity thesis. In Section 3, I outline Woods’s case against the parity thesis. In Section 4, I raise a quick response to Woods’s argument: I suggest that the argument is unsuccesful because it presupposes that expressivism is transparent to ordinary speakers—that is, it presupposes that ordinary speakers recognize that our moral beliefs are identical with noncognitive, desire-like attitudes.³ In Sections 5 and 6, I consider Woods’s preferred solution to Moore’s paradox, and argue that it is incomplete since it accounts only for the absurdity of asserting a Moorean proposition, and not for the absurdity of believing one. I defend a popular alternative solution to the paradox that avoids this problem by explaining the absurdity of asserting a standard Moorean proposition like ‘it is raining but I don’t believe it’ in terms of the absurdity of believing one. In the remainder of the paper (Sections 7, 8, and 9), I draw on this solution to the paradox to provide a theoretical explanation of why the absence of transparency offers expressivists a way


³ Most contemporary expressivists continue to call moral judgments “beliefs” despite denying that they are the same attitudes as ordinary, descriptive beliefs (which are not, of course, noncognitive attitudes). This is part of the quasi-realist project of attempting to “earn the right” to use all the same language as the cognitive moral realists. (See Blackburn 1984.) I will adopt this same practice myself in what follows, but for an intriguing discussion of quasi-realism, see Dreier (2004).
to block Woods’s argument against the parity thesis based on Moore’s paradox. I conclude with some brief, exploratory remarks about the transparency of our moral attitudes.

2. **Expressivism, Expression, and the Parity Thesis**

Before we get to Woods’s argument, a few introductory remarks about expressivism are in order. Like Woods, I will talk as if defenders of expressivism accept a single view, though of course ‘defenders of expressivism’ picks out a motley crew of philosophers whose views differ in subtle ways. For the purposes of this paper, let ‘expressivism’ be the metaethical view according to which (i) moral utterances such as ‘murder is wrong’ do not express beliefs that could be true or false in a robust, non-deflationary sense; and (ii) instead, moral utterances express noncognitive, desire-like attitudes such as approval or disapproval. For simplicity, let us assume that the relevant noncognitive attitude is disapproval, so that ‘murder is wrong’ expresses the speaker’s disapproval of murder. Then, according to the expressivist, to judge that murder is wrong not to have a belief about murder, but rather to disapprove of murder. And to express this judgment is not express a belief but rather to express disapproval.

Central to the expressivist view is the idea that we can *express* our attitudes without *reporting* that we have them. When I assert ‘murder is wrong,’ I am not reporting that I disapprove of murder, but rather expressing my disapproval. This is what distinguishes expressivism from cognitivist speaker subjectivism, according to which a sincere assertion of ‘lying is wrong’ reports an attitude, and therefore has the same meaning—or at least, the same

---

4 See, for example, Allan Gibbard (1990) and Simon Blackburn (1984; 1992; 1998)
truth conditions—as a sincere assertion of ‘I disapprove of lying’. But what exactly is it for an assertion to express a noncognitive, desire-like attitude rather than report it? The critical expressivist move is to appeal to the parity thesis. The answer the parity thesis offers is that we are already familiar with the idea that assertions express without reporting attitudes; namely, assertions of ordinary, descriptive statements express cognitive attitudes such as belief. If I believe that it is raining, uttering ‘it is raining’ is different than stating ‘I believe that it is raining,’ though both bear some relation to the same belief. The former expresses my belief that it is raining, while the latter reports it. To accept the parity thesis is just to maintain that sincere assertions of moral statements express noncognitive, desire-like attitudes in exactly the same way as ordinary, descriptive statements express cognitive attitudes like beliefs, however that might be. Whatever relation holds between a sincere assertion of ‘it’s raining’ and the speaker’s belief that it’s raining, the same one holds between a sincere assertion of ‘murder is wrong’ and the speaker’s disapproval of murder. And the same relation also holds between a sincere utterance of ‘hooray!’ and a speaker’s noncognitive, positive attitudes. As we will see, the parity between these three different types of expression—expressing beliefs, expressing moral disapproval in expressivist fashion, and expressing more mundane noncognitive attitudes such as whatever ‘hooray!’ expresses—plays an essential role in Woods’s argument.

---

5 For a careful discussion of the differences between expressivism and subjectivism, see Mark Schroeder’s (2008a; 2008b; 2014).
3. Woods versus the Parity Thesis

The standard case of Moore’s paradox arises for propositions of the form \(<p, \text{ but I don’t believe that } p>\).\(^6\) Call propositions of this form Moorean propositions. Moore observed that assertions of Moorean propositions seem “absurd” or self-defeating in much the same way as assertions of contradictions, even though Moorean propositions are not contradictory.\(^7\) For instance, it’s easy to imagine a scenario in which it’s raining but I fail to believe that it’s raining. For Moore, it is a “paradox” that it is absurd to assert such a possible truth.\(^8\) And the puzzle of Moore’s paradox is to explain this absurdity.

Woods thinks that the presence or absence of the absurdity characteristic of standard

---

\(^6\) Woods’s argument focuses on the \textit{omissive} form of Moore’s paradox, so called because an assertion of \textquoteleft{}\textit{p}, but I don’t believe that \textit{p}\textquoteright{} reports the specific \textit{omission} of true belief. In this paper, I follow Woods’s lead, but it is worth pointing out that Moore also observed that to say, \textquoteleft{}I believe that he has gone out, but he has not\textquoteright{} would also be ‘absurd’ (1944: 204). This is the \textit{commissive} form, so called because an assertion of \textquoteleft{}p, but I believe that not-\textit{p}\textquoteright{} reports the \textit{commission} of a specific mistake in belief.

\(^7\) More carefully, Moorean propositions are paradoxical because even though they can be true, assertions of them are absurd in a way that cannot be fully explained in terms of a \textit{semantic contradiction} generated by the words of the proposition.

\(^8\) Moore put it like this: “It is a paradox that it should be perfectly absurd to utter assertively words of which the meaning is something which may quite well be true—is not a contradiction” (2013: 209).
cases of Moore’s paradox is an adequate test of the truth of the parity thesis. Here is my reconstruction of Woods’s argument against the parity thesis:

1. If the parity thesis is true, a sincere assertion of ‘murder is wrong, but I don’t disapprove of murder’ will produce the same absurdity as a sincere assertion of ‘p, but I don’t believe that p’.
2. It is not absurd to assert sincerely ‘murder is wrong, but I don’t disapprove of murder’.
3. So, the parity thesis is false.

Woods seems to think that premise 2 is the crux of the argument, since he spends the majority of his paper making the case that it is not absurd to assert ‘murder is wrong, but I don’t disapprove of it’. Much more important for his argument, however, is premise 1, according to which Moore’s paradox is a good test of the parity thesis. And unfortunately, Woods does not have much to say in its favor. His defense for this claim focuses on a set of statements that, according to Woods, generate the absurdity characteristic of standard cases of Moore’s paradox.

Woods considers:

(A) Go Red Sox! I don’t support the Red Sox.

(B) Fuck the Yankees! I don’t have any negative attitude towards the Yankees.

(C) Murder is wrong, but I don’t believe that murder is wrong.

(D) It’s raining, but I don’t believe that it’s raining. \(^9\)

Woods claims that (A–D) all generate the absurdity characteristic of standard cases of Moore’s paradox. Now (C) and (D) are standard Moorean propositions, so it is no surprise that they

\(^9\) (A–C) are lifted directly from Woods’s paper (2014: 5). (D) is a standard case of Moore’s paradox.
generate the absurdity characteristic of standard cases of Moore’s paradox. (A) and (B), however, contain interjections in the place of statements in the indicative. In this way, (A) and (B) are not standard Moorean propositions. Even so, Woods contends that sincere utterances of (A) or (B) reproduce the very same absurdity as sincere assertions of (C) or (D). Woods explains,

These examples indicate that we can reproduce cases of the incoherence present in the standard Moore-paradoxical constructions without using sentences in the indicative. Ejaculations like ‘Go Red Sox!’ are not usually considered assertions, but they give rise to the same incoherence when paired with a denial of possessing the relevant backing attitude of approval. Sincere utterance of ‘Go Red Sox’ thus plausibly requires the possession of some degree of enthusiasm and approval for the team. (2014: 5)

The suggestion in this passage is that because sincerely uttering ‘Go Red Sox!’ expresses the speaker’s support of the Red Sox, pairing this utterance with a sincere assertion of ‘I don’t support the Red Sox’ reproduces the very same absurdity we find in standard cases of Moore’s paradox. In other words, (A) and (B) are defective because when a sincere utterance expresses that the speaker is in a mental state $p$, conjoining this utterance with the denial of being in $p$ yields the absurdity characteristic of standard cases of Moore’s paradox like (C) and (D). (A) and (B) are defective because when a sincere utterance expresses that the speaker is in a mental state $p$, conjoining this utterance with the denial of being in $p$ yields the absurdity characteristic of standard cases of Moore’s paradox like (C) and (D).

---

10 I use ‘utter’ rather than ‘assert’ here because interjections such as ‘Go Red Sox!’ express noncognitive attitudes that lack propositional content. Utterances express attitudes that may but need not have propositional content. Assertions, on the other hand, are utterances that necessarily express propositional content. Henceforth, I’ll use ‘utterances’ for any set of propositions that includes (A) and (B) to respect this distinction.
(B) play an important role in Woods’s argument, and in what follows, I shall refer to them as
emotive Moorean propositions for the simple reason that the attitudes they express are more
similar to those expressed by interjections like ‘Boo!’ and 'Hooray!'.

The fact that (A) and (B) reproduce the same absurdity as (C) and (D) is supposed to
show that the presence of Moorean absurdity is a product of the expression relation that holds
between a speaker’s sincere utterances and the attitudes those utterances express. It is also
supposed to show that nothing about the generation of Moorean absurdity depends on whether
the attitude expressed by the utterance is cognitive (the attitude of believing that it’s raining) or
noncognitive (the attitude of supporting the Red Sox).

The next move of Woods’s argument is to claim that the characteristic Moorean absurdity
present in (A–D) is absent in:

(E) Murder is wrong, but I don’t disapprove of murdering.

(F) Murder is wrong, but I’m not against murdering.

(G) Murder is wrong, but I’m in favor of murdering.\footnote{Again, these are lifted directly from Woods’s paper (2014: 5).}

Woods remarks that although we might not hire the person who uttered (E–G) to babysit our
children, “we do not find utterances of these propositions defective” in the way that we find
utterances of (A–D) defective (Woods 2014: 5). Although Woods does not say much to
characterize the “defect” that (A–D) possess but (E–G) lack, what he seems to have in mind is a
bit of linguistic data about the way we “hear” utterances of Moorean propositions in standard
cases. The linguistic data is that in standard cases, the absurdity of a Moorean assertion is
immediately discernible by or obvious to the speaker who utters it, her audience, and those of us
reading the article. Woods thinks that the parity thesis not only predicts the absurdity of assertions of (E–G), but also that the absurdity of these assertions should be just as “audible” and immediately discernible as it is in (A–D).

The absence of this absurdity in (E–G) is supposed to be a problem for expressivism insofar as the view is committed to the thesis that moral assertions express noncognitive attitudes in exactly the same way as ordinary, descriptive assertions express beliefs. The absurdity we find in standard cases of Moore’s paradox is a symptom of the fact that ordinary, descriptive assertions express beliefs. So if the parity thesis is true, assertions of so-called expressivist Moorean propositions such as (E–G) should generate the same absurdity as assertions of standard Moorean propositions. After all, (E–G) each conjoin an assertion that expresses the speaker’s being in a mental state $p$ with one that denies that the speaker is in that state. Moreover, recall that Woods takes himself to have established that (A) and (B) produce the very same absurdity as (C) and (D). This effectively blocks one natural explanation of why assertions of (E–G) are not absurd, namely, that they are not absurd because the attitude expressed by asserting ‘murder is wrong’ is noncognitive. For, as noted above, the attitude expressed by ‘Fuck the Yankees!’ is also noncognitive, and yet, Woods claims, it is just as absurd to utter (B) as it is to utter (D). Woods concludes from the fact that it is not absurd to assert expressivist Moorean propositions that it is not the case that moral assertions express noncognitive attitudes in exactly the same way as ordinary, descriptive assertions express beliefs. Thus, the parity thesis fails.

4. A Shortcut to the Big Problem

There is a deep problem with Woods’s argument, which is that the first premise of the argument is false. Moore’s paradox is not a good test of the parity thesis because we should
expect assertions of expressivist Moorean propositions to generate the same absurdity as standard Moorean propositions only if the truth of expressivism were transparent to ordinary speakers. If expressivism were transparent in the requisite way, then competent users of moral language would recognize that sincere moral assertions express noncognitive attitudes. They would recognize, for instance, that whenever they believe that murder is wrong, they disapprove of murdering. Or, at the very least, they would recognize that by making sincere moral assertions they incur commitments to having certain noncognitive attitudes. But expressivism is not transparent in this way. Not even self-avowed expressivists think so.

This failure of transparency matters because—as Woods himself acknowledges—‘murder is wrong, but I don’t believe that murder is wrong’ does generate the standard Moorean absurdity. If the correct analysis of moral judgment is broadly expressivist, then the state someone is in when she believes that murder is wrong just is the attitude of disapproving of murdering. If expressivism is true, the difference between ‘murder is wrong, but I don’t believe that murder is wrong’ and ‘murder is wrong, but I don’t disapprove of murdering’ is like the difference between ‘Superman is dreamy, but I don’t believe that Superman is dreamy’ and ‘Superman is dreamy, but I don’t believe that Clark Kent is dreamy’. Ignorant of an important fact about Superman’s true identity, Lois Lane does not realize that whatever she believes of Superman is also true of Clark Kent. This ignorance explains why ‘Superman is dreamy, but I don’t believe that Superman is dreamy’ would strike Lois Lane as absurd while ‘Superman is dreamy, but I don’t believe that Clark Kent is dreamy’ would not. Prior to learning that
Superman is Clark Kent, Lois Lane does not recognize that if Superman is dreamy, then so too is Clark Kent. This fact would not be transparent to her, and we do not expect it to be.\footnote{In fact, it is exactly this failure of transparency that makes possible the whimsical love story that develops between the two.}

Expressivists may appeal to a similar failure of transparency to account for the difference between ‘murder is wrong, but I don’t believe that murder is wrong’ and ‘murder is wrong, but I don’t disapprove of it’. They might claim that, with respect to this difference, ordinary speakers are like Lois Lane before she realizes exactly who’s who. Like Lois Lane, ordinary speakers are ignorant of an important fact about what it is to believe that murder is wrong – namely, that what it is to believe that murder is wrong just is to disapprove of murder.\footnote{Cases of \textit{a posteriori} identity are not the only thing that can lead to a failure of transparency. Such failures can also arise in cases of \textit{a priori} identities, so long as these identities are nonobvious. For example, someone might believe that something is a right triangle without believing that it is a shape whose sides are related according to the Pythagorean theorem. This is something that has been stressed by analytic naturalists such as Frank Jackson (1998) and Steven Finlay (2014), who claim that the truth of a moral statement is determined by the (highly nonobvious) meanings of its relevant terms. So nothing turns on whether expressivists believe that the identity of moral judgments and non-cognitive attitudes is \textit{a posteriori} or \textit{a priori}, and I focus on \textit{a posteriori} identity in the main text only because it is the more intuitive case of the two. Thanks to an anonymous referee for encouraging me to clarify this point.} And as it did in the case of Lois Lane, this ignorance explains why ‘murder is wrong, but I don’t believe that murder is wrong’ strikes ordinary speakers as absurd while ‘murder is wrong, but I don’t disapprove of
murdering’ does not. Things might be different if we lived in a world in which all competent users of moral language accepted the truth of expressivism, or in the very least recognized that sincere moral assertions expressed non-cognitive attitudes. But we do not live in this world. So, it is no surprise that ordinary speakers do not recognize that a sincere assertion of ‘murder is wrong’ expresses the very same attitude that a sincere assertion of ‘I disapprove of murdering’ reports. Thus, like other problematic arguments against expressivism, Woods’s argument rests on the mistaken assumption that if expressivism is true, competent users of moral language must realize that moral judgment necessarily involves some relevant noncognitive attitudes.  

In the remainder of this paper, I elaborate on this line of reasoning in order to explain how and why the failure of transparency surfaces in Woods’s argument from Moore’s paradox. This explanation is worthwhile, for although others have engaged with Woods’s argument, none has focused carefully on the issue of transparency. Since Woods’s case against the parity thesis

To name a few, see Frank Jackson and Phillip Pettit (1998), Terence Cuneo (2006), Jussi Suikkanen (2009), and Jonas Olson (2010). For responses to some of these arguments, see Schroeder (2014) and James Dreier (2004). To be fair, Woods has more to say about the topic of transparency than these other authors, claiming that “ordinary people are … aware that moral assertions are commonly used to condemn and give vent to attitudes” (Woods 2014: 10). This is plausible, but the sort of transparency I’m concerned with requires more. It requires that ordinary people are aware that sincere moral assertions necessarily express (or “give vent to”) non-cognitive attitudes such as condemnation. Thanks to an anonymous referee for urging me to discuss this distinction.

See the citations in Footnote 2.
focuses on the absurdity of standard Moorean propositions à la (D), it is worth thinking carefully about his explanation in standard cases of Moore’s paradox. In the next section, I review Woods’s explanation of these cases argue that this explanation is ultimately incomplete since it fails to account for all the phenomena that need accounting for. After that, I will return to emotive Moorean propositions like (A) and (B) to see whether the explanation I offer of the standard case provides insight there, too.

5. Woods’s Explanation of Moore’s Paradox

Given that his entire argument against the parity thesis is based on Moore’s paradox, it is surprising how little Woods says about what generates the absurdity in standard cases like (C) and (D). It might well be true that the expression relation that holds between a sincere assertion of ‘p’ and the speaker’s belief that p is what accounts for the absurdity in standard cases of Moore’s paradox, but Woods hasn’t given us much to see why this is so. From what little he does say, Woods seems to think that what explains the absurdity in standard cases of Moore’s paradox has to do primarily with the effects of asserting a Moorean proposition. Woods states,

The proper explanation of [Moore’s paradox] is that when I assert p, I somehow commit myself to believing p, but not by asserting that I believe p. Because I have committed myself to believing that p, when I go on to say ‘I don’t believe that p’, I have undercut my immediate prior commitment and left my audience in a muddle as to what I was up to in so asserting. (2014: 2)

Woods is probably right that audiences have trouble making sense of assertions of ‘p, but I don’t believe that p’, but this explanation makes it seem as if the deep problem with Moorean propositions is that assertions of them leave the audience confused about what the speaker is
trying to do by asserting what she asserts and so unsure of whether to attribute to her a commitment to \( p \) or a commitment to not-\( p \).\(^{16}\) This explanation of Moore’s paradox is fine as far as it goes, but the trouble is that it does not go far enough. For although Moore did not seem to notice that Moorean propositions also generate a similar paradox for beliefs, it is easy to imagine a scenario in which an individual consciously believes ‘\( p \)’, but I don’t believe that \( p \)’ but does not assert it because, for example, she has a nasty case of laryngitis that prevents her from speaking. She does not confuse her audience—indeed, she has no audience—but her belief is no less absurd simply because she does not express it out loud. But it is hard to see exactly why this is so, if Woods is right that Moorean propositions are absurd because of their effect on the audience. He elaborates,

\[
\ldots \text{when I sincerely assert } p, \text{ I commit myself to, though I do not assert, my belief that } p.\]

\[
\text{It is this feature that gives rise to the incoherence of my utterance of the form } 'p', \text{ but I}
\]

\(^{16}\) I qualified with ‘probably’ here for the simple reason that I am not actually sure just how immediately obvious the absurdity of Moore’s paradox is to ordinary speakers and audiences. I took it upon myself to toss an assertion of a Moorean proposition into a casual conversation at my local climbing gym and observe how my audience reacted. Perhaps unsurprisingly, my climbing partners were exceptionally good at applying the principle of charity and updating their interpretation of what I’d said accordingly (“oh, she must mean that before she looked out the window, she believed that it wasn’t raining when in fact it was”). Getting people to see the absurdity of what I’d said actually took a bit of poking and prodding. No doubt, my sample size is very small, but I suspect that many audiences are like mine in this way.
don’t believe that $p$’. The incoherence is a result of my incurring a commitment I explicitly deny I meet. (Woods 2014: 2)

This is as close as Woods comes to recognizing that what explains the absurdity in standard cases of Moore’s paradox will bottom out at something psychological, such as “incurred commitments.” Unfortunately, all that Woods says about what ultimately explains the absurdity in standard cases of Moore’s paradox is that there is something incoherent about incurring a commitment I explicitly deny I meet. So, Woods’s solution to Moore’s paradox depends on, but leaves entirely unexplained, the idea that there is something incoherent about the commitments a speaker incurs by asserting a Moorean proposition. And worse still, Woods never explains what exactly a commitment is, what it is to incur one, and what (if anything) distinguishes incurred commitments from beliefs. So we cannot simply appeal to these concepts in order to put together, on Woods’s behalf, an account of why it is incoherent to incur commitments that seem to contradict each other. I conclude that Woods’s solution to Moore’s paradox is at best incomplete.

6. Belief and Moore’s Paradox

Call assertions of Moorean propositions Moorean assertions and conscious believings of them Moorean beliefs.  17 In the previous section, we saw that Woods’s solution to Moore’s

17 The reference to conscious beliefs in Moorean propositions is necessary since it appears to be a consensus position in the literature that such beliefs are only absurd when they are consciously held (Rosenthal 1995a; 1995b; Kriegel 2004). This makes sense, given the following sort of example (adapted from Williams 2015a; 2015b). A painful conversation with a psychologist helps Peter exhume a repressed belief that his father physically abused him that persists in the
paradox fails to generalize from Moorean assertion to Moorean belief. Moorean beliefs are absurd whether or not they are given linguistic expression through the speech act of assertion. Since the absurdity generated by Moore’s paradox is not purely a linguistic phenomenon, purely linguistic solutions to Moore’s paradox will fail to generalize in a way that accounts for the absurdity of Moorean beliefs. And this is a problem. An adequate solution to Moore’s paradox must explain the absurdity of both Moorean assertions and Moorean beliefs (Shoemaker 1995: 213).

In fact, an adequate solution to Moore’s paradox will do more than generalize from assertion to belief; it will also explain the absurdity of Moorean assertions in terms of the absurdity of the Moorean beliefs that they express. There is, after all, nothing wrong with face of his conscious conviction that he did not. Before Peter’s visit to his psychologist, Peter held the pair of beliefs ‘my father physically abused me’ and ‘I don’t believe that my father physically abused me.’ But Peter was not conscious or aware of the former belief, and so it seems unduly harsh to judge that Peter was guilty of absurdity even before he walked into his psychologist’s office. Indeed, it is controversial whether Peter could even have been attributed an unconscious belief in the Moorean proposition ‘my father abused me but I don’t believe it,’ since such a conjunctive belief arguably requires someone to hold both conjuncts in their mind at once—that is, for the belief to be conscious. In any event, I will assume that only conscious Moorean beliefs are absurd, and that an explanation of Moore’s paradox need not have anything to say about unconscious beliefs. But for a discussion of the possibility that people can have unconscious beliefs in Moorean propositions, and that there is something at least “mildly” problematic about such beliefs, see Williams (2010).
uttering Moorean propositions when they do not express Moorean beliefs. An honest flight attendant who does not believe the schedule he’s required to report to the passengers stuck on a delayed flight might conclude his announcement of the schedule with the disclaimer, “Personally I don’t believe it.” What is important for my purposes is just that the absurdity of asserting Moorean propositions can be derived exclusively from the absurdity of believing them. In the remainder of the paper, I follow Sydney Shoemaker’s suggestion that, “an explanation of why one cannot assert a Moore-paradoxical sentence will come along for free, via the principle that what one can believe constrains what one can assert” (1995: 213). With Shoemaker, I assume that an account of the absurdity of Moorean beliefs is explanatorily prior to an account of the absurdity of Moorean assertions. More specifically, I assume that an account of the absurdity of consciously held Moorean beliefs is explanatorily prior to an account of the absurdity Moorean assertions. Or, in any event, expressivists who adopt this strategy can handily avoid Woods’s objection.

7. **The Absurdity in the Standard Case of Moore's Paradox, (D)**

Most accounts of the absurdity of Moorean belief diagnose it in terms of a manifestation of a certain kind of irrationality. Someone who believes a Moorean proposition violates a rational requirement that governs the relation between the belief that \( p \) and the belief that \( \neg p \). Namely:

[No Contradictions] One cannot rationally hold the conscious belief \((p & \neg p)\).

\footnote{This sort of case comes from Wittgenstein, who discusses a railway official who does not believe the schedule he must report to the passengers. He announces, “Train No. … will arrive at … o’clock. Personally I don’t believe it” (Wittgenstein 1980: 93).}
Accounts that adopt this strategy attempt to show that given certain principles governing conscious belief, someone who believes a Moorean proposition has a belief with self-contradictory content. According to a simple version of this view—inspired by Shoemaker (1995), Thomas Baldwin (1990), Uriah Kriegel (2004), David Rosenthal (1995a; 1995b), and Jonathan Williams (2015a; 2015b)—believing a Moorean proposition violates No Contradictions because of the following two principles governing conscious belief:

**[Conjunction]** One consciously believes\(^\text{19}\) that \((p \& q)\) if and only if one both consciously believes that \(p\) and one consciously believes that \(q\).\(^\text{20}\)

**[Transparency of Belief]** If one consciously believes that \(p\), then one consciously believes that one believes that \(p\).\(^\text{21}\)

---

\(^{19}\) Providing an explanation of what it is for a belief to be conscious is part of, well, a notoriously Hard Problem, so I would prefer not to commit myself to a particular explanation of conscious belief, at least if I can help it. The claim that conscious belief collects over conjunction (in the sense that I explain just down the page) will seem more or less controversial depending on one’s preferred explanation of what it is for a belief to be conscious.

\(^{20}\) It is implausible that belief is closed under entailment *in general*, but the idea that believing a conjunction implies believing the first conjunct is an especially plausible form of doxastic closure. Or, at any rate, philosophers working on Moore’s paradox (and surrounding issues) seem to take for granted that suitably qualified (conscious, first-order, *de se*) beliefs are closed under entailment in this way. See Williams (2015a; 2015b) for more on this.

\(^{21}\) This comes from Rosenthal (1995b) but Baldwin anticipates it in his (1990).
A word on Conjunction and Transparency of Belief is in order. According to Conjunction, conscious belief both *distributes* and *collects* over conjunction. Suppose I consciously believe that my shoes are green and that Phoenix is the capital of Arizona. If conscious belief *distributes* over conjunction, then it follows that I consciously believe that my shoes are green and I consciously believe that Phoenix is the capital of Arizona. If conscious belief *collects* over conjunction, then if I consciously believe that the Arcade Fire is from Montreal and I consciously believe that San Francisco is in California, it follows that I consciously believe that the Arcade Fire is from Montreal and that San Francisco is in California.\(^\text{22}\) According to Transparency of Belief, on the other hand, if I consciously believe that Istanbul is in Turkey, then I consciously believe that I believe that Istanbul is in Turkey.\(^\text{23}\)

\(^{22}\) For example, one might worry that the Lottery Paradox causes trouble for the claim that conscious belief collects over Conjunction, but these sorts of cases are not relevant to the issue at hand.

\(^{23}\) It does *not* follow from any suitably qualified version of Transparency of Belief that if I consciously believe that I believe that Istanbul is in Turkey, then I consciously believe that I consciously believe that I believe that Istanbul is in Turkey. Thus, Transparency of Belief may be rendered *narrow* in that the antecedent may be restricted to first-order beliefs to ward off the threat of this infinite series of beliefs. I left out this plausible qualification (as well as another, according to which the content of the belief in the antecedent is necessarily *de se*) because exactly how best to formulate these qualifications is controversial, and because this does not matter much for my purposes. See Kriegel (2004) and Williams (2015b) for discussion. Also worth noting is that Transparency of Belief does not license us to replace the believed content in
With these two principles in hand, it is possible to show that someone who consciously believes a Moorean proposition violates No Contradictions. For example, suppose I consciously believe ‘it’s raining, but I don’t believe that it’s raining’. By Conjunction, I thereby have a conscious belief that it’s raining and a conscious belief that I don’t believe that it’s raining. By Transparency of Belief, I therefore have the conscious belief that I believe that it’s raining. By Conjunction, I thereby have a conscious belief with the self-contradictory content: ‘I believe that it’s raining and I don’t believe that it’s raining’. In other words, someone who consciously believes ‘it’s raining, but I don’t believe that it’s raining’ must, by Conjunction and Transparency of Belief, consciously believe both that she believes that it’s raining and that she doesn’t believe that it’s raining. But that’s a contradiction! So in standard cases, Moorean beliefs are absurd because they are manifestations of irrationality: they violate No Contradictions. And because of the expression relation that holds between sincere assertions and beliefs—because I express my belief that $p$ by sincerely asserting that $p$—if Moorean beliefs are absurd then Moorean assertions are as well. This is the absurdity that is characteristic of standard cases of Moore’s paradox such as (D).

This basic structure lies behind many popular solutions to Moore’s paradox, though, of course, not everyone accepts this argument exactly as it is written. Philosophers squabble over the antecedent of Transparency of Belief with a co-referential expression in the consequent. For example, while Transparency of Belief implies that if I consciously believe that superman can fly, then I consciously believe that I believe that superman can fly, it does not imply further that I consciously believe that I believe that Clark Kent can fly. Thanks to an anonymous referee for raising this possibility.
exactly how to formulate the principles governing conscious belief, and in particular, exactly how to qualify Transparency of Belief to make plausible the transition from believing that $p$ to believing that one believes that $p$. Shoemaker, for instance, restricts the beliefs governed by the principle to those that are “available” to the believer in a particular way. When beliefs are available in this way, the believer is disposed to do things such as say “I believe that $p$.” (Shoemaker 1995: 218, 225–226). According to Shoemaker, these sorts of dispositions themselves constitute believing that one believes that $p$, which is supposed to suffice to explain why the transition is a kosher one.\(^{24}\) Kriegel (2004), on the other hand, along with Baldwin (1990), Rosenthal (1995a; 1995b), and Williams (2006; 2010) all focus explicitly on conscious beliefs, and maintain that for the transition from believing that $p$ to believing that one believes that $p$ to fall under the scope of the principle, the belief in the antecedent of Transparency of Belief must be first-order, and the belief in the consequent of the Transparency of Belief must be \textit{de se}.

One might wish to reject my response to Woods on the grounds that Conjunction and Transparency of Belief cannot be the real explanation of the absurdity of Moorean beliefs because they are implausibly strong and controversial without suitable qualifications of the sort just mentioned. For my purposes here, however, I do not need to assume that the principles hold

\(^{24}\) It also explains why Shoemaker offers Transparency of Belief as a \textit{biconditional}: necessarily, if one is rational, then one believes that $p$ if and only if one believes that one believes that $p$. Notice also that for Shoemaker, Transparency of Belief is not merely a descriptive claim about conscious belief, but a rational requirement. I return to the possibility of treating this principle as a rational requirement later.
in full generality, but only that they hold in the specific instances where Moorean propositions sound absurd, since this is the sort of linguistic data Woods’s argument relies on. And perhaps they need not hold even in all such cases. If what we want to explain is why expressions of Moorean beliefs sound absurd, all we need to hold is that the audience to whom the absurdity is apparent believes that the speaker forms beliefs in accordance with Conjunction and Transparency of Belief in these cases. In other words, Transparency of Belief can explain why expressions of Moorean beliefs sound absurd if the principle is a claim about the beliefs a linguistically competent audience would attribute to the speaker on the basis of her utterance. In this case, though we are no longer explaining the absurdity of a Moorean assertion in terms of the absurdity of the belief it expresses, we are still explaining the apparent absurdity of a Moorean assertion in terms of the absurdity of the belief it apparently expresses, thus retaining the spirit of Shoemaker’s suggestion that we explain the absurdity of a Moorean expression in terms of the absurdity of Moorean belief.²⁵

8. The Absurdity in the Emotive Cases of Moore’s Paradox, (A)-(B)

One advantage of the explanation just given is that it can be straightforwardly generalized to provide a unified explanation of Woods’s emotivist Moorean propositions, (A) and (B). Recall Woods’s claim that it is possible to reproduce the same absurdity present in standard Moore-paradoxical constructions using utterances such as ‘Go Red Sox!’ rather than sentences in the indicative. For example, Woods thinks that a sincere utterance of ‘Go Red Sox!’ conjoined with ‘I don’t support the Red Sox’ yields the absurdity we find when we conjoin a sincere utterance of ‘p’ with ‘I don’t believe that p’.

²⁵ Thanks to an anonymous reviewer for pressing me on this point.
My explanation of the absurdity in standard cases of Moore’s paradox can extend to these cases if we replace Transparency of Belief with the more general principle:

**[Transparency of Attitudes]** If one consciously holds an attitude, then one consciously believes that one has that attitude.

Transparency of Attitudes entails Transparency of Belief, but applies to all consciously held attitudes, rather than only consciously held beliefs. So suppose I sincerely utter ‘Fuck the Yankees! I don’t have any negative attitudes toward the Yankees’, thereby expressing two distinct attitudes: (i) my consciously held negative attitudes toward the Yankees and (ii) my conscious belief that I don’t have any negative attitudes toward the Yankees. Applying Transparency of Attitudes to (i) yields a conscious belief that I have negative attitudes toward the Yankees. So, I consciously believe that I have negative attitudes toward the Yankees and I consciously believe that I don’t have any negative attitudes toward the Yankees. By Conjunction, I thereby have a conscious belief with self-contradictory content: I consciously believe ‘I have negative attitudes toward the Yankees and I don’t have any negative attitudes toward the Yankees,’ and this violates No Contradictions. Thus, if Transparency of Attitudes holds, we have a straightforward explanation of why emotive Moorean propositions generate the same absurdity we find in standard cases of Moore’s paradox: they violate No Contradictions and therefore are manifestations of irrationality.

Transparency of Attitudes is stronger than Transparency of Belief, so if it is implausible that Transparency of Belief holds in full generality, then this should be even more implausible in the case of Transparency of Attitudes. But, once again, it is enough to appeal to the fact that this principle holds only in the specific instances where emotive Moorean propositions are absurd, or
at the very least that in cases where emotive Moorean assertions sound absurd, we may explain this apparent absurdity by pointing to the fact that audiences attribute attitudes in accordance with Transparency of Attitudes in these contexts. For example, it seems clear enough that when someone sincerely utters ‘Go Red Sox!’ or ‘Fuck the Yankees!’ we do indeed attribute to the speaker a conscious belief that they hold a positive or negative noncognitive attitude towards a baseball team, as Transparency of Attitudes predicts. It is an interesting question why Transparency of Attitudes holds for some sincere expressions of attitudes but not others, or why we attribute beliefs in this way, and I will briefly return to this question in the next section. But what matters for now is that in cases where we do form or attribute attitudes in accordance with Transparency of Attitudes we have an explanation of the absurdity or apparent absurdity characteristic of Moore’s paradox. On the other hand, when Transparency of Attitudes fails, we should not expect the absurdity characteristic of Moore’s paradox, and we have an explanation of why: when Transparency of Attitudes fails, there is no way to derive a belief with self-contradictory content.

9. Expressivism and the Moral Case of Moore's Paradox

We are now in a good position to see how and why the failure of the transparency of expressivism to ordinary speakers surfaces in an analysis of Moore’s paradox. Suppose first we attempt to derive this absurdity by relying only on Transparency of Belief. Consider the expressivist Moorean proposition (E), ‘murder is wrong, but I don’t disapprove of murdering’. The problem is that expressivist Moorean propositions like (E) do not express beliefs with self-contradictory content. This is because the belief that one believes that murder is wrong is not the negation of the belief that one doesn’t disapprove of murdering. Following the steps of our
previous derivation, we are able to show that someone who consciously believes (E) believes ‘I believe that murder is wrong and I don’t disapprove of murdering’. But that’s not a contradiction! So unlike a standard Moorean proposition, someone who believes an expressivist Moorean proposition does not necessarily violate No Contradictions, at least if we constrain ourselves to Transparency of Belief.

What about Transparency of Attitudes? Believing an expressivist Moorean proposition would violate No Contradictions if both expressivism and Transparency of Attitudes were true. For according to expressivism, what it is to consciously believe that something is wrong is to consciously disapprove of it. And according to Transparency of Attitudes, when one consciously disapproves of something, one consciously believes that one disapproves of it. So together, expressivism and Transparency of Attitudes imply:

[Transparency of Moral Judgment] If one consciously believes that \( x \) is wrong, then one consciously believes that one disapproves of \( x \).

For example, whereas Transparency of Belief implies that if one consciously believes that murder is wrong, then one consciously believes that one believes that murder is wrong, Transparency of Moral Judgment implies further that one consciously believes that one disapproves of murder. And when combined with Conjunction, Transparency of Moral Judgment does yield a conscious belief with self-contradictory content: someone who consciously believes ‘murder is wrong, but I don’t disapprove of murdering’ must, by Conjunction and Transparency of Moral Judgment, consciously believe ‘I believe that I disapprove of murdering and that I don’t disapprove of murdering’. But that’s a contradiction, and believing it violates No Contradictions. Thus, if Transparency of Moral Judgment plausibly governs our attitudes, then Woods is right:
we should expect expressivist Moorean propositions to generate the absurdity as standard Moorean propositions. And furthermore, as before, we should expect the appearance of absurdity even in cases where Transparency of Moral Judgment fails to hold as a matter of fact, so long as it at least describes the attitudes a linguistically competent audience would attribute to a speaker on the basis of her utterance. If, when someone sincerely utters an expressivist Moorean proposition, her audience attributes to her attitudes in accordance with Transparency of Moral Judgment and Conjunction, then this utterance will appear absurd, since the audience will attribute to the speaker a belief with self-contradictory content.

But as we have already seen, no self-respecting expressivist would accept Transparency of Moral Judgment, either interpreted as a descriptive claim about the attitudes people actually have or a principle governing the attributions of attitudes, and for good reason. Transparency of Moral Judgment licenses replacing the explicitly moral statement ‘I believe that murder is wrong’ with its expressivist analysis ‘I disapprove of murder’. So, according to Transparency of Moral Judgment, whenever someone consciously believes that something is wrong, she thereby believes that she disapproves of it, or at least her audience attributes attitudes to her in this way. In other words, if Transparency of Moral Judgment were true, then it would be transparent to someone who believes that murdering is wrong, or transparent to her audience, that she disapproves of murdering. But expressivists explicitly deny this. They explicitly deny that the truth of expressivism is transparent, as it would be if we necessarily believed either that we disapprove of everything that we believe is wrong or that others necessarily disapprove of everything that they believe is wrong (see Blackburn 1992: 172–173; 1996). And furthermore, expressivists allow that even competent users of moral language fail to form attitudes in
accordance with Transparency of Moral Judgment, and that we fail to attribute attitudes to
competent users in his way.\textsuperscript{26} So Transparency of Moral Judgment is neither true as a matter of
fact, nor is it true as a necessary condition for linguistic competence, nor is it true as a principle
governing our attribution to (linguistically competent) speakers. And since Transparency of
Moral Judgment is essential to explain why believing an expressivist Moorean proposition
violates No Contradiction, we have a perfectly good explanation of why these propositions fail to
generate the characteristic absurdity. This is where the assumption of transparency surfaces in an
analysis of Moore’s paradox.\textsuperscript{27}

Still, one might wish to revive Woods’s objection by insisting we think instead of
Transparency of Moral Judgment as a \textit{rational requirement}:

\begin{quote}
\textbf{[Rational Transparency of Moral Judgment]} If one is rational and if one consciously believes
that $x$ is wrong, then one consciously believes that one disapproves of $x$.
\end{quote}

\textsuperscript{26} Woods seems to think that being linguistically competent with moral language requires
satisfying Transparency of Moral Judgment. Responding briefly to the objection that ordinary
speakers find expressivist Moorean assertions felicitous because they do not know that
expressivism is the correct account of moral discourse, Woods suggests that competent speakers
tacitly acknowledge that by making moral assertions we incur commitments to having certain
noncognitive attitudes, and that, given the truth of expressivism, this is not

\textsuperscript{27} Thanks to two anonymous referees as well as the editors of \textit{Ergo} for encouraging me to clarify
precisely where the failure of transparency arises in the analysis of Moore’s Paradox.
Rational Transparency of Moral Judgment is a putative rational requirement linking conscious moral beliefs to conscious beliefs about one’s own noncognitive attitudes. According to Rational Transparency of Moral Judgment, someone who consciously believes that murder is wrong will, if she is rational, consciously believe that she disapproves of murdering. While we cannot use Rational Transparency of Moral Judgment to show that someone who consciously believes an expressivist Moorean proposition has a conscious belief with self-contradictory content, we can use it to show that someone who consciously believes an expressivist Moorean proposition necessarily violates a requirement of rationality: either she fails to satisfy Rational Transparency of Moral Judgment or else she fails to satisfy No Contradictions. According to this view, then, Moorean assertions are absurd because they are manifestations of this irrationality. And thus, if Rational Transparency of Moral Judgment is true, we should expect expressivist Moorean propositions and standard Moorean propositions to generate the same absurdity.

28 Declan Smithies offers an alternative account of Moore’s paradox which, rather than attempting to derive a conscious belief with self-contradictory content, attempts to show that belief in the first conjunct justifies belief in the negation of the second conjunct. This is an interesting strategy, but for my purposes, what matters is that Smithies argument also relies on a belief in the first conjunct of a Moorean proposition being a negation of the second conjunct. See Smithies (2012).

29 In fact, one might think of Transparency of Belief or Conjunction as rational requirements, too. This is Shoemaker’s own view (1995). (For more on Shoemaker’s view, see Footnote 27.) Nothing turns on this.
Contemporary defenders of expressivism have not, to my knowledge, said anything about Rational Transparency of Moral Judgment, but there are good reasons for them to reject it (and the associated principle that we attribute people attitudes in accordance with it). Unlike Transparency of Moral Judgment, Rational Transparency of Moral Judgment does not rely on the assumption of transparency, but it does require that failures of transparency are irrational. And it does not take much to see why one might wish to reject this interpretation of the principle on the grounds that it undermines the rationality of our moral practices. Defenders of expressivism claim that Transparency of Moral Judgment is false: they claim that ordinary people who are competent users of moral language systematically fail to grasp fully the nature of the attitudes that their moral judgments (or the moral judgments of others) express. But to reject Transparency of Moral Judgment while accepting Rational Transparency of Moral Judgment is to label all of these people irrational. This is a bad thing because it certainly seems as though people are able to engage rationally in our moral practices, irrespective of their metaethical views. And it is likely to be especially unattractive to defenders of expressivism, who take pains to show that their view accommodates all the features of our moral practices.\textsuperscript{30} It would be a Pyrrhic victory for

---

\textsuperscript{30} One of these features is precisely that we express our disapproval of murder not by shouting ‘Boo to murder!’ but rather by asserting ‘murder is wrong’. Defenders of expressivism claim that their view can accommodate our use of these moral assertions. Another of these features is the tight connection between sincere assertion and belief according to which sincerely asserting that \( p \) implies that the speaker believes that \( p \). Defenders of expressivism also claim that their view can accommodate that sincerely asserting a moral proposition such as ‘murder is wrong’ implies
expressivism if it could accommodate these important features of our moral practices only by labeling nearly all of its participants irrational, or, on the attributional version of the principle, by holding that we go around labeling people irrational in this way. So expressivists have ample reason to reject Rational Transparency of Moral Judgment, independently of its implications for Moore’s paradox.

To sum up, I have argued that there is a straightforward explanation of why expressivist Moorean propositions do not generate the absurdity characteristic of standard cases of Moore’s paradox. One does not necessarily violate a requirement of rationality by believing an expressivist Moorean proposition because it need not be transparent to someone who believes that murdering is wrong that she disapproves of murdering, even if she is rational and a competent user of moral language; furthermore, audiences do not attribute attitudes in this way. So Transparency of Moral Judgment fails as a descriptive claim, as a requirement for linguistic competency or rationality, and as a principle by which we attribute attitudes to speakers. All of this is compatible with the plausibility of Transparency of Belief, which is all we need to generate the absurdity in the standard case. Thus, we may explain why standard Moorean propositions are absurd and expressivist Moorean propositions are not without rejecting the parity thesis.

In our explanation of the absurdity of emotive Moorean propositions, however, we relied not just on Transparency of Belief, but Transparency of Attitudes. And, as we have seen, if expressivism were true, then Transparency of Moral Judgment would follow from Transparency that the speaker believes that murder is wrong. The critical expressivist move here is to adopt a highly deflationary understanding of belief that falls out of endorsing semantic minimalism.
of Attitudes. So we return to the earlier question of why Transparency of Attitudes holds for some conscious attitudes and not others, as expressivists must insist if they are to acknowledge the absurdity of emotive Moorean propositions without claiming that expressivist Moorean propositions are absurd as well. Why is it that some consciously held attitudes—for example, beliefs in typical cases, and negative and positive attitudes expressed by cheering on or cursing at sports teams—are transparent whereas others—most notably, the noncognitive attitudes that are expressed by moral utterances—are not? Is this just an ad hoc distinction, posited only to save the expressivist research program? Or are there more principled grounds for distinguishing between cases where Transparency of Attitudes does and does not hold?

Although I cannot provide a full explanation of this phenomenon here, I suspect that a plausible explanation will have something to do with our social and linguistic practices, and the different competency and sincerity conditions that attach to these practices. When someone utters an interjection like, ‘Fuck the Yankees!’ and we take her to be sincere, we do attribute to her a strong negative attitude toward the Yankees, which, we assume is just as transparent to her as it is to us. Someone who says, “Fuck the Yankees!” without consciously believing that she has negative attitudes towards the Yankees, is, we think, either insincere or else incompetent when it comes to cheering for or against sports teams. But it is precisely this that expressivists deny in the case of moral language. Expressivists deny that, as a matter of social and linguistic practice, we necessarily attribute to sincere, competent moral speakers who utter ‘Murder is wrong’ an attitude of disapproval toward murdering that is necessarily transparent to her. Whereas someone cannot sincerely and competently engage in sports discourse without understanding that cheering for and against sports teams expresses positive or negative attitudes, expressivists allow that
someone can sincerely and competently engage in moral discourse without believing that her expressions of moral attitude amount to expressions of disapproval rather than, say, ordinary beliefs. Otherwise, expressivists would have to insist that anyone who uses moral language without subscribing to expressivism is necessarily either incompetent or insincere. Perhaps some hold these views in private company, but it is not the official party line.

10. Conclusion

Woods argues that Moore’s paradox reveals that a central tenet of expressivism, the parity thesis, is false. The arguments in this paper have been primarily defensive. I have shown that defenders of expressivism who accept the parity thesis have at their disposal a good response to the argument Woods advances against their view. The response is to reject the first premise of the argument and deny that Moore’s paradox is a good test of the parity thesis. In standard cases, Moorean beliefs are absurd because by believing ‘\( p \), but I don’t believe that \( p \)’, one has a conscious belief with self-contradictory content, which violates a rational requirement; and a similar analysis holds for emotive Moorean propositions. On the other hand, one does not necessarily violate a rational requirement by believing an expressivist Moorean proposition because Transparency of Moral Judgment fails as a descriptive claim, a condition for linguistic competence, and a rational requirement. It need not be transparent to someone who believes that murdering is wrong that she disapproves of murdering, even if she is rational and a competent user of moral language. So the parity thesis is out of the Woods: expressivists may continue to maintain that moral assertions express noncognitive attitudes in just the way that ordinary, descriptive assertions express beliefs.
Nevertheless, Woods’s argument invites—indeed requires—both friends and foes of expressivism to think carefully about the transparency of our moral attitudes. What explains why some sincere expressions of attitudes (such as interjections, and perhaps pejoratives) are transparent whereas others are not? What explains why some failures of transparency seem to evince linguistic incompetence or irrationality whereas others do not? These are interesting and important questions for expressivists, to be sure, but they are no less pressing for those of us who find ourselves unsure of what view to accept. I have suggested the beginnings of an answer rooted in our social and linguistic practices, but more carefully developed answers are needed to help clarify the features of moral discourse that any plausible metaethical view must accommodate.

---

31 See Joyce (2006: 51–58) for an interesting discussion of transparency and pejoratives.
Works Cited


DOES INTERNALISM SUPPORT NONCOGNITIVISM?

1. **Introduction**

One fundamental feature of normativity is the action-guiding or practical nature of normative judgments. Normative judgments typically motivate the people who make them. For example, suppose Alex claims that she morally ought to donate to Oxfam, yet refuses to donate when the opportunity arises just a few minutes later. Many of us would find Alex’s refusal puzzling. We would look for an explanation: perhaps Alex wants to hold off on donating until her next paycheck, when she has a bit more disposable income, or maybe she has a stronger desire to spend her money elsewhere. Without any such explanation, though, we may come to doubt Alex’s sincerity. We may come to think that Alex did not really judge that she ought to donate. Motivational internalism (henceforth, “internalism”) is the thesis that necessarily, if a person makes a sincere normative judgment of the form 'I ought to phi' then she is motivated to act accordingly. This view ties normative judgment and motivation together with a necessary connection: if someone is not motivated by a judgment she makes, then whatever else is true of that judgment, it cannot be a normative judgment, because normative judgments necessarily motivate.¹

¹ Externalists, on the other hand, claim that while it may be true that most people tend to be motivated to comply with their normative judgments, what explains this phenomenon is not the nature of normative judgments, but rather, some contingent fact about us (e.g. the fact that most of us tend to care about what’s good, right, and so on).
Like other necessary connections, there is keen philosophical interest in accounting for and explaining the one that internalists allege holds between normative judgment and motivation. The conventional wisdom among metaethicists is that noncognitivism has an especially easy time accounting for and explaining internalism, and that this is an advantage of the view. Cognitivism and noncognitivism are competing views about what kind of state of mind a normative judgment is. Cognitivists think that normative judgments are beliefs, or belief-like states with cognitive content. The difference between a judgment like ‘I donate to the Red Cross’ and ‘I ought to donate to the Red Cross’ lies, according to cognitivists, in their different contents, not in their attitude-types. Noncognitivists deny this. They think that normative judgments are desires, or desire-like states. A judgment such as ‘I donate to the Red Cross’ differs from a judgment like ‘I ought to donate to the Red Cross’ not because the judgments have different contents, but because the former is a cognitive, belief-like state, and the latter is a non-cognitive, desire-like attitudes. This is supposed to provide a simple, elegant, and direct explanation of the necessary connection between normative judgment and motivation. Desire-like attitudes, after all, have a necessary connection to motivation. So if normative judgments just are desire-like attitudes, this explains why normative judgments have a necessary connection to motivation as well.

In this paper, however, I argue that the conventional wisdom is wrong. Noncognitivists still haven’t provided a plausible version of internalism that gives them the distinctive explanatory advantage they purport to have. An important upshot is that the argument from internalism to noncognitivism is far less compelling than noncognitivists have let on. Although those who defend noncognitivism on other grounds have good reasons to accept some version of
internalism, one cannot argue from internalism to noncognitivism. The debate between these two views must be settled on other grounds.

2. **Noncognitivism**

Noncognitivism is a brand of metaethical antirealism according to which normative judgments do not express cognitive states such as beliefs, or make statements that could be true or false in a robust, non-deflationary sense. Instead, normative judgments express desires or desire-like states. One supposed advantage of noncognitivism is its elegant and direct explanation of internalism. The explanation goes roughly like this. Noncognitivism maintains that sincere normative judgments of the form 'I ought to phi' or 'phi-ing is right' express desires, or “desire-like attitudes” such as approval, disapproval, acceptance, and so on, rather than belief-like attitudes that are 'cognitive', 'representational' or 'descriptive' in character. According to a popular Humean theory of motivation, having a desire, in the presence of other relevant beliefs, is both necessary and sufficient to be to some extent motivated to perform some action. When combined with the Humean theory of motivation, noncognitivism seems to provide a straightforward explanation of motivational internalism. Normative judgments are always to some extent motivating because normative judgments just are desires or desire-like attitudes, and such attitudes are necessarily motivating.²

² It is important to distinguish noncognitivism from cognitivist speaker subjectivism. According to that view, a sincere assertion of ‘lying is wrong’ has the same meaning—or at least, the same truth conditions—as a sincere assertion of ‘I disapprove of lying’. Cognitivist speaker subjectivism is a view about the content of normative judgments. Noncognitivism, on the other hand, is a view about what kind of mental state normative judgments are and says nothing about
Defenders of noncognitivism usually arrive at the view because they are unhappy with the alternative view, cognitivism. Cognitivism is a family of metaethical views according to which the meaning of normative sentences like 'lying is wrong' are like the descriptive meanings typical of non-normative declarative sentences (such as ‘lying is common’). Cognitivism combines this thesis about normative language with a corresponding thesis about normative thought according to which normative judgments, such as, say, judging that lying is wrong, are not so different from ordinary descriptive judgments such as, say, judging that lying has a certain property, the property of wrongness. But all sorts of difficulties arise when we try to get clear on exactly what the property of wrongness is. What kind of property is wrongness? How does it relate to the natural properties instantiated by the action being described as wrong? Is it reducible to those natural properties? How do we come to know whether things have the property? Is there a way to confirm whether something does or does not have the property? And how in the world could judging that something has a property—natural or non-natural—be motivating? These are difficult questions to answer. The difficulty of answering them led Mackie to the radical position that normative judgments are systematically false because they presuppose the existence of some rather queer entities that are never in fact instantiated—normative properties.  

But similar concerns about the alleged queerness of normative properties led other philosophers to reject the presupposition on which they depend, namely, the idea that normative judgments predicate properties. This is the strategy adopted by noncognitivism.

For a careful discussion of the differences between these two views, see Mark Schroeder’s (2008a, 2008b, 2014).

3 See Mackie (1977), especially pp. 23, 40, 49.
3. The Noncognitivist Argument

Noncognitivism is also supposed to have this important advantage when it comes to explaining moral motivation, or the motivational efficacy of normative judgments. Earlier I sketched a bare bones picture of how this is supposed to go. If normative judgments are desires, then the fact that they motivate us the way that they do is not so mysterious. Common sense has it that desire and motivation are inextricably linked; that to desire something is a matter of being motivated to attain it; that behind motivated behavior lies a desire. In other words, common sense subscribes to the standard theory of desire according to which to desire that P is to be disposed to bring it about that P: that providing motivation is precisely what desires are for, or at least one of the things desires are for. So we can and should explain the motivational efficacy of normative judgment by appealing to this feature of desires. And one way to do that, while holding onto the idea that there is some kind of conceptual necessity between normative judgment and motivation, is to hold that normative judgments just are desires, or desire-like states. Thus there seems to be some sort of argument from motivational internalism to noncognitivism.

Philosophers often point to Hume as the first proponent of this strategy, and tend to cite one passage in particular. Hume states,

---

4 More carefully, the claim is that intrinsic desires—desires for things for their own sake, rather than for the sake of something else—motivate. I brush over this distinction in this main text, but see, for example, Schroeder (2004).

5 But see Schroeder (2004) for a comparison of various theories of desire.
Since morals, therefore, have an influence on the actions and affections, it follows, that they cannot be deriv’d from reason; and that because reason alone, as we have already prov’d, can never have any such influence. Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular. The rules of morality, therefore, are not conclusions of our reason. (T 3.1.1.6/457)

It is an interesting question whether Hume is best interpreted as a noncognitivist or as a proponent of some other view, but one that I will leave to Hume scholars. What matters for the purposes of this paper is Hume’s claim that “morals excite passions, and produce or prevent actions.” This claim gets interpreted as a statement of the thesis of motivational internalism, which gets combined with the Humean theory of motivation, and trotted out as an argument for noncognitivism.\(^6\) Thus, we arrive at what Russ Shafer-Landau calls “the noncognitivist argument.” It goes:

\(^6\) It is worth pointing out that the word ‘morals’ in this line from Hume is ambiguous. Hume might mean by ‘morals’ moral *judgments*, in which case the claim that ‘morals excite passions’ looks like motivational internalism. But Hume might also mean by ‘morals’ moral *properties*, in which case the claim that ‘morals excite passions’ looks like existence internalism, or the thesis that there is a necessary connection between something’s having a certain moral status and motivation. See Darwall (1983: 54-55) for the distinction between judgment and existence internalism.
1. Necessarily, if you judge that you ought to do something, then you will do it.\(^7\)

2. Taken by themselves, beliefs neither motivate nor generate motivationally efficacious states.

3. Thus, normative judgments are (at least partly\(^8\)) noncognitive states.\(^9\)

I think Shafer-Landau’s noncognitivist argument is helpful for understanding how different views in contemporary metaethics account for moral motivation. But it is controversial: some deny its validity, and others its second premise. There is, however, a different way of

\(^7\) Actually Shafer-Landau’s version of the argument employs a slightly weaker version of motivational internalism according to which, necessarily, if one judges that one ought to phi, then one is motivated to some extent to phi. More on this in what follows.

\(^8\) I say ‘at least partly’ because the conclusion of the argument allows both pure and hybrid forms of noncognitivism. Pure noncognitivists think normative judgments are (and express) noncognitive attitudes, whereas hybrid noncognitivists think normative judgments are (and express) both cognitive and noncognitive attitudes. In what follows, I drop ‘at least partly’ but the conclusion of the argument should be understood to apply to both pure and hybrid noncognitivism. Defenders of hybrid noncognitivism include Copp (2001); Ridge (2006); Barker (2002); Boisvert (2008). For a nice discussion of the virtues and vices of hybrid noncognitivism, see Schroeder (2009).

\(^9\) Shafer-Landau admits that no one lays the argument out as explicitly as this, but nevertheless maintains that it is an accurate representation of an argumentative strategy endorsed by Stevenson (1937: 14–31); Aiken (1950); Hare (1952: 79–93); Nowell-Smith (1954: 36–43); Harman (1975); and Mackie (1977: 27–42).
understanding how motivational internalism is supposed to support noncognitivism. The thought is that noncognitivism provides the best explanation of a fundamental feature of normativity, namely, some sort of motivational internalism. The argument between cognitivism and noncognitivism is then to be understood as an argument over which view best explains motivational internalism. In other words, the argument for noncognitivism from internalism is an abductive one. It goes more like:

1. Necessarily, if you judge that you ought to do something, then you will do it.
2. The best explanation of (1) is that normative judgments are noncognitive states.
3. Therefore, normative judgments are (at least partly) noncognitive states.

The first premise of the abductive argument is the same as the first premise of Shafer-Landau’s deductive argument, but the second premise asserts that the truth of this premise is best explained by noncognitivism rather than asserting that it follows from noncognitivism in conjunction with the Humean theory of reasons. The abductive argument is therefore less controversial than the deductive argument and also, I think, better captures the sort of reasoning employed by most actual defenders of noncognitivism. It is therefore the argument I will focus on here, though, in the end, we will see, that my argument against the abductive argument will also show what is wrong with the deductive argument as well.

Of course, most internalists do not endorse anything as strong as the version of internalism at work in the abductive argument as I have stated it. If this version is true, then, necessarily, there is no gap between judging that you ought to phi and phi-ing, or more carefully, being sufficiently motivated to phi. The kernel of truth in internalism—what seems right about the thesis regardless of your metaethical allegiances—is that normative judgments reliably
motivate the people who make them. This is a far cry from the claim that it is impossible to fail to act on one’s normative judgments. Nevertheless, Gibbard embraces this version of strong internalism in defending quasirealist expressivism. We also see similarly strong theses in R.M. Hare’s defense of prescriptivism and in John Mackie’s famous queerness argument.\textsuperscript{10} Since the abductive argument for noncognitivism works best given strong internalism, this is a good place to start. But strong internalism is implausible; some weaker version is needed. And we shall see that the noncognitivist move is to modify strong internalism, qualifying or weakening the thesis into something more plausible and that can avoid various objections and counterexamples. What is important for my argument is the way that internalism is qualified or weakened, and whether such modifications leave the second premise intact. In the end, we will see that only implausibly strong versions of internalism can sustain the abductive argument for noncognitivism. Plausible versions of internalism do not support noncognitivism, since cognitivism can explain them just as well.

4. \textbf{Strong Internalism}

When combined with the Humean theory of motivation, strong internalism, if it is true, makes for a powerful abductive argument for noncognitivism. My focus will be on Allan Gibbard’s view in \textit{Thinking How to Live}. Gibbard’s view is particularly useful for my purposes because in the process of developing and defending his quasirealist expressivism, he accepts,\textsuperscript{11}

\textsuperscript{10} Mackie (1977)

\textsuperscript{11} At least Gibbard defends this sort of strong internalism for “a crucial sense of ‘ought’” (2003: 152-154).
Strong internalism: necessarily, if you sincerely judge that you ought to do something, then you do it.

According to this version of internalism, someone who sincerely judges that she ought to phi is overridingly motivated to phi. Gibbard’s defense of this thesis also helps make vivid the intuition that drives the abductive argument. So let’s begin with some basic components of Gibbard’s view and then proceed to the defense of strong internalism.

Expressivists like Gibbard deny that we can explain the meaning of a normative judgment in terms of their truth conditions, content, or character. Instead, expressivists tell us what someone is doing when he or she makes a normative judgment, which, on Gibbard’s view, is to express a certain state of mind: a planning state. Planning states are motivational states that are sort of like intentions or conditional preferences. So, for Gibbard, the meaning of a normative judgment is explained in terms of the planning state it expresses. A normative judgment like ‘I ought to phi right now’ expresses a plan to phi, and what it is for me to think that I ought to phi right now is for me to plan to phi. Gibbard’s view is thus supposed to be well placed to explain strong internalism. Normative judgments express plans, and plans are motivational states. Gibbard himself states,

My theory thus yields internalism in a strong form: if I think that something is now the thing to do, then I do it. My hypothesis about ordinary ought judgments is that they are judgments of what to do, of what is the thing to do. I don’t, then, think that I ought right now to defy the bully unless I do defy him. If I fail to defy him, then as a matter of the very concept of ought, I don’t believe I ought to. (2003: 153)
So the necessary connection between normative judgment and motivation falls out of Gibbard’s account of meaning: the state that is necessarily connected to a normative judgment just is the state expressed when that normative judgment is sincere. And a normative judgment that one ought right now to phi is sincere only when the person who makes the judgment phis.

Unfortunately, as Gibbard himself recognizes, strong internalism “seems plain wrong.”\(^\text{12}\) The trouble is that it seems possible to conceive of someone who judges that she ought to phi but fails to phi. The examples are all too familiar. Suppose Lisa recently quit smoking. She is doing great and hasn’t had a cigarette in a whole month. But after a long and difficult day at work, she meets some friends at the bar for drinks and soon enough, her tipsy friend absentmindedly offers her a cigarette. She sincerely judges that she ought to turn down the cigarette, but in her tired and slightly intoxicated state, Lisa cannot help herself. She puts a cigarette in her mouth and lights up. Strong internalism tells us that Lisa either does not \textit{really} judge that she ought to turn down the cigarette or else she is mistaken about the concept of ought. But this seems wrong. It would be nice if the fact that you sincerely judge that you ought to do something guarantees that you do it. But we do sometimes fail to do what we sincerely judge that we ought to do. And when we do, it seems wrong to insist that the motivational failure is possible only in cases of insincerity or conceptual confusion. Much more plausible is the claim that motivational failure is possible even in the absence of insincerity and conception confusion because we are capable of experiencing conflicting motivations. We are somewhat motivated to do what we judge we ought, but also \textit{more} motivated to do something else, where that something else is incompatible with acting in accordance with our initial ought-judgment. Sometimes we are strong-willed and do what we

judge we ought, but sometimes we are akratic or weak-willed: we fall short and fail to do what we judge we ought to do. This is so familiar to our everyday lived experience as persons that it is hard to take seriously a view that rules this sort of thing out as a matter of conceptual competence.

All this spells trouble for the argument from strong internalism to noncognitivism. If it is possible to fail to do what you judge you ought to do, then we should reject any version of internalism according to which this is not possible. We should, in other words, reject Gibbard’s version of strong internalism. Noncognitivism purports to give the best explanation of a wide range of normative phenomena. But if the explanation of moral motivation on offer commits defenders of noncognitivism to denying the possibility of akrasia, then this is hardly an advantage at all. In fact, since strong internalism is the internalist first premise of both the deductive and abductive arguments, both arguments should be rejected as failing to establish the conclusion that normative judgments are noncognitive states. Noncognitivism might have an explanatory advantage when it comes to accounting for strong internalism. But as we have seen this doesn’t count for much! Since accepting strong internalism requires denying the possibility of akrasia, it is not the sort of internalism we want our metaethical theories to account for. The question is whether we can weaken or qualify the thesis in a way that still gives noncognitivism a distinct explanatory advantage when it comes to explaining the necessary connection between normative judgment and motivation. With that question in mind, let’s move on to consider some more plausible versions of internalism.
5. **Weak Internalism**

Strong internalism told us that necessarily, whenever you judge that you ought to do something, you do it. In other words, it told us that sincere normative judgments guarantee that the person making them is sufficiently motivated to do what she judges she ought to do. Problematically, this thesis did not allow for the possibility of failing to do what you judge you ought to do. And, indeed, Gibbard seems to notice the implausibility of strong internalism. In light of it, he retreats to the view that there is just “a crucial sense of ‘ought’” for which it holds and another for which there is a weaker (but still necessary) connection between normative judgment and motivation.\(^\text{13}\) In particular, he suggests,

Weak internalism: necessarily, if you sincerely judge that you ought to do something, then you are *somewhat motivated* to do it.

Whereas strong internalism posits a necessary connection between normative judgment and action, weak internalism posits one between normative judgment and *motivation*. According to weak internalism, someone who makes a sincere normative judgment is always to some extent somewhat motivated to comply with it. Akrasia is possible because being to some extent motivated to do something is different than being all things considered motivated to do something. Motivations come in different strengths and are apt to conflict. So, when Lisa judges that she ought to refuse the cigarette, says the weak internalist, she *is to some extent* motivated to comply, it’s just that her weak motivation stands no chance against the strong craving for a cigarette. So, weak internalism is supposed to allow us to explain motivational failures by

---

\(^{13}\) Ibid.
pointing to conflicting motivations rather than insisting on insincerity or conceptual incompetence.

Unfortunately weak internalism also faces serious objections, this time concerning the conceptual possibility of amoralism. In this literature, the amoralist is the person who makes normative judgments while remaining wholly unmotivated to comply with them. \(^{14}\) Many have argued that amoralists are conceivable, and thus, weak internalism must be false: it is not true that necessarily, a person who judges that she ought to do something is to some extent motivated to do that thing. \(^{15}\) In fact, many argue that there are documented cases of psychopathy, suggesting that amoralism is not merely a conceptual possibility but an actual phenomenon we can encounter in the actual world. \(^{16}\) So the conceivability of amoralism and the existence of psychopaths put pressure on the claim that sincere normative judgments are always, as a matter of necessity, accompanied by motivation.

But less extreme characters cause trouble for weak internalism, too. Many think that certain familiar psychological conditions such as depression, exhaustion, apathy, or other emotional disturbances can undermine the supposedly necessary connection that holds between normative judgment and motivation, and that the absence of motivation is plausibly a symptom of the very condition from which they suffer. \(^{17}\) Take someone suffering from severe depression

\(^{14}\) See Brink (1989, 1997); Svavarsdóttir (1999 and 2006)

\(^{15}\) Shafer-Landau (2003: 146)

\(^{16}\) Nichols (2004: 65-82)

\(^{17}\) Brink (1989: Chapter 3); Mele (1996); Miller (2008); Roskies (2003); Stocker (1979); Svavarsdóttir (1999)
or extreme exhaustion. Such a person might judge, sincerely, that he ought to get out of bed and make an appearance at his son’s birthday party, but find that he is completely unmotivated to do so. So although weak internalism is no doubt much more plausible than strong internalism, it is still contentious whether the thesis is weak enough to allow for certain sorts of motivational failures we think are important and familiar.

Some advocates of weak internalism respond by claiming that amoralists, psychopaths, and people suffering from psychological maladies are making normative judgments that they are motivated to comply with, and insist that, appearances notwithstanding, they are still to some extent motivated. In other words, the move is to claim that such individuals really do make sincere normative judgments that satisfy the weak internalist constraint, it is just that the motivation they feel to comply with them is very faint or weak and thus easily overridden by other strong competing desires. On this strategy, what ultimately explains the motivational failure is other weighty conflicting desires rather than the absence of a necessity relation between sincere normative judgment and pro tanto motivation.

A second, more popular strategy is to maintain that amoralists, psychopaths, and people suffering from psychological maladies are not really making normative judgments. Someone who appears to make normative judgments while remaining wholly unmotivated to comply with them either lacks competence with normative concepts or else is speaking insincerely. If she is insincere, then she uses normative language only in what Hare called the “inverted commas” sense.¹⁸ When she uses normative language, she is not making a normative judgment herself, but rather is “alluding to the value-judgments of other people,” claiming that, “so-and-so falls within

a class of actions which is generally held (but not by [her]) to be obligatory.”¹¹⁹ In other words, this strategy is simply to deny that amoralists, psychopaths, and people suffering from psychological maladies make normative judgments. Although such individuals appear to make normative judgments, this response requires denying that we can take such appearances at face value. The absence of motivation is evidence, on this view, of a disqualifying defect in the judgment.

While it makes sense that someone already convinced of weak internalism would retreat to one of these responses, the rest of us are likely to find these move pretty unsatisfying. The problem posed by amoralists, psychopaths, and people suffering from psychological maladies is supposed to be that these are people who make genuine normative judgments while remaining entirely unmotivated to comply with them. The first response attempts to save the view that normative judgments and motivation are connected as a matter of necessity from this problem by insisting that, appearances notwithstanding, these people are in fact motivated, albeit very, very weakly. The second response instead saves the view by claiming that such people do not really make normative judgments. But if accounting for motivational failure requires defenders of weak internalism to posit motivational states almost entirely drained of their characteristic motivational effects, or to claim that apparently genuine normative judgments are not really genuine normative judgments, it seems reasonable to reject such suggestions as *ad hoc*. Or, in any event, it requires the noncognitivist to give up the abductive argument.

¹¹⁹ Hare (1952: 124–126, 163–165). Gibbard says similar things when he defends internalism against objections concerning the possibility of amoralism and psychopathy (2003: 153).
To see why, recall how the abductive argument is supposed to work. The first premise is internalism, the phenomena to be explained. The second premise is the claim that noncognitivism best explains the first premise. And the third is the conclusion that normative judgments are noncognitive states. But once the non-cognitivist retreats to a version of the first premise that the cognitivist rejects, it now seems like this argument rests on shaky ground. We need agreement on an *explanandum* in order for one side to claim that they explain it better than the other. Otherwise, we have reached a stalemate. At one point, it seemed as though the two sides agreed on what needed to be explained, but offered competing explanations. But now it seems like the disagreement has spread from *explanans* to *explanandum*. For the argument now seems to be:

1. Weak internalism, but a version according to which amoralists, psychopaths, and people suffering from psychological maladies that *appear* unmotivated by their moral judgments are either a tiny bit motivated or do not really make normative judgments
2. Noncognitivism best explains premise (1)
3. Normative judgments are noncognitive attitudes

And here’s the rub: this version of the abductive argument no longer supports noncognitivism since the explanation on offer commits noncognitivists to controversial claims about what the phenomenon is that calls out for explanation. If the first premise is a highly controversial and arguably *ad hoc* version of internalism, then, defenders of cognitivism can and should deny that this is an apt characterization of the thing that needs explaining. And they can therefore reject this explanation, in just the way that they can reject an explanation that commits us to strong internalism.
Thus far, I have identified a general problem with maintaining that there is a necessary connection between normative judgments and action and/or *pro tanto* motivation in the face of counterexamples in the form of various motivational failures. The general problem is that doing so commits defenders of noncognitivism to a controversial *explanandum*, and one which cognitivists are happy to reject. What was once an argument that noncognitivism best explains an agreed upon feature of normative judgments—namely, their distinctive and reliable connection to motivation—has now become an argument that noncognitivism best explains a highly controversial characterization of this feature. This defeats the abductive argument from weak internalism, but it also clearly defeats its deductive cousin. Again, cognitivists can and should reject its first premise if this premise is cast as a version of internalism with such controversial commitments.

6. **Typical Internalism**

In order for the abductive argument to go through, then, we need to find some version of internalism that (a) both cognitivists and noncognitivists can accept, and (b) noncognitivists can explain better than cognitivists. The problem for noncognitivism is that this proves to be an impossible pair of constraints to satisfy. For example, consider:

**Typical Internalism**: *Typically, if you sincerely judge that you ought to do something, then you are somewhat motivated to do it.*

This plausibly satisfies (a)—there is wide agreement that we at least tend to be motivated by our normative judgments—but does not seem to satisfy (b). The trouble is that cognitivists can explain typical internalism just as well as noncognitivists can. All that is required is giving some explanation of why we tend to be motivated to comply with our normative judgments, and there
are lots of ways to explain these tendencies that do not involve maintaining that normative judgments are or express noncognitive, desire-like attitudes. Consider a proposal from Peter Railton. He states,

The [cognitivist] can point to the widespread tendency of individuals to see their own attitudes and conduct as morally acceptable, one result of which is that moral self-assessment—and willingness to accept others’ evaluations of oneself—is largely restrained within the boundaries of what individuals find themselves with reason or motive to do. The [cognitivist] can further point to the widespread reluctance of individuals, at least among acquaintances and when tempers are cool, to make moral evaluations of others in the absence of any substantial reason or motive for them to act accordingly. It would appear possible to account for these and related facts by postulating a nomological rather than semantic connection between moral evaluation and motivation.

(Railton 1993: 297)

On Railton’s view, normative judgments are beliefs, and they motivate when they are accompanied by relevant desires. Normative judgments are typically accompanied by such desires because of the content of normative judgments. Generally, normative judgments concern things we care about; they are about things that promote wellbeing, cause harm to others, or enable us to engage in mutually beneficial cooperation. On Railton’s view, then, individuals suffering from amoralism, psychopathy, and other psychological maladies make genuine normative judgments; they just happen to lack the desires that typically accompany such judgments. But, as Railton notes,
[M]ost people are not like this. They are not wholly unmoved by the suffering of others, even their enemies, or by the prospect of alienating themselves from others through deception. Nor are they entirely without interest in defending their conduct from a disinterested perspective. Nor do they find compensating benefits to offset fully the likelihood of ostracism or punishment for immoral behavior. These facts about most people most of the time are central to explaining why morality has emerged as a persistent and important phenomenon in human societies. (Railton 1993: 322)

When a normative judgment fails to motivate, then, rather than denying that the judgment is genuinely normative, cognitivists like Railton can explain this in terms of an absent desire—a desire that typically accompanies normative judgments, at least in most people, but doesn’t as a matter of necessity.

Of course, noncognitivism can accept and explain typical internalism, too. But gone is the distinctive explanatory advantage it once had. This is because typical internalism eliminates the necessary connection between normative judgment and motivation, and replaces it with a contingent one. And if the reliable connection between normative judgment and motivation is indeed only contingent, then it is far from clear why we should, in light of it, be moved to accept that normative judgments are or express desires. For what drove the intuition motivating the abductive argument in its first instance was the idea that both normative judgments and noncognitive, desire-like attitudes bear some sort of conceptually necessary connection to motivation. This necessary connection to motivation is, it seems, what is supposed to drive the conclusion that normative judgments are or express these noncognitive, desire-like attitudes. A merely contingent connection is not tight enough to give the noncognitivist the explanatory
advantage she needs. We rejected weak internalism because the connection it posits between normative judgment and motivation is too strong; typical internalism is less controversial, but the connection it posits between normative judgment and motivation is too weak to give noncognitivism a distinctive explanatory advantage. Railton and other cognitivists can plausibly explain a contingent connection by pointing to the content of normative judgments to secure a reliable connection to noncognitive, desire-like states that motivate. What the noncognitivist needs, then, is something in between weak and typical internalism, something that preserves the strength of the connection without sacrificing the explanatory advantage like typical internalism.

7. **Conditional Internalism**

Defenders of internalism and noncognitivism alike have tried to offer such a version of internalism. The move is to defend a more qualified version of weak internalism. On this sort of view, there is a necessary connection between normative judgment and motivation, but the connection is a defeasible one. More carefully,

Conditional Internalism: Necessarily, in conditions C, if you judge that you ought to do something, then you are to some extent motivated to do it.\(^{20}\)

Conditional internalism maintains that the necessary connection between normative judgment and motivation holds only under certain conditions. There are lots of different ways to specify the conditions. Perhaps the most common specification maintains that the connection holds only when the person making the judgment meets some normative standard.

For example, Michael Smith has famously defended a version of normative conditional internalism according to which necessarily, if you judge that you ought to do something, then

---

\(^{20}\) Björnsson, et al. (2015: 7)
you will be to some extent motivated to do it unless you are practically irrational. On Smith’s view, then, the connection between normative judgment and motivation is mediated by rationality. If someone is perfectly rational, then she will be motivated to comply with her normative judgments. Why? Because, according to Smith, when someone who is perfectly rational forms the belief that she ought to phi, this causes her to form an desire to phi, which then motivates her to phi. So, on Smith’s view, normative judgments are beliefs, and they motivate when they are accompanied by the relevant desires. In the perfectly rational person, normative beliefs unfailingly generate the relevant intrinsic desires, and thus the perfectly rational person’s normative judgments are unfailingly accompanied by motivation. But, importantly, on Smith’s view, motivational failures are possible. It is just that failing to form the relevant intrinsic desire to phi while nonetheless judging that you ought to phi constitutes an important form of practical irrationality, not unlike failing to form an intention to do something while nonetheless taking yourself to have sufficient reason to do it.

That Smith provides a cognitivist explanation of normative conditional internalism does not, of course, imply that the noncognitivist cannot provide an alternative explanation of the same phenomenon. But the noncognitivist will have to locate the irrationality in a slightly different place. Whereas Smith claims that irrationality can defeat the connection between normative judgments and motivation by resulting in failure to form a desire to accompany one’s normative judgment, the noncognitivist will have to claim instead that irrationality results in the noncognitive, desire-like attitudes that constitute our normative judgments failing to play their

---

21 Smith (1994: 61)
standard motivational role.\textsuperscript{22} Thus, both cognitivists and noncognitivists can accept normative conditional internalism and provide a satisfying explanation of it. The cognitivist can simply claim that, in rational people, normative judgments are accompanied by desires, whereas the noncognitivist can claim that, in rational people, normative judgments motivate. So far, there is no advantage for the noncognitivist. Instead, we find that cognitivists and noncognitivists who accept normative conditional internalism must appeal to different substantive constraints on rationality, with cognitivists championing the view that rationality requires that normative judgments be accompanied by desires, and noncognitivists arguing that rationality requires that the noncognitive, desire-like attitudes that normative judgments constitute be motivating.

A noncognitivist might insist that her explanation is better, because she identifies the correct constraint on rationality, whereas cognitivists identify the wrong one. But the problem with this response is that noncognitivism is supposed to be a metaethical view about what someone does when she judges that something is required by rationality, and one which is compatible with but neutral between different first-order views of the substantive content of rationality. Thus, while a noncognitivist could perhaps argue that her rational constraint is better than the cognitivists, most noncognitivists would be loath to give up their claim that their metaethical theory is compatible with and neutral between various first-order theories of rationality, and rest the defense of their metaethical view on a first-order claim about the substantive content of rationality. Perhaps for this reason, none have tried.

Other versions of conditional internalism try to cash out the conditions in non-normative terms, or at least in less explicitly normative terms. One such version takes the relevant constraints.

\textsuperscript{22} Toppinen (2016); Dreier (2015)
conditions to be those in which the person who makes the judgment is psychologically normal.\footnote{Gibbard (2003: 155), Svavarsdóttir (1999: 176-177); Zangwill (2003; 2009)} On this sort of view, if someone is psychologically normal, then she will be motivated to comply with her normative judgments. This makes room for the motivational failures that caused trouble for weak internalism. Amoralists, psychopaths, and people suffering from depression might not be motivated to comply with their normative judgments because amoralism, psychopathy, and depression are abnormal psychological conditions that defeat the necessary connection that holds between normative judgment and motivation, leaving the judgment without its standard motivational role. For now, I will set aside doubts one might have about the possibility of cashing out some notion of psychological normality in non-normative terms in a way that does not just amount to a claim of typicality.\footnote{Dreier (1990: 12-14, 21-22)} If it cannot be done, then the thesis ultimately collapses into a version of either normative conditional or typical internalism, and thus is vulnerable to the same objections. But even if it can be done, problems remain.

The trouble is that once again both cognitivists and noncognitivists can accept this version of conditional internalism and provide a satisfying explanation of it. This is because the claim that in the psychologically normal person, normative judgments are necessarily connected to motivation does not commit us to a particular explanation of why this connection holds, and in particular, is silent on whether it holds because the content of normative judgments is such that they tend to be accompanied by desires, or because the states of mind normative judgments express are noncognitive, desire-like states that are assumed to have some sort of conceptual connection to motivation. The cognitivist can simply claim that, in psychologically normal
people, normative judgments are accompanied by desires, whereas the noncognitivist can claim that, in psychologically normal people, normative judgments (that is, noncognitive, desire-like attitudes) motivate.\textsuperscript{25} In order for the abductive argument for noncognitivism to go through, though, recall that we need a version of internalism that (a) both cognitivists and noncognitivists can accept, and (b) noncognitivists can explain better than cognitivists. We abandoned strong and weak internalism because both overdrew the connection between normative judgment and motivation by failing to allow for the possibility of certain familiar motivational failures in order to satisfy (a). Conditional internalism can satisfy (a), but so far, just like typical internalism, it looks like no version of it can satisfy (b).

There are many varieties of conditional internalism, and I cannot go through them all here. What the discussion has revealed, though, is that in explaining any particular version of conditional internalism, the cognitivist and the noncognitivist will appeal to different styles of explanation. The cognitivist will claim that normative judgments motivate when they are accompanied by desires, and fail to motivate when those desires are absent. The noncognitivist will claim that normative judgments motivate when noncognitive attitudes motivate, and fail to motivate when such attitudes fail to motivate. And this suggests that, ultimately, the cognitivist and the noncognitivist are committed to two different versions of conditional internalism:

1. Necessarily, in conditions where you have an accompanying desire to do something, if you judge that you ought to do something, then you are to some extent motivated to do it.

2. Necessarily, in conditions where noncognitive states motivate, if you judge that you ought to do something, then you are to some extent motivated to do it.

The disagreement between cognitivist and noncognitivists can only be resolved by determining which is a better description of the phenomenon of moral motivation. The task for noncognitivists hoping to gain support from the abductive argument is to provide an account of the conditions in which desires fail to motivate. Presumably, the noncognitivist will want to show that the conditions in which desires fail to motivate are the very same as the conditions in which normative judgments fail to motivate. If they are, that seems to lend support to (2), the noncognitivist-style interpretation of conditional internalism. On the other hand, the task for cognitivists hoping to squash the noncognitivists’ appeal to the abductive argument is to show that the non-cognitivist cannot make good on this claim. Furthermore, if the cognitivist can identify cases in which normative judgments fail to motivate, but we would expect a noncognitive attitude with the same content to motivate, this would provide an argument for cognitivism. Whether the phenomenon of moral motivation captured by the various forms of internalism that have been discussed in this paper provides support for either noncognitivism or cognitivism therefore turns on which is the better description of the phenomenon. Do normative judgments motivate in conditions where they are accompanied by desires, or in conditions where desires (or desire-like attitudes) motivate? Until this question is answered, noncognitivists cannot claim that internalism provides any support for their view; but nor can cognitivists claim that the failure of internalism provides the reverse.

8. Conclusion

We look to our metaethical theories to explain the fundamental features of normativity. One such feature metaethicists have latched onto is the tight connection between normative judgment and motivation. An alleged advantage of noncognitivism is the ease with which it explains this
tight connection: it claims that normative judgments are (at least partly constituted by) desires, or desire-like attitudes, and that this is all we need to say to explain why such judgments motivate. The problem is that noncognitivists have not provided a plausible characterization of this phenomenon that allows them to retain the distinctive explanatory advantage that they tout. They either posit a strong and controversial connection between judgment and motivation that cognitivists have no trouble rejecting, or they posit a weaker connection that cognitivists can explain just as well. With their favored versions of internalism in hand, noncognitivists may tell each other nice, internally coherent stories about how normative judgments motivate. But they should not expect their stories to convince anyone else. Cognitivists have equally good stories to share amongst each other, and the story one prefers will almost certainly depend on what metaethical view one already accepts.
Works Cited


COMPLETE WORKS CITED


