Hybrid Expressivism: Virtues and Vices*

Mark Schroeder

I. INTRODUCTION: HYBRID METAETHICAL THEORIES

A. A Tale of Two Flowcharts

If you open any textbook on metaethics, one of the first things that you are likely to see is a flowchart.[1] The advertised purpose of this flowchart is to ascertain, by means of your answers to three or four binary questions, where you lie in the space of possible metaethical theories. And its first question usually goes as follows: "Do you think that moral sentences express beliefs or that they express desire-like attitudes?" If you say "beliefs," then you count as a cognitivist, and you will be expected to answer such questions as what those beliefs are about, whether and how we find out about such a thing, how we manage to refer to it in thought and language, and why finding out about it should bear any special connection to motivating us. If you say "desire-like attitudes," then you fall into the expressivist camp, or so it is said, and you are duly forewarned that famous problems accounting for the logical and inferential features of moral sentences await you. The flowchart tells us that the differences between these two camps are deep and fundamental and that we can think of other differences between views as downstream from them.

This first question of such flowcharts, however, contains multiple

* Special thanks to audiences at Cornell University and at the State University of New York at Buffalo in February 2008 and to the research group on moral psychology and moral realism at the Institute for Advanced Studies at the Hebrew University in Jerusalem, organized by David Enoch. Thanks especially to Daniel Boisvert, Steve Finlay, David Copp, Jake Ross, Sigrun Svavarsdottir, Andrew Alwood, Zach Abrahams, Mark van Roojen, Mike McGlone, Nick Sturgeon, Jamie Dreier, Mike Ridge, David Enoch, Ruth Weintraub, Hagit Benbaji, Yonatan Shemmer, and David Braun; to the cast of characters at PEA Soup; and to two anonymous referees and the editors of *Ethics*.

[1] For example, see Alexander Miller, *An Introduction to Contemporary Metaethics* (Cambridge: Polity, 2003), 8.

Ethics 119 (January 2009): 257–309

presuppositions.[2] The growing family of *hybrid* metaethical theories calls our attention to the presupposition that we must choose. When asked whether moral sentences express beliefs or desires, the hybrid theorist asks, "Why not both?" And it is a natural thought. After all, if cognitivists avoid expressivists' problems with logic and inference because they associate moral sentences with ordinary descriptive contents, and if expressivists can offer elegant explanations of the motivating power of moral judgments and the pull of the Open Question argument because they hold that moral judgments are not just beliefs, then maybe a view according to which moral sentences express both kinds of states of mind could claim both of these kinds of advantages—or, at least, something in the neighborhood.

David Copp was an influential early proponent of this recent strand of thought, and Michael Ridge has been one of its high-profile advocates, claiming that his hybrid theory can solve problems facing ordinary expressivism "on the cheap" and that his view offers "the best of both worlds."[3] Stephen Barker sketches what amounts to a hybrid theory, and Daniel Boisvert gives one of its most detailed expositions.[4] Frank Jackson and David Alm offer hybrid suggestions, and Richard Joyce and Matthew Chrisman and Dorit Bar-On offer views in the same broad family as well.[5] Moreover, some pure cognitivists like Stephen Finlay have shown that they can take advantage of expressivist resources in surprising ways, and some pure expressivists like Allan Gibbard have gone to special lengths to show how they can appeal to cognitivist ideas.[6]

[2] See Mark Schroeder, "Expression for Expressivists," *Philosophy and Phenomenological Research* 76 (2008): 86–116, for a different presupposition than I discuss in this article.

[3] David Copp, "Realist-Expressivism: A Neglected Option for Moral Realism," *Social Philosophy and Policy* 18 (2001): 1–43; Michael Ridge, "Ecumenical Expressivism: Finessing Frege," *Ethics* 116 (2006): 302–36, "Ecumenical Expressivism: The Best of Both Worlds?" *Oxford Studies in Metaethics* 2 (2007): 51–76, "Epistemology for Ecumenical Expressivists," *Proceedings of the Aristotelian Society* 81 (2007): S83–S108, and "The Truth in Ecumenical Expressivism," in *Reasons for Action*, ed. David Sobel and Stephen Wall (Cambridge: Cambridge University Press, forthcoming), chap. 11, http://www.michaelridge.com/mr/publish/pdf/truth%20in%20ecumentical%20expressivism.pdf.

[4] Stephen Barker, "Is Value Content a Component of Conventional Implicature?" *Analysis* 60 (2002): 268–79; Daniel Boisvert, "Expressive-Assertivism," *Pacific Philosophical Quarterly* 89 (2008): 169–203.

[5] Frank Jackson, "Non-Cognitivism, Validity, and Conditionals," in *Singer and His Critics*, ed. Dale Jamieson (Oxford: Blackwell, 1999), 18–37; David Alm, "Moral Conditionals, Non-Cognitivism, and Meaning," *Southern Journal of Philosophy* 38:355–77; Richard Joyce, *The Evolution of Morality* (Cambridge, MA: MIT Press, 2006); Dorit Bar-On and Matthew Chrisman, "Ethical Neo-Expressivism," in *Oxford Studies in Metaethics*, ed. Russ Shafer-Landau (Oxford: Oxford University Press, forthcoming), vol. 4.

[6] See esp. Stephen Finlay, "The Conversational Practicality of Value Judgment," *Journal of Ethics* 8 (2004): 205–23, and "Value and Implicature," *Philosophers' Imprint* 5 (2005): 1–20, http://www.philosophersimprint.org/005004/; and Allan Gibbard, *Thinking How to*

All of this amounts to some kind of clear trend, but no two of these theorists' views are alike, and their differences turn out to be both important and educative. The time has come to begin to sort out the differences among these kinds of views, to begin to categorize the advantages and costs facing each, and to set the agenda which must be pursued by those who seek to develop each kind of view. That is where this article comes in. By offering a way of beginning to classify some of these views, by means of sorting out some of the major questions on which they differ, I will show that—as well as why and how—the resources and challenges of these views differ widely from one another. In one of this article's central arguments, I will offer a general argument that hybrid theories, which have a significant advantage over ordinary, pure expressivist views in explaining moral inferences, are thereby debarred from having a significant advantage over ordinary cognitivist views when it comes to explaining the widespread fact of moral motivation. In other words, hybrid theorists can't simply have it both ways, getting the advantages of cognitivism in explaining logic and inference and the advantages of expressivism in explaining moral motivation. Along the way, I'll also be posing specific challenges to hybrid theories with certain sorts of shapes, distinguishing among which advantages hybrid theories can obtain and which they cannot, and trying to set an agenda for what needs to be done in order to make good on what turns out to be the most promising sort of hybrid theory.

The task of this article is therefore expository, organizational, and clarificatory as much as it is argumentative, but the moral will also be cautiously skeptical. I salute hybrid theories for liberating our conception of the space of possible metaethical views—for freeing us from the flowchart, if you will. But their advantages have been in many cases quite exaggerated and not always, I think, as well understood as one would like. My hope is that making explicit the kinds of choice points that we face in developing a hybrid theory can help us to better understand which benefits and costs we garner along the way and why, and hence that I'll be forgiven for introducing what amounts to yet another flowchart. In doing so I don't mean to oversimplify the range of options

Live (Cambridge, MA: Harvard University Press, 2003). Some readers have wondered why I don't include in this list Terry Horgan and Mark Timmons, "Cognitivist Expressivism," in their *Metaethics after Moore* (Oxford: Oxford University Press, 2006), 255–98. The question arises because Horgan and Timmons describe their view as both expressivist and cognitivist. By my lights, however, Horgan and Timmons's view is a perfect example of ordinary, pure expressivism and is not a hybrid theory at all. The only distinctive feature of their expressivist theory which allows them to call it cognitivist is that they offer an object-language semantics for 'believes that' which allows moral complements—a move anticipated much earlier by Simon Blackburn. See Mark Schroeder, *Being For: Evaluating the Semantic Program of Expressivism* (Oxford: Oxford University Press, 2008), esp. chap. 10, for further discussion.

but rather simply to focus attention on some broad kinds of differences which affect our commitments.

B. Four Important Questions

Ordinary, pure expressivists like Simon Blackburn and Allan Gibbard hold that the meaning of a sentence consists in the mental state it expresses and that moral sentences express only one state—a desire-like one.[7] So since 'stealing is wrong' and 'murder is wrong' mean different things, they must express different desire-like states of mind. And since different people mean the same thing by 'stealing is wrong', each must express the same desire-like state of mind. For example, an ordinary, pure expressivist might say that 'stealing is wrong' expresses disapproval of stealing, no matter who the speaker, and that 'murder is wrong' expresses disapproval of murder—again, no matter who the speaker. These are different states of mind because one can be in the former without being in the latter, and conversely.

Nothing about the hybrid move, however, requires the hybridist to agree with either of these views of the pure expressivist. Some hybrid theorists do share with pure expressivists the view that 'stealing is wrong' and 'murder is wrong' express different attitudes—one toward stealing and one toward murder. But others—in fact, as we'll see as we go along, the majority—think that 'stealing is wrong' and 'murder is wrong' express the very same desire-like state. Not just both states of disapproval but the very same state, in the sense that if you are in the desire-like state expressed by the former, then you are ipso facto in the desire-like state expressed by the latter. (Such theorists do not hold that this is an attitude either toward stealing or toward murder.)

Similarly, some hybrid theorists share with pure expressivists the view that when different speakers say 'stealing is wrong' with the same meaning as one another, they express the same desire-like state. But others deny this and claim that different speakers can and do express different desire-like attitudes from one another, even by use of the very same sentence. So, for example, Michael Ridge's view differs from pure expressivists on both of these counts.[8] He holds that every sentence containing the word 'wrong' expresses approval of some 'ideal observer' or another and that which one it expresses approval of depends on which ideal observer the speaker of the sentence actually approves of.

So these are two dimensions in which hybrid views can vary; many

[7] See, e.g., Simon Blackburn, *Spreading the Word* (Oxford: Oxford University Press, 1984), *Essays in Quasi-Realism* (Oxford: Oxford University Press, 1993), and *Ruling Passions* (Oxford: Oxford University Press, 1998); and Allan Gibbard, *Wise Choices, Apt Feelings* (Cambridge, MA: Harvard University Press, 1990), and *Thinking How to Live.*

[8] Ridge, "Finessing Frege."

such dimensions will turn out to be important, but for the primary purposes of this article I am going to draw attention to just two more. My third dimension of variation is this: ordinary, indexical-free sentences have the same descriptive content for different speakers and hence are associated for each speaker with the same belief—belief in that uniform descriptive content. In contrast, ordinary sentences containing indexicals or other context-dependent elements may vary in their descriptive content from speaker to speaker and hence may be associated with different beliefs for different speakers—beliefs in those disparate descriptive contents. I will be interested in the difference between hybrid theories according to which the belief expressed by a sentence varies from speaker to speaker (in the sort of way that happens with sentences containing indexicals), or whether there is a single belief associated with each moral sentence which is the same for each speaker, bracketing the effects of other indexicals (as in 'I've done something wrong').

Finally, the fourth dimension of variation in which I'll be interested is whether a hybrid view takes the descriptive content of a sentence to be derivative from the desire-like attitude expressed.[9] I capture these four dimensions of variation by the following questions, which are going to provide the map for our journey through the hybrid terrain. Wherever the questions seem less than fully precise, they are to be understood in company with the preceding elucidation.

Q1: Do different sentences containing the word 'wrong' express different desire-like states?
Q2: Do different speakers express different desire-like states with the same sentence?
Q3: Does a given sentence have a different descriptive content for different speakers?
Q4: Does the descriptive content of a sentence depend on the desire-like state it expresses?

Figure 1 depicts the space of answers to these questions and classifies a group of theorists who have stated detailed enough views to place them in the figure. Other theorists, such as Richard Joyce, have defended hybrid views but not in sufficient detail to locate them in the figure.[10]

[9] This will become clearer later, in Sec. VII.

[10] Joyce, *Evolution of Morality*. It should be noted that although I've classified Copp by the specific hybrid view that he has spelled out, he is an advocate of the hybrid approach more generally and does not wish his advocacy of it to hang on the success of his other views. Also, as noted above, I'm not as inclined to classify Gibbard and Finlay as hybrid theorists. Gibbard's view differs from that of Jackson's in "Non-Cognitivism, Validity, and Conditionals," e.g., in not appealing to the properties picked out by moral terms in order to underwrite his answer to the Frege-Geach Problem; on the contrary, he appeals to his

Q3	Q4	Q1 Y		Q1 N	
		Q2 Y	Q2 N	Q2 Y	Q2 N
Y	Y		Jackson	Ridge	
	N			Barker Finlay Copp	
N	Y		Gibbard		
	N				Boisvert

FIG. 1

C. My Argumentative Strategy

My strategy in this article is simple and essentially follows the structure of figure 1. In Sections II and III, I will take up the question of how hybrid theorists are to answer question Q1. Only views which answer no to question Q1, I'll be arguing, have better prospects for explaining moral inferences than ordinary, pure expressivists have. Then in Sections IV–VI, I'll turn to questions Q2 and Q3. I'll argue in Section IV that in order to explain moral motivation, hybrid views must give these two questions the same answer; in Section V that answering yes to Q2 makes pressing the question of what the hybrid theorist means by 'express'; and in Section VI that answering yes to Q3 leads to a deep

answer to the Frege-Geach Problem in order to argue that moral terms pick out properties. See Laura Schroeter and François Schroeter, "Is Gibbard a Realist?" *Journal of Ethics and Social Philosophy* 1 (2005), http://www.jesp.org/PDF/Gibbardvol1no2.pdf, for discussion of this aspect of Gibbard's view. Finlay's view has some features similar to Barker's, but it appeals only to explanatory resources to which any cognitivist theory of a certain kind can appeal and is more flexible in some other ways. See Finlay, "Value and Implicature," for comparative discussion. From here forward, I won't be paying special attention to either of these views. Finally, Alm (in "Moral Conditionals") has suggested an interesting view that he calls the "Descriptive Meaning View," according to which the descriptive content of sentences depends on the desire-like attitudes that they express, but the desire-like attitude expressed by a complex sentence in turn depends on the descriptive contents of its parts. This means that his answer to question Q4 is somewhat complicated, so I leave him off of the table. Unfortunately, I won't be able to do justice to his view here.

problem with attitude ascriptions. So, I'll argue (with one important qualification) that workable hybrid theories should say no to both.

Then in Section VII, we'll turn to question Q4, which is probably the central distinctive feature of the account due to Ridge.[11] I take it that this is worth investigating because Ridge claims his view to contrast with competing views, such as that of Stephen Barker, by being a better candidate to be the true heir to pure expressivist views and because Ridge has been one of the most high-profile exponents of hybridism. I'll show that the difference between Ridge's view and Barker's is in fact rather subtle and will argue that the main effect of his positive answer to question Q4 is actually to rob Ridge's account of the very simple assumptions that are needed in order to get the trivial solution to the original expressivist problem with logic that motivates hybrid theories in the first place.

The net result is an argument that if you want to defend a hybrid view with the prospect of doing better than ordinary, pure expressivism at explaining the inferential relations among moral sentences, the most promising prospects lie in the lower right-hand corner of figure 1. You should answer no to each question, which is the view advanced by Daniel Boisvert, who has emphasized, in favor of his view, that pejorative terms seem to have precisely this kind of dual content—for example, the offensive racial slur 'n——r' clearly has a descriptive content, but it is also associated with a contemptuous attitude.[12] In accordance with a no answer to question Q1, the contemptuous attitude it is associated with is the same, no matter what sentence it figures in and even when embedded under negations, inside conditionals, or in questions. In accordance with a no answer to question Q2, it is the same attitude for different speakers—always contempt for a certain racial group. In accordance with a no answer to question Q3, the descriptive content is the same for every speaker. And in accordance with a no answer to question Q4, the contemptuous attitude associated with this word derives from the descriptive content of the word, together with the attitudes prevalent in the social background in which the word was originally used. So the descriptive content does not depend on the associated attitude.

So in Sections VIII and IX, I will look at precisely which sorts of advantages may possibly accrue to hybrid views in the lower right-hand corner of the figure. I will argue that there are many sorts of advantages which cannot accrue to this sort of view but also that this view does allow for certain subtle but potentially attractive improvements over ordinary cognitivist realism in the way that it handles both Open Ques-

[11] Ridge, "Finessing Frege" and "Best of Both Worlds?"

[12] Boisvert, "Expressive-Assertivism."

tion phenomena and in making good on the claim that judgment motivational internalism is both necessary and a conceptual truth.[13] The trick is that both of these relatively subtle but potentially attractive advantages are contingent on a certain hypothesis about how the expressive meaning of moral words interacts with intentional attitude verbs.

The unfortunate thing is that the analogy with pejoratives should not inspire confidence in this hypothesis because the hypothesis arguably fails for pejoratives and clearly fails for a wide range of other sorts of expressive words, including epithets and honorifics. I'll explain this in Section IX and will hence argue that, to go forward, the most promising sort of hybrid theory is going to need to proceed by way of a better set of models than those provided by pejoratives—so hybrid theorists even of this variety still have their work cut out for them.

II. DO HYBRID THEORIES HELP EXPLAIN VALIDITY OR INFERENCE?

A. Pure Expressivism and Frege-Geach

Traditional, *pure* expressivism promises a particular way of accounting for the semantics of moral sentences. Rather than accounting for their semantics by saying what they are about or what their truth conditions are, the expressivist program is to account for their semantics by saying what kind of thoughts they express. I think of it this way: for pure expressivists, the right kind of semantics for moral sentences doesn't assign them propositions as values.[14] It assigns them mental states instead. So for a pure expressivist, 'murder is wrong' and 'stealing is wrong' must express different states of mind because that is what their meaning something different consists in. The pure expressivist's answer to Q1 is yes.

It is a well-known obstacle to traditional expressivism that it has a difficulty accounting for the semantics of complex sentences with moral parts.[15] The problem is very general. Every natural language construc-

[13] For the record, I do not think that judgment motivational internalism is a conceptual truth—but the idea that it is often plays a role in the motivation of expressivist and hybrid theories. The question here is whether such a view can even succeed at obtaining its putative advantages; whether those advantages are more than putative is a topic for another occasion.

[14] See my "Expression for Expressivists" and chap. 2 of *Being For.*

[15] See, e.g., Peter Geach, "Ascriptivism," *Philosophical Review* 69 (1960): 221–25, and "Assertion," *Philosophical Review* 74 (1965): 449–65; John Searle, "Meaning and Speech Acts," *Philosophical Review* 71 (1962): 423–32; Bob Hale, "Can There Be a Logic of Attitudes?" in *Reality, Representation, and Projection,* ed. John Haldane and Crispin Wright (New York: Oxford University Press, 1993); James Dreier, "Expressivist Embeddings and Minimalist Truth," *Philosophical Studies* 83 (1996): 29–51; Mark van Roojen, "Expressivism and Irrationality," *Philosophical Review* 105 (1996): 311–35; and Nicholas Unwin, "*Quasi*-Realism, Negation, and the Frege-Geach Problem," *Philosophical Quarterly* 49 (1999): 337–52.

tion works equally well with moral arguments as with descriptive arguments and yields complex sentences with the same sorts of semantic properties: negation, conjunction, disjunction, conditionals, and quantifiers; alethic, epistemic, and deontic modals; subjunctives, tense, generics, and habituals; binary quantifiers, complement-taking and infinitive-taking verbs, qualifiers like 'yesterday', and more. Expressivists can't take advantage of standard kinds of accounts of the semantics of these kinds of constructions because standard accounts don't treat atomic sentences as having mental states for their semantic values.

So expressivists owe us an alternative semantic account of each and every one of these constructions, which predicts their expected semantic properties. To date, as I've argued elsewhere, pure expressivists have not yet managed even to supply a semantics for 'not' that explains why atomic moral sentences are inconsistent with their negations.[16] So the problem is about as far from being discharged as problems come.

The most famous version of this challenge is to supply a semantics for conditionals—by assigning them a mental state which they express—that explains why they validate *modus ponens*. Now, valid arguments have at least three important features. First, if their premises are true, then their conclusion is true as well. Some have understood this condition to be the biggest obstacle for expressivists; this is apparently why some have held that minimalism about truth would be sufficient to make the problem go away.[17] But since all existing expressivist views have troubles that arise even before we get to this condition, I'll set it aside. A second important feature of valid arguments is that it is inconsistent to accept each of their premises and deny their conclusion. Call this the *inconsistency property*. Some traditional expressivist accounts of conditionals are designed specifically to explain the inconsistency property of *modus ponens*.[18]

A third important feature of valid arguments is that accepting their premises commits someone, in some sense, to accepting their conclusion. This doesn't mean that something is necessarily going wrong if

[16] Mark Schroeder, "How Expressivists Can and Should Solve Their Problem with Negation," *Noûs* 42 (2008): 573–99, and *Being For*, chap. 3. See also Nicholas Unwin, "*Quasi*-Realism" and "Norms and Negation: A Problem for Gibbard's Logic," *Philosophical Quarterly* 51 (2001): 60–75; and James Dreier, "Negation for Expressivists: A Collection of Problems with a Suggestion for Their Solution," in *Oxford Studies in Metaethics*, ed. Russ Shafer-Landau (Oxford: Oxford University Press, 2006), 1:217–33.

[17] See, e.g., Daniel Stoljar, "Emotivism and Truth Conditions," *Philosophical Studies* 70 (1993): 81–101; Huw Price, "Semantic Deflationism and the Frege Point," in *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives*, ed. S. L. Tsohatzidis (London: Routledge, 1994); and Paul Horwich, "Gibbard's Theory of Norms," *Philosophy & Public Affairs* 22 (1993): 67–78; but also Dreier, "Expressivist Embeddings," for discussion.

[18] For example, Gibbard, *Wise Choices, Apt Feelings* and *Thinking How to Live*.

someone accepts the premises and doesn't accept the conclusion—she may simply not have put the premises together, or though she has put them together, she may also have strong evidence against the conclusion and be as yet unsure how to proceed. But it does mean that something is going wrong if someone accepts the premises and has considered the argument but simply declines to accept its conclusion—even when explicitly confronted with the argument from her existing beliefs and even in the absence of any countervailing evidence. Valid arguments must not be like that; their whole interest, to us, is that we can create rational pressure on people to either accept the conclusion of the argument or else give up on one of its premises—not just to not deny the conclusion unless they give up one of the premises. I'll call this stronger feature of valid arguments the *inference-licensing property*. Some traditional expressivist accounts of conditionals are designed specifically to explain the inference-licensing property of *modus ponens*.[19]

B. Inconsistency and the Hybrid 'In'

The best way to see the appeal of hybrid views, and of the idea that by incorporating some descriptive element they can gain a significant step on pure expressivist theories, is to focus on the inconsistency property of arguments like the following (I'll use *modus ponens* just to illustrate):[20]

P1: 'P' expresses[21] → BF(P)
P2: 'P ⊃ Q' expresses → BF(P ⊃ Q)
C1: 'Q' expresses → BF(Q)
~C1: ' ~ Q' expresses → BF(~ Q)

(Here I adopt the convention that small caps are used to denote mental states, so, e.g., 'BF(P)' denotes the belief that P.) I'll assume that accepting a sentence is a matter of being in the mental state that it expresses and that denying a sentence is accepting its negation. So someone who accepts the premises and denies the conclusion of the argument from P1 and P2 to C1 believes that P, believes that P ⊃ Q, and believes that ~Q. It is inconsistent to have these beliefs (because their contents are inconsistent). So such a descriptive argument has the inconsistency property.

The basic attraction of hybrid theories stems from the fact that nothing changes if we put the person who accepts the premises and

[19] For example, Blackburn, *Spreading the Word*.

[20] I include the negation of the conclusion in the diagram in order to illustrate the inconsistency property.

[21] Officially, I don't think that cognitivist views are committed to holding that moral sentences express beliefs in the same sense as pure expressivists think that they express desire-like attitudes. See my "Expression for Expressivists." I'll bracket that worry and the complications it results in, for the purposes of this article.

denies the conclusion in some other mental states as well. So, for example, if moral sentences express both beliefs and desire-like attitudes, then a moral *modus ponens* argument might look something like the following:

P3: ‘N’ expresses $\rightarrow$ BF(P); ATT(1)
P4: ‘N $\supset$ O’ expresses $\rightarrow$ BF(P $\supset$ Q); ATT(2)
C2: ‘O’ expresses $\rightarrow$ BF(Q); ATT(3)
~C2: ‘ ~ O’ expresses $\rightarrow$ BF(~ Q); ATT(4)

(Here I designate the desire-like attitude expressed by each sentence using the generic ‘ATT(*n*)’, for nothing about the hybrid account of the inconsistency property of this *modus ponens* argument turns on what theory we adopt about what attitudes these are. And I make the minimal assumption about the beliefs expressed by each sentence, that they have contents which are isomorphic to the logical structure of the sentence, without yet taking any view about the relationship between them—about this hybrid theorists will take several different views.)

Any hybrid theory that satisfies this simple constraint can explain the inconsistency property of moral *modus ponens* arguments. Accepting the premises and denying the conclusion involves, inter alia, believing that P, believing that P $\supset$ Q, and believing that ~Q. And these are inconsistent beliefs (because their contents are inconsistent). Of course, on the hybrid view, someone who accepts the premises and denies the conclusion also, as it happens, has several desire-like attitudes: ATT(1), ATT(2), and ATT(4). But this is simply irrelevant for the explanation of the inconsistency property, which turns only on the beliefs involved. So capturing the inconsistency property is trivial for such views and follows from the inconsistency property of the ordinary descriptive argument with the same descriptive contents. The additional attitudes associated with sentences neither help nor hinder—it doesn’t even matter what those desire-like attitudes are.

Apparently emboldened by such reasoning, Michael Ridge has recently suggested that hybrid theorists should define validity to be the inconsistency property: an argument is by definition valid, he says, just in case anyone who accepts the premises and denies the conclusion at one and the same time has inconsistent beliefs.[22] This is the key, he claims, to his solution to the Frege-Geach Problem “on the cheap.” It is the hybrid ‘in’.

C. Excluded Middle and Inference Licensing

If an argument has the inconsistency property, then it is inconsistent for someone who accepts the premises to deny the conclusion. So if

[22] Ridge, “Finessing Frege.”

there is some kind of rational pressure to either accept or deny the conclusion, or at least to do one if the other is not an option, then someone who accepts the premises will be under rational pressure to accept the conclusion. So for any argument with a conclusion which obeys excluded middle, having the inconsistency property is sufficient for having the inference-licensing property. But unfortunately, if accepting a moral sentence requires more than just having a belief, then we should not expect moral sentences to obey excluded middle.

The reason for this is simple. Take the sentences 'O' and '~O', from above. 'O' expresses BF(Q) and ATT(3), and '~O' expresses BF(~Q) and ATT(4). So accepting 'O' requires believing that Q and having the desire-like attitude ATT(3), and accepting '~O' requires believing that ~Q and having the desire-like attitude ATT(4). So even though 'Q' obeys excluded middle, it may be perfectly coherent to decline to accept either 'O' or '~O', and not because you aren't sure which way to go. This is because you may believe that Q but not have ATT(3), or you may believe that ~Q but not have ATT(4). Since moral sentences need not obey excluded middle, showing that valid moral arguments have the inconsistency property is insufficient to establish that they have the inference-licensing property. But valid arguments do license inference. So it follows that Ridge validity is not a sufficient condition for validity.

The point I mean to be making in this section is a simple one. I am not saying that hybrid views have a special problem when it comes to accounting for the inference-licensing property. I am just noting that their account of the inconsistency property (in which the descriptive contents alone did the work) is insufficient in order to do this. Because on the hybrid view coming to accept the conclusion of an argument with a moral conclusion involves coming to accept both a belief and also a desire-like attitude, we need to know something about which desire-like attitudes are expressed by moral sentences, in order to establish the inference-licensing property. It is to that task that I turn in Section III. I will be arguing that a hybrid view allows for a more promising explanation of the inference-licensing property just in case it answers no to question Q1.

III. INFERENCE-LICENSING AND THE ANSWER TO Q1

A. Where Does the Attitude Come From?

Since the argument from P3 and P4 to C2 is valid, someone who accepts its premises should be committed to its conclusion. In order to be neutral on exactly how to understand the inference-licensing property, I've been careful about exactly how to understand this sense of 'commitment', but we can say at least this much. There must be something wrong going on with someone who accepts both premises and merely declines to accept the conclusion, even when confronted with the argument and in the absence of reasons to deny the conclusion.

It is easy, of course, to see where the commitment to having the belief expressed by the conclusion comes from. Someone who accepts both premises, after all, must have the beliefs expressed by each, and these beliefs commit to the belief expressed by the conclusion in the very same way as they do for any ordinary descriptive argument. So the question that we need to answer is, why can't someone accept the premises and have the belief expressed by the conclusion, but not go on to have the desire-like attitude expressed by the conclusion, and hence simply decline to accept the conclusion? What could be wrong with that?

Just as a reminder, our picture looks like this (here I've bolded ATT(3) because our question is why someone can't simply have the other five mental states, but not this one):

P3: 'N' expresses → BF(P); ATT(1)
P4: 'N ⊃ O' expresses → BF(P ⊃ Q); ATT(2)
C2: 'O' expresses → BF(Q); **ATT(3)**

As I see it, there are four real distinct possibilities. Either ATT(3) is distinct from all five of the other mental states (and their mereological compositions) or it is not. If it is not, in fact, distinct from the other attitudes, then the answer to the puzzle is simple: someone who is in the other five states ipso facto is in the state ATT(3). But if it is distinct from the other states, then somehow a commitment to ATT(3) must come from some combination of the other mental states. And that commitment must derive either from only the beliefs, from only the desire-like attitudes, or from some combination of the beliefs with the desire-like attitudes.

So first consider the possibility that the commitment to ATT(3) derives only from some combination of the three beliefs. For all I know, this could be true. But it is a dialectically unpromising strategy for the hybrid theorist to pursue. This is because if there was some easily defensible explanation of why anyone who had those descriptive beliefs would be committed to having a desire-like attitude like ATT(3), then motivational internalism would be easy for ordinary cognitivists to explain, putting expressivism in both its pure and hybrid forms completely aside.[23] It is only because this looks hard to explain or defend that the

[23] This is Michael Smith's tactic in *The Moral Problem* (Oxford: Blackwell, 1994). Note that motivational internalism comes in different strengths and that a strategy like Smith's only explains a weaker version, which says that an agent who judges that doing A is wrong will be motivated not to do A so long as she is practically rational, and can't explain a version of internalism which omits the clause about practical rationality. But versions of internalism which omit this clause are quite difficult to defend.

problem gets going in the first place. So my argument here is conditional: if this kind of strategy seemed like a promising one, that would significantly dampen much (though perhaps not all) of the motivation for hybridism in the first place.

Next consider whether the commitment to ATT(3) might derive from the desire-like attitudes alone. There may be a way of making this work as well.[24] But dialectically speaking, it is not a particularly promising strategy either. After all, this is precisely what it would take for an ordinary, pure expressivist view to account for the inference-licensing property. So if a commitment to ATT(3) could be explained on the basis of the desire-like attitudes expressed by the premises alone, then hybridism wouldn't offer special advantages over ordinary expressivist views when it comes to explaining the inference-licensing property after all.[25] Notice, again, that I am not claiming that no theory can be given that would explain how ATT(1) and ATT(2) commit to ATT(3). In fact, I've explored exactly that question in depth on another occasion. All I am claiming is that if we could explain that, then we would again not need the hybrid view. We could do what we need to do, at least as far as moral conditionals go, with the resources of classical expressivism.

The third possibility is that the commitment to ATT(3) derives from some combination of the desire-like attitudes with the beliefs. But if anything, this strategy looks even more complicated than the last. On the face of it, explaining rational relationships purely among desires or purely among beliefs should be easier than explaining rational relationships among sets of both beliefs and desires. So, again, it is very hard to see how this strategy could lead to a hybrid solution to Frege-Geach "on the cheap," as Ridge advertises. More to the point, it is hard to see how it yields a clear improvement over ordinary, pure expressivism, "on the cheap" or not.[26]

[24] See, e.g., my *Being For.*

[25] Notice, e.g., that Jackson's hybrid view appears on the left-hand side of fig. 1; it requires a different sort of account of the inference-licensing property, for which there is no space to go into here. Other hybrid theories which answer yes to question Q1 may also be possible.

[26] There is a model for understanding how desire-like attitudes can, together with beliefs, commit to further desire-like attitudes. It is the instrumental model. So suppose that our argument is from 'stealing is wrong' and 'if stealing is wrong, then getting your little brother to steal is wrong' to 'getting your little brother to steal is wrong'. If 'getting your little brother to steal is wrong' expresses the belief that getting your little brother to steal is K and expresses the desire not to get your little brother to steal, then this desire could turn out to be an instrumental desire, derivative from the desire to not do what is K. So if the conditional premise expresses the desire to not do what is K, then we can get commitment to the desire expressed by the conclusion, even though the conclusion expresses a different desire. As we'll see, however, this picture differs very little from the standard answer that I push for in the main text, and it faces essentially all of the same

That dispenses with all of the options on which ATT(3) is actually distinct from each of the other mental states. Moreover, if it is identical to one of the other states, we can also narrow down which it must be. It would be unpromising to think that it is identical with one of the beliefs, because if there really are beliefs such that having that belief is identical to having some desire-like attitude, then those states would be what are known as *besires*, and, again, old-fashioned cognitivism would have the day with the motivational problem, without any help from expressivism, pure or hybrid. So I conclude that in order to get additional leverage on explaining the inference-licensing property unavailable to pure expressivist views, hybrid views must hold that ATT(3) must really be identical with either ATT(1) or ATT(2).

If so, then it is impossible to accept the premises of the argument without having the desire-like attitude expressed by its conclusion. But accepting the conclusion is merely being in both states that it expresses. So someone who accepts the premises and does not accept the conclusion must not have the belief that it expresses. But then what is going wrong with them is whatever goes wrong in ordinary descriptive cases when someone accepts the premises of a valid argument and merely declines to draw the conclusion. So on the hypothesis that ATT(3) is identical to either ATT(1) or ATT(2), we do get a simple explanation of the inference-licensing property. This answer doesn't drop out of hybridism alone, however, or even out of hybridism plus the hypothesis that the descriptive contents are isomorphic to the structure of the sentences. To get this explanation, we need a particular hypothesis about which desire-like attitudes are expressed.

B. Which One?

If the desire-like attitude expressed by a *modus ponens* argument is always also expressed by one of its premises, then there ought to be something in general that we can say about which of the premises must express the same attitude as the conclusion. So let's suppose that for a *modus ponens* argument, ATT(3) is always identical with ATT(1)—the attitude expressed by the minor (nonconditional) premise. Now consider the following argument:

P5: 'N ⊃ O'
P6: '(N ⊃ O) ⊃ O'
C3: 'O'

This argument is another instance of *modus ponens*, simply substituting 'N ⊃ O' for 'N'. So if the attitude expressed by the conclusion of a

problems in the remainder of the article; I omit separate discussion of it for reasons of space.

modus ponens argument is always also expressed by the minor premise, then that must be true in this case as well. So the attitude expressed by ‘O’ must also be expressed by ‘$N \supset O$’. But, of course, ‘$N \supset O$’ was the major (conditional) premise of our original argument. So if the minor premise always expresses the same attitude as the conclusion does, then the major premise must too. I conclude that if there is any general guarantee that the attitude expressed by the conclusion is always expressed by at least one of the premises, it must guarantee that the conditional premise always expresses the same attitude as the conclusion.

This should be no surprise; we can get to the same result by different reasoning. Take, for example, the case of a descriptive-normative *modus ponens* argument, rather than one that has a normative minor premise. For example: ‘the tax reduction would increase revenues’; ‘if the tax reduction would increase revenues, then the tax reduction ought to be passed’; ‘the tax reduction ought to be passed’. If the minor premise like this one does not express a desire-like attitude at all, it can’t express the one that is expressed by its conclusion, and so if someone who accepts the premises of this argument ipso facto has the desire-like attitude required to accept its conclusion, it must be in virtue of accepting the conditional premise. Hence, if there is a general explanation of why someone who accepts the premises of a *modus ponens* argument is committed to its conclusion, it must be that the conditional premise expresses the same desire-like attitude as the conclusion.

C. The Answer to Q1

Notice, moreover, that there was nothing special about *modus ponens* that allowed us to establish this result. If arguments of the form of *modus tollens* are valid, they must also license inference, and if they have moral conclusions which express desire-like attitudes, then there must be a question of where the commitment to those desire-like attitudes comes from. We will have all of the same choices as before, with all of the same merits. And just as the conditional premises of *modus ponens* arguments can serve as the minor premises in alternative *modus ponens* arguments for the same conclusion, the same goes for the conditional premises of *modus tollens* arguments.[27] For example, for the argument from ‘Q’ and ‘$\sim P \supset \sim Q$’ to ‘P’ there is the alternative argument from ‘$\sim P \supset \sim Q$’ and ‘$\sim P \supset \sim (\sim P \supset \sim Q)$’ to ‘$P$’. So the same goes: in general, the major premise of a *modus tollens* argument must express the same desire-like attitude as its conclusion.

[27] In the following argument I ignore cancellation of double negations. But it is easy to see for similar reasons that ‘$\sim\sim P$’ must express all and only the same desire-like attitudes as ‘P’, given that the inference from each to the other is valid and so must be inference licensing.

This observation leads to a very general one. Compare the following two arguments constructed with arbitrary choice of moral sentences P and Q:

$$\begin{array}{ll} \sim P & \sim Q \\ \sim P \supset Q & \sim P \supset Q \\ \hline Q & P \end{array}$$

The argument on the left is of the form of *modus ponens*, so by the previous reasoning, its major premise, '$\sim P \supset Q$', must express the same desire-like attitude as its conclusion, 'Q'. Similarly, the argument on the right is of the form of *modus tollens*,[28] so by the previous reasoning, its major premise, '$\sim P \supset Q$', must express the same desire-like attitude as its conclusion, 'P'.

It follows that as long as we assume that sentences express at most one desire-like attitude, then since '$\sim P \supset Q$' expresses the same attitude as 'P' and the same attitude as 'Q', it follows that 'P' and 'Q' must express the same attitude as each other. And this for arbitrary choice of 'P' and 'Q'! So if any sentences that can figure as conclusions of nontrivial valid arguments express desire-like attitudes at all, they must all express the very same one! We can avoid this conclusion only by assuming that complex sentences like '$\sim P \supset Q$' may express more than one desire-like attitude—but always the ones expressed by their parts. But then the hybrid view turns out to be much more surprising than the suggestion that perhaps sentences express pairs of mental states rather than single mental states. On this view, it will turn out that sentences may express arbitrarily many mental states. All that is required is to construct arbitrarily complex sentences with parts that express all of those distinct attitudes, and there is no restriction in English on how complex sentences may be.

As I understand Michael Ridge's view, he endorses the conclusion that arbitrary moral sentences P and Q (holding fixed the speaker and the time) express the very same desire-like attitude: approval of a certain sort of ideal observer (although perhaps a different one for different speakers and for the same speaker at different times).[29] Ridge thinks that each moral predicate has a descriptive content that is analyzed in terms of this ideal observer, in different ways for different predicates.[30]

[28] See n. 27 above.

[29] See Ridge, "Finessing Frege," "Best of Both Worlds?" and "Truth in Ecumenical Expressivism." For discussion of the point in parentheses, see Mark van Roojen, "Expressivism, Supervenience, and Logic," *Ratio*, n.s., 18 (2005): 190–205.

[30] Although compare "Epistemology for Ecumenical Expressivists," in which Ridge seems to suggest the view that sentences containing 'knows' express a different desire-like attitude.

He doesn't say so explicitly, but, to generalize from his remarks, Ridge apparently thinks that 'murder is wrong' and 'skiing is great' express the very same desire-like attitude, relative to the same speaker and the same time.

Most hybrid theorists (e.g., Barker, Boisvert, and Copp) don't go so far as to hold that arbitrary moral sentences P and Q always express the same desire-like attitude. Instead, most such theorists go in for the view that arbitrary sentences containing a given moral predicate all express the same desire-like attitude. In the same way that any use of a pejorative like 'n——r' expresses the same contemptuous attitude no matter where it appears in a sentence—under negation or disjunction, in a conditional, a question, or what have you—these theorists suppose that each moral predicate is associated with a desire-like attitude (perhaps holding fixed the speaker and time), such that every sentence in which that word appears expresses that desire-like attitude. And this guarantees, of course, that a conditional premise will express the same desire-like attitude as the conclusion, since any moral words appearing in the conclusion must also appear in the conditional premise. So unlike Ridge's view, these views associate sentences with potentially more than two mental states. But like Ridge's view, they answer no to question Q1, holding that different sentences containing 'wrong' express the very same desire-like state of mind.

You might think that there is yet a further possibility.[31] For all that I've shown, you might suppose, a hybrid theorist might still answer yes to question Q1, holding that there is a different desire-like attitude expressed by each and every atomic moral sentence. So long as her view also holds that every complex sentence (including conditionals) expresses every mental state expressed by each of their parts, she can still obey the constraint that is brought out by my pair of the *modus ponens* and the *modus tollens* arguments, and she is not thereby forced to conclude that 'stealing is wrong' and 'murder is wrong' express the same desire-like attitude. That is true. But unfortunately, the point of the argument that the attitude expressed by 'Q' must be expressed by 'P ⊃ Q' was based on the assumption that someone who accepts 'P ⊃ Q' must be in each mental state that it expresses. But it is not in fact plausible, if 'murder is wrong' and 'stealing is wrong' express distinct attitudes—one toward stealing and one toward murder—that someone who accepts 'if stealing is wrong, then murder is wrong' must have the attitude associated with 'murder is wrong'. That, after all, was Geach's point in the first place, in raising the Frege-Geach Problem. So it is because for this to help us explain the inference-licensing property,

[31] Thanks to Andrew Alwood for making me be explicit about this important step of the argument. For this point, also see Boisvert, "Expressive-Assertivism."

'P ⊃ Q' must express an attitude that would be had by anyone who accepted it, that hybrid theorists cannot simply say that this is an attitude that you have when you accept 'murder is wrong' but not when you merely accept other sentences, such as 'stealing is wrong'.

My discussion in Sections II and III has clearly not been intended to argue hybrid theorists into any particular corner; the point has been rather to explain the trade-offs that we face at this choice point in developing a hybrid theory and why hybrid theorists are generally (though not always) led to such a different view about the attitudes expressed by moral sentences than that of ordinary, pure expressivists. We'll see in what follows that this difference turns out to be important. My moral so far is this: if you want your hybrid theory to give you more leverage on explaining the inference-licensing property than ordinary, pure expressivism, then by far the most promising option, and possibly the only real option, is to answer no to question Q1 and hold that different sentences containing 'wrong' all express the very same desire-like state.

IV. EXPLAINING MORAL MOTIVATION

A. Hybrid Theories and Moral Motivation

In Sections IV–VI, I will be arguing that hybrid theories face significant problems unless they say no to both questions Q2 and Q3. In Sections V and VI, I'll pose one specific problem for each, arguing in Section V that a yes answer to Q2 leads to a problem about how moral sentences can express desire-like attitudes at all, and in Section VI that a yes answer to Q3 leads to a serious problem about attitude ascriptions. But first, in Section IV, I'll argue that problems for either yes answer are really also problems for the other because, given that the answer to question Q1 is no, hybrid theorists need to answer questions Q2 and Q3 in the same way in order to be able to explain moral motivation.

First I'll give an intuitive gloss on the problem, and then I'll outline a simple hybrid theory in order to illustrate how the problem works in detail. To see the intuitive problem, first observe that if moral inferences work as we observed in Section III, then someone who makes a moral inference does not come to have a new desire-like attitude but only comes to acquire a new belief. But if beliefs motivate only in connection with desire-like attitudes (an assumption which is controversial but shouldn't be for hybrid theorists), then someone who draws a moral conclusion by inference will be motivated to act on it only if the belief that she forms is somehow connected to some desire-like attitude that she already has.

Now, any ordinary externalist view can explain why someone would be motivated by a desire-like attitude that she happened by coincidence to have or which people simply generally happen to have. So if hybrid theorists are to have any edge over ordinary externalists in explaining moral motivation, then it must be the desire-like attitude expressed by the sentence which motivates her to act. This attitude, after all, is the only one that we can be sure that she has, if she accepts the sentence. But that means that the belief that you come to accept when you accept a moral sentence must be related to the desire that it expresses in some systematic way—such that the desire that it expresses would motivate you to act on that belief.

That is why if moral sentences express different desires in the mouths of different speakers, they must express correspondingly different beliefs as well. If they did not and all expressed the same belief, then the different desires that they expressed couldn't motivate agents to act on that same belief in the same way. Similarly, this is why if moral sentences express different beliefs in the mouths of different agents, they must express different desires as well. If they did not and all expressed the same desire, then that desire wouldn't be able to motivate people to act on any of the variety of different beliefs that they might form when they come to accept that sentence. So once hybrid theorists answer no to question Q1, they need to give the same answer to questions Q2 and Q3.

B. A Detailed Example to Work With

I'll now outline a simple hybrid theory, in order to illustrate how this constraint works in practice; the illustration will come in handy as we look at more detailed hybrid views in Sections V–VII as well. The theory I outline will be for an idealized, simplified language with the expressive power of propositional logic, and its semantics will work by assigning each sentence to a pair of attitudes, a belief and a desire, relative to a context of utterance. I'll ignore the kinds of complications that would be required in order to expand this language very far; my point here is only about how motivation will get explained, although since the discussion so far has been very abstract, the example will also enable us to get a more concrete look at what hybrid theories are typically like.

First assume that all sentences containing the word 'wrong' express a desire not to do what is K, at least relative to a given context of utterance (different ways of implementing this theory will choose different values of 'K' or will make it vary from context to context, depending on the speaker). Then say that atomic sentences containing 'wrong', of the form 'A is wrong', express the belief that A is K. And then say that complex sentences express the belief in the proposition that is formed by compositionally composing the contents of the beliefs expressed by its parts, in accordance with the structure of the sentence. That is, if 'N' is a sentence expressing the belief that P and 'O' is a sentence expressing the belief that Q, then '~N' expresses the belief that ~P, 'N&O' expresses the belief that P&Q, 'N ∨ O' expresses the belief that P ∨ Q, and 'N ⊃ O' expresses the belief that P ⊃ Q.

Now consider the following argument, where 'M' and 'S' are abbreviations for 'murder' and 'stealing' and small caps again denote mental states, this time with the attitude made explicit as a particular desire:

P7: 'S is wrong' → BF(S is K); DES(to not do what is K)
P8: 'If S is wrong then M is wrong' → BF(S is K ⊃ M is K); DES(to not do what is K)
C4: 'M is wrong' → BF(M is K); DES(to not do what is K)

The mental states listed after the arrows are the ones expressed by each sentence, according to the rudimentary view just sketched. Now suppose that Al begins by accepting the premises of this argument but not accepting its conclusion. So he believes that stealing is K and that if stealing is K, then murder is K, and he has a desire to not do what is K but does not yet believe that murder is K. Because he does not yet believe that murder is K, his desire to not do what is K does not yet motivate him not to murder.

But when Al goes through this argument and draws its conclusion, he forms the belief that murder is K. Because we answered no to question Q1, his desires do not themselves change when he comes to accept the conclusion of the argument for the first time, but his motivations do change because, now that he believes that murder is K, this engages with his desire to not do things that are K and motivates him to not murder. This kind of simple hybrid view, because it respects the constraint that the desire expressed by a sentence is related in a systematic way to the belief that it expresses, is able to explain why new moral conclusions can motivate us.

The hybrid views offered by Ridge, Barker, Copp, and Boisvert all work in precisely this way.[32] Ridge's and Barker's views differ from Boisvert's in assigning a different value of 'K' relative to contexts of utterance with different speakers, and Copp is neutral on this question. But each view that assigns a different value of 'K' in the beliefs expressed also assigns a different value of 'K' in the desire expressed.[33] And it is now

[32] Ridge, "Finessing Frege"; Barker, "Value Content"; Copp, "Realist-Expressivism"; and Boisvert, "Expressive-Assertivism."

[33] Finlay's view, as explained in "Conversational Practicality of Value Judgment'" and in "Value and Implicature," is somewhat more complicated; he holds that in certain kinds of contexts speakers pragmatically convey that they have a certain desire-like attitude and that which one the speaker conveys that she has in these sorts of contexts varies along with the descriptive content of the sentence. So Finlay doesn't believe that speakers "express" (pragmatically convey that they have) desire-like attitudes in every context of ut-

easy to see why this is so. If the belief expressed were allowed to vary across different speakers while the desire expressed was held fixed, then there would no longer be any guarantee that someone who came to accept the conclusion would thereby be motivated, because there would no longer be any systematic connection between the belief she came to have in accepting the conclusion and the desire that she already had, by virtue of accepting the premises—and similarly if the desire were allowed to vary across different speakers while the belief was held fixed.

I've now argued that hybrid theorists who answer no to question Q1 must give the same answer to both question Q2 and question Q3. That means that a problem for a yes answer to either is a problem for a yes answer to the other. In Sections V and VI, I'll raise one serious problem with a yes answer to each. The former is really a problem about how different such a hybrid view is from ordinary, pure expressivism, in terms of whether it can still take advantage of the expressivist's resources and framework. The latter is really a problem about how similar such a hybrid view is to ordinary contextualist views—so similar that it still inherits some of their principal difficulties.

V. A PROBLEM IF THE DESIRE-LIKE ATTITUDE VARIES ACROSS SPEAKERS (YES TO Q2)

A. The Expression Relation

According to a yes answer to question Q2, different speakers express different desire-like attitudes by the same sentence. So, for example, for Jeremy the value of 'K' might be 'fails to maximize happiness', so that when Jeremy uses a sentence containing the word 'wrong', he expresses the desire to not do what fails to maximize happiness, while for Immanuel the value of 'K' might be 'does not follow from a universalizable maxim', so that when Immanuel uses a sentence containing the word 'wrong', he expresses the desire to not do what does not follow from a universalizable maxim. If we put the yes answer to Q2 along with our previous no answer to Q1, therefore, it leads to the view that the desire-like attitude expressed by moral sentences varies across speakers but not across sentences.

This is striking because, as we noted earlier, it contrasts with what ordinary, pure expressivists like Blackburn and Gibbard think about the *expression* relation. According to Blackburn and Gibbard, the desire-like attitude expressed by moral sentences varies across sentences but is constant across speakers. This is what Blackburn and Gibbard have to say because they hold that moral sentences express only desire-like at-

terance, but when they do, he holds that they vary systematically with the descriptive content of the sentence, thus obeying the constraint argued for in this section and the last.

titudes and that the meaning of a sentence is determined by the attitude that it expresses. Since 'murder is wrong' and 'stealing is wrong' mean different things, it follows that they must express different states, and so since they express only desire-like states, it follows that the desire-like states that they express must differ. Similarly, since Jeremy's use of 'murder is wrong' and Immanuel's use of 'murder is wrong' mean the same thing, they have to express the same desire-like state.

In holding this view of expression, Blackburn and Gibbard ascribe to it the same properties as the expression of ordinary, descriptive sentences. Ordinary, indexical-free descriptive sentences also express the same belief for different speakers but different beliefs for different sentences. For example, 'grass is green' expresses a different belief from 'snow is white'—the former expresses the belief that grass is green, and the latter expresses the belief that snow is white—and each sentence expresses the same belief, no matter who its speaker is. Given that pure expressivists use the word 'express' in such a way that the attitude expressed varies across sentences but not speakers and that even hybrid theorists must agree that for indexical-free descriptive sentences the attitude varies across sentences but not speakers, it is quite striking for the hybrid theory to tell us that for moral sentences the desire-like attitude expressed varies across speakers but not sentences.

In fact, this is a striking enough fact to merit a serious question as to what hybrid theorists even mean by 'express'—not only that the expression relation has such different properties from what ordinary expressivists held it to have but also that it has such different properties between moral and descriptive sentences. Could it really be that they are talking about the same thing? What conception of the expression relation could hybrid theorists have in mind? There is in fact a serious challenge for hybrid theorists to say what the expression relation could possibly be that it behaves so differently between moral and descriptive language in this way. The problem is one about how different hybrid views like this are from pure expressivist views. They are different enough that it is no longer clear that they can take advantage of the resources of ordinary expressivist views. They are going to have to say something else instead, and just what, it turns out, may be a tricky question in and of itself. (In the following sections I use Ridge's hybrid account to illustrate this problem, but I could just as well have illustrated the very same problem with Barker's account.)

B. Gibbardian Expression, to Illustrate the Basic Problem

In *Wise Choices, Apt Feelings*, Gibbard offered an account of the expression relation appealed to by his theory.[34] An utterance expresses a mental

[34] Gibbard, *Wise Choices, Apt Feelings*, 84–86.

state, for Gibbard, when a speaker utters that sentence in order to convey to her audience that she is in that mental state by means of linguistic conventions. According to Gibbard, 'grass is green' is conventionally associated with conveying the information that the speaker believes that grass is green (and hence, derivatively, with the information that grass is green, since this can be inferred together with the assumptions that the speaker is sincere and that she is reliable).

Unfortunately, Gibbard's view cannot be what Ridge or Barker mean when they talk about 'expression'—I'll use Ridge's view just as an example in order to illustrate why not. Recall that on Ridge's view, which mental state a speaker expresses when she utters a moral sentence is not a function of which sentence she has uttered; it is rather a function of which 'ideal observer' the speaker in fact approves of. Therefore, no moral sentence is conventionally associated with conveying the information that the speaker approves of any particular ideal observer, and so no speaker can take advantage of such a convention in order to convey that she approves of some particular ideal observer to her audience by uttering the sentence. So it follows that she can't express, in Gibbard's sense, her approval of that particular ideal observer, as Ridge's view requires.

You may think that this reasoning moves too fast. Perhaps there is no particular ideal observer, such that convention associates any moral sentence with conveying the information that the speaker approves of it (him; her). That much follows from the fact that different speakers use the same sentence to express approval of different ideal observers. But it doesn't yet follow that there are no conventions that a speaker can take advantage of in order to convey to her audience that she approves of a particular ideal observer. Perhaps there is instead a rule which determines which information is conventionally conveyed, as a function of the speaker of the sentence.

Fortunately, Kaplan's 'dthat' operator allows us to conveniently state, for Ridge's case, what this function would be.[35] Each moral sentence would be conventionally associated with conveying the information that is the descriptive content of the following sentence:

'I approve of dthat(the ideal observer whom I approve of)'.

Unfortunately, this is not progress. No sentence could be designed to convey this information from speaker to audience because in order to ascertain which information was being conveyed, the audience would

[35] Recall that 'dthat' is a directly referential-term-forming operator, so that the semantic contribution of 'dthat(the F)', if *x* is the F, is just *x*. The 'dthat' operator gives us a convenient way to describe the rules by which context-dependent expressions acquire their referents.

have to already know it. The problem is not that a rule like this could not succeed at conveying some information or other. The problem is that it will not succeed at conveying the information that the speaker approves of dthat(the ideal observer whom the speaker approves of). That information will be either already known or not ascertainable.

And that means that no sentence could Gibbard-express what Ridge needs it to, even by means of a context-dependent convention like this one. (Of course, Ridge would have independent problems with appealing to Gibbard's account of expression because he also claims that neither speaker nor audience need know which ideal observer the speaker approves of or has expressed approval of, whereas Gibbard's account clearly requires that both the speaker and the audience know these things.[36] But I take that to be an incidental feature of Ridge's view.) The central problem with appealing to Gibbard's account of expression is that the rule by which Ridge wants to associate sentences with the attitudes expressed makes any such information that they could convey vacuous.

Of course, Ridge doesn't in fact appeal to Gibbardian expression in his view—I've merely used it to illustrate the basic problem because it is the most familiar account of expression in the literature. In Section V.C, I'll show that the same problems arise if we appeal instead to Wayne Davis's account of expression (which Ridge seems to favor), to the idea of expression as conventional implicature (as Stephen Barker favors), or to assertability-conditional expression, as outlined in my "Expression for Expressivists."[37]

C. Davisian, Conventional Implicature, and Assertability-Conditional Expression

Wayne Davis's account of expression leaves us in precisely the same place as Gibbard's, and for precisely the same reasons. Here is Davis's definition:

> Definition: *S expresses* Ψ *iff S performs an observable act as an indication of occurent* Ψ *without thereby covertly simulating an unintentional indication of* Ψ.[38]

Davis clarifies that this use of 'indication' is in the evidential sense, in which a turn signal indicates, by being evidence that, someone is about

[36] See Ridge, "Finessing Frege," 313–15: "The speaker may not have a very clear idea" and following, particularly the reference to "je ne sais quois" at 315.

[37] Wayne Davis, *Meaning, Expression, and Thought* (Cambridge: Cambridge University Press, 2003), as endorsed in Ridge, "Finessing Frege" and "Best of Both Worlds?"; and Barker, "Value Content."

[38] Davis, *Meaning, Expression, and Thought*, 59. The italics are from the original text; Davis uses italics for all of his definitions.

to turn.[39] And he is explicit that by doing A as E, he means that 'S did A in order to provide E, and moreover intended his doing A to be E'.[40] So it follows from this definition that a speaker only expresses a desire-like attitude by uttering a moral sentence if she intends that utterance to provide evidence that she has that desire-like attitude.

Davis's account of expression leaves us with the same problems that Gibbard's did. In order for a given speaker's utterance to be an 'indication' of her occurrent approval of some particular ideal observer, there must be some general rule which tells us, given that speaker's circumstances, which ideal observer she is likely to approve of, given that she has uttered that sentence. But on Ridge's view, the circumstances under which you count as expressing approval of a given particular ideal observer are that that is the ideal observer of whom you actually approve. So the general rule only tells us that, given that the speaker approves of such and such an ideal observer, her utterance indicates that she approves of that particular observer. But that's no kind of indication at all. Just as with Gibbard's account, Ridge's view simply debars the speaker from using the sentence to convey the information that she is in the mental state that she needs to be able to express. So she can't express it on the Davisian account.[41]

Stephen Barker's suggestion that moral sentences conventionally implicate that the speaker is in the mental state expressed is unavailable to Ridge for exactly the same reason. A conventional implicature is some information that is both conventional and conveyed. But we saw in the last section that so long as the speaker is to count as expressing approval of whatever ideal observer she actually approves of, no conventional rule could result in information that would actually be conveyed to the audience. To know which information was to be conveyed, the audience would have to already know which ideal observer the speaker approved of, in which case there would be nothing left to convey.

Finally, it is worth noting that although all three of the accounts of the expression relation that I have surveyed so far have traded on the idea that in expressing a mental state, you are somehow conveying the information that you are in it, very similar problems arise for accounts of the expression relation on which it is not a matter of conveying anything. For example, on the proposal I made in "Expression for Expressivists," the mental state expressed by a sentence is the mental state that it is the semantic assertability condition of the sentence that the

[39] Ibid., 25.

[40] Ibid., 49.

[41] It is worth noting, incidentally, that Davis's account also clearly requires the speaker to know which mental state she is expressing, which Ridge explicitly denies.

speaker be in. That is, for a sentence, S, to express a mental state, M, S must be semantically associated with a rule of the form:

> Assert S only if you are in M!

To take advantage of this account, Ridge would need to claim that every moral sentence is semantically associated with the following rule:

> Assert me only if you approve of dthat(the ideal observer you approve of)!

Just as the informationally based accounts of expression foundered on the problem that Ridge's account made the required information vacuous, the assertability-conditional account founders on the problem that Ridge's account makes the associated rule vacuous. Since this rule cannot be broken, it is hard to see what semantic role there could be for it to play.[42]

D. Summing Up, and a Contrast

Clearly there is unfinished work for hybrid theorists in saying just what they mean when they say that a sentence expresses a desire-like attitude. The main lesson to be drawn is not so much the problem for the particular theories of Barker and Ridge as the more general lesson that the more that such theorists differ from ordinary, pure expressivists in terms of their commitments, the clearer it is that they cannot simply take advantage of what works for ordinary, pure expressivists. In this section I've focused particularly on the view of Michael Ridge, but I could just as well have conducted the discussion with Stephen Barker as my example because, like Ridge, Barker seems to think that the desire-like attitude expressed by a given utterance of a moral sentence not only varies from speaker to speaker but is fixed by which desire-like attitude that speaker actually has. This was the feature of Ridge's view that raised the biggest problems.

That also means that the 'problem' raised in this part of the article has not been completely general, which brings us to the "important qualification" that I anticipated in Section I.C. A view according to which the desire-like attitude expressed by a sentence varies from speaker to speaker but is not a strict function of the desire-like attitude that speaker actually has could gain leverage in finding an appropriate account of expression. Stephen Finlay's end-relational theory (which is really a resourceful version of ordinary cognitivism rather than a

[42] Technically, these rules can be broken, but you are guaranteed to only be bound by one you are not breaking—this is analogous to the fact that indexical validities like 'I am here' are guaranteed to express truths, even though whatever truth they express is contingent.

specifically hybrid view, even though he does talk about moral sentences 'expressing' desire-like attitudes) appears to have just these kinds of resources.

According to Finlay, every moral sentence requires an *end-relational* semantics. Things are good only relative to some end, wrong only relative to some end, and so on. In some contexts we can talk about what is good or bad relative to an end that we understand is shared by neither the speaker nor the audience. But in other contexts—particularly contexts in which the end is not made explicit—it is understood that the speaker and the audience have the ends relative to which the thing is being said to be good (wrong, etc.). In such cases, this ends up being conversationally implicated—that is, pragmatically conveyed in a way that is not built into any particular rule of the language. So on Finlay's view, a speaker 'expresses' a desire-like attitude—that is, conversationally implicates that she has a certain end—only in certain contexts, but not in others. Since the attitude expressed depends on the end relative to which the sentence is independently being understood, rather than being fixed by a rigid rule which pays attention only to what the speaker actually approves of, Finlay's account can allow that information is genuinely conveyed by such conversational implicatures, and so it does not run into the problems of the last few sections.

So my point is not that you can't say yes to question Q2 and hold that the desire-like attitude expressed by a moral sentence depends on the speaker or can vary across contexts. It is that a yes answer to question Q2 makes particularly pressing the question of whether you can still take advantage of the resources of ordinary, pure expressivism, and I have shown that for Barker and Ridge in particular, this leaves some significant obstacles unsurmounted. The ultimate extent of these difficulties turns, of course, on whether there are significant reasons why Barker and Ridge are tempted to conclude that the speaker always expresses the desire-like attitude that she actually has—an important question, but unfortunately one I won't be able to investigate here.

VI. A PROBLEM IF THE BELIEF VARIES ACROSS SPEAKERS (YES TO Q3)

A. The General Problem with Attitude Ascriptions

Whereas a yes answer to question Q2 holds that the desire-like attitude expressed by a moral sentence can vary across speakers or in general between contexts of utterance, a yes answer to question Q3 holds that the belief expressed can vary across speakers or in general between contexts of utterance—even for sentences containing no other indexicals or context-dependent terms. In Section V, I argued that a positive answer to question Q2 leads to a problem with understanding what the expression relation could be, a problem that is particularly acute in the cases of Barker and Ridge. Given the conclusion of Section IV, that militates against a positive answer to question Q3 as well or at least indicates some of the work cut out for it. Now I'll be arguing for a direct problem for a positive answer to question Q3: that it leads to an old problem with attitude ascriptions. Whereas the problem in Section V was for hybrid theories that were too different from ordinary, pure expressivism to take advantage of its resources, the problem in Section VI arises for hybrid theories that are too similar to ordinary contextualist theories to escape their difficulties.

Ordinary contextualist theories are just ordinary cognitivist theories according to which the content of sentences containing moral words like 'wrong' varies between contexts of utterance, even in sentences which contain no other context-dependent terms, such as indexicals. One of the simplest and most familiar contextualist theories is old-fashioned speaker subjectivism, according to which 'murder is wrong' either means or is truth-conditionally equivalent to 'I disapprove of murder'. Jamie Dreier offered one important updated contextualist theory, and the most sophisticated contemporary contextualist theory is developed by Stephen Finlay.[43]

The general problem with attitude ascriptions is simple. We can work, just for illustration, within the framework established in Section IV.B. Suppose, then, that for Jeremy the value of 'K' is 'fails to maximize happiness', whereas for Immanuel the value of 'K' is 'does not follow from a universalizable maxim'. Now consider the following exchange (I won't go so far as to call it a "dialogue," and the problem does not require that it is):

Jeremy: Lying to save a friend's life is not wrong.

Immanuel: Jeremy said that lying to save a friend's life is not wrong.

Immanuel: But lying to save a friend's life is wrong.

Immanuel: So Jeremy said something that is not true.

Given our stipulations about the value of 'K' for each speaker, we can spell out the descriptive components of the argument as follows (I leave '??' in the place that I want to examine in the argument):

Jeremy$_{\text{descr}}$: Lying to save a friend's life does not fail to maximize happiness.

[43] James Dreier, "Internalism and Speaker Relativism," *Ethics* 101 (1990): 6–26; Stephen Finlay, "Conversational Practicality of Value Judgment," "Value and Implicature," and also "Oughts and Ends," *Philosophical Studies* (forthcoming), and "A Confusion of Tongues," unpublished book manuscript.

Immanuel$_{\text{descr}}$: Jeremy said that lying to save a friend's life is not ??.[44]

Immanuel$_{\text{descr}}$: But lying to save a friend's life does not follow from a universalizable maxim.

Immanuel$_{\text{descr}}$: So Jeremy said something that is not true.

The problem about attitude ascriptions can be summed up as the question of whether to replace '??' with 'fails to maximize happiness' or with 'does not follow from a universalizable maxim'.

Suppose, first, that we choose 'fails to maximize happiness'. That seems motivated because it makes sense of why Immanuel would attribute it to him, given the belief that he actually expressed. But then it makes Immanuel's inference look like a terrible one. 'Jeremy said something that is not true' clearly should not follow from 'Jeremy said that lying to save a friend's life does not fail to maximize happiness' and 'lying to save a friend's life does not follow from a universalizable maxim'.

So suppose, on the other hand, that we choose 'does not follow from a universalizable maxim'. That solves the problem about Immanuel's inference. Now it is a good inference to go from 'Jeremy said that lying to save a friend's life does not follow from a universalizable maxim' and 'lying to save a friend's life does not follow from a universalizable maxim' to 'Jeremy said something that is not true'. But it makes it extremely puzzling why Immanuel would attribute such a view to Jeremy in the first place, given that this is not the belief that he actually expressed. It is particularly puzzling, given that Immanuel may fully realize that he and Jeremy share the same views about which actions do not follow from universalizable maxims and which ones fail to maximize happiness, even though they still disagree about what is right and wrong.

This is a problem. There must be *some* descriptive content to Immanuel's report of what Jeremy has said. But neither candidate for what this descriptive content could be makes sense of the full range of what Immanuel does with it—to both infer it from the observation of Jeremy's utterance and to employ it in inference. This is a problem that is faced by ordinary (purely cognitivist) contextualist theories which postulate surprising sorts of context dependence in the word 'wrong'. It is just an example of such a problem, but it is a nice example, because it makes clear that there is a problem, no matter which way we go. The worry

[44] For the purposes of this section I set aside the question of whether attitude ascriptions also include, as part of their descriptive content, that their subject has or has expressed the associated desire-like attitude. We'll return to consider this question and its consequences in Secs. VIII and IX.

here is that hybrid theories which say yes to question Q3 will simply inherit this difficulty, along with the others facing ordinary cognitivist contextualist theories.

Notice, moreover, that unlike the most pressing problem in Section V, this problem does not trade on the assumption that the descriptive content of a moral sentence is fixed by the attitudes that the speaker actually has. The problem still arises for more flexible views (such as that of Stephen Finlay) because it is not a problem about whether the view can assign the right descriptive content to Immanuel's report of what Jeremy has said, but rather that there is no single assignment that can make sense of both of Immanuel's inferences—to what Jeremy has said and, from that, to the conclusion that he has said something untrue.

For example, since Finlay's contextualist view is flexible and treats 'wrong' more like 'local' than like 'I', he can claim that sometimes a sentence like Immanuel's report of Jeremy's view can report the descriptive content associated with Jeremy's use of the word—in which case the former inference is a good one—and sometimes the very same sentence can report the descriptive content associated with Immanuel's use of the word, in which case the latter inference is a good one. But Finlay is forced, without some further maneuver, to conclude that the actual set of inferences that Immanuel goes through trades on an equivocation—each inference works relative to some context of utterance, but there is no context that licenses both.[45] Yet I take it to be a clear datum that in the original dialogue, Immanuel goes through an immaculate set of observations (bracketing his questionable view about the moral question at stake).

B. A Possible Reply

(This and the following section pursue a possible, somewhat complicated reply to the problem posed in Section VI.A, based on an idea which would, if it worked, show how hybrid theories could have extra resources available to solve the problem unavailable to ordinary contextualist theories.[46] Both sections may easily be skipped by readers who would like to stick to the main thread.)

In setting up the problem in the previous section, I assumed that the following argument would be invalid:

P9: Jeremy said that lying to save a friend's life does not fail to

[45] For a further maneuver, see Gunnar Björnsson and Stephen Finlay, "Normative Contextualism Defended," working paper (University of Gothenburg and University of Southern California, 2008).

[46] Thanks to Mike Ridge and to the PEA Soupers for fruitful discussion of the objection in Sec. VI.A and for making the worry in Sec. VI.B concrete. See http://peasoup.typepad.com/peasoup/2007/10/a-problem-for-s.html.

maximize happiness.

P10: Lying to save a friend's life does not follow from a universalizable maxim.

C5: Jeremy said something that is not true.

That assumption seemed warranted—arguments of the form 'Jeremy said that P', 'but Q'; 'therefore Jeremy said something that is not true' are not, in general, truth preserving, unless P implies ~Q. But it could be that lying to save a friend's life does not fail to maximize happiness, even though it does not follow from a universalizable maxim. In fact, anyone who thinks that Kantian ethical theory succeeds in upholding absolutist principles where utilitarianism fails will think that the failure of this implication is quite important. So the argument does seem clearly to be invalid.

But it is important to note that the success of this objection trades on an assumption about what 'true' means. To contrast, compare the following (structurally similar) argument:

P11: Jeremy said that he is tall.

P12: I am not tall.

C6: Jeremy said something that has the property that there is a sentence that could be used in his context of utterance to say it, which is not true relative to my context of utterance.

This argument is truth preserving, even though 'he is tall' and 'I am tall' have different descriptive contents. Its truth-preservingness derives, in addition to the background facts about how 'he' and 'I' vary from context to context, entirely from the predicate in C6: 'has the property that there is a sentence that could be used in his context of utterance to say it, which is not true relative to my context of utterance'. So here is the point: my assumption of the nonvalidity of the argument from P9 and P10 to C5 traded on assuming that 'true' differs importantly in meaning from the predicate which appears in C6.

The reason it is so important to note this potential loophole in the argument is that Michael Ridge has, in fact, offered an account of the semantics for 'true' which is supposed to make the argument from P9 and P10 to C5 truth preserving by a precisely analogous move.[47] According to Ridge's account, the descriptive contents of the sentences can be spelled out something like as follows:

P9*: Jeremy has uttered some sentence which expresses, relative to his context of utterance, the ordinary descriptive belief that lying to save a friend's life does not fail to maximize happiness and the desire not to do actions which fail to max-

[47] Ridge, "Truth in Ecumenical Expressivism."

imize happiness.

P10*: Lying to save a friend's life does not follow from a universalizable maxim.

C5*: Jeremy has uttered some sentence, S, with the property that there is some proposition, *p*, such that for anyone who is expressing the desire-like attitude that I hereby express, S expresses the belief whose content is *p*, and *p* is not true.

Now, that is a mouthful, and it still doesn't make the argument truth preserving. But let's allow Ridge the free background assumption that every sentence which expresses, relative to Jeremy's context of utterance, the descriptive belief that lying to save a friend's life does not fail to maximize happiness and the desire not to do what fails to maximize happiness has the feature that, relative to the context of anyone who is expressing the same desire-like attitude as Immanuel is in uttering C5*, it expresses the descriptive belief that lying to save a friend's life does not follow from a universalizable maxim. That assumption is yet another mouthful, but if we give it to Ridge on the assumption that it is guaranteed by some sort of facts about meaning in the language, then, holding it fixed, C5* does follow from P9* and P10*. I'll let you verify this yourself, since any explanation of mine is likely to make it sound more complicated, rather than less.

C. A Problem and an Observation

Without stopping for a complete evaluation of Ridge's semantics for 'true', let me make two observations here. The first is that as I have stated it, Ridge's account predicts that too many arguments are truth preserving. In particular, it appears to work for the following argument:

P13: Jeremy said that dthat(the desire-like attitude that he is thereby expressing) is the desire not to do what fails to maximize happiness.

P14: dthat(the desire-like attitude that I am hereby expressing) is not the desire not to do what fails to maximize happiness.

C7: Jeremy said something that is not true.

This is because for anyone who is expressing the desire-like attitude that Immanuel expresses, the sentence 'dthat(the desire-like attitude that I am hereby expressing) is the desire not to do what fails to maximize happiness' expresses the belief that the desire not to do what does not follow from a universalizable maxim is the desire not to do what fails to maximize happiness—which premise P14 explicitly denies. But 'dthat(the desire-like attitude that I am hereby expressing) is the desire not to do what fails to maximize happiness' is precisely what Jeremy uttered and which premise P13 reports. So by Ridge's account of truth, it follows that Jeremy said something that is not true.

I take this to be a serious problem, and not least because by such reasoning any speaker can immediately infer that everyone expresses the same desire-like attitudes as they do. But it is clearly part of Ridge's view that different speakers express different desire-like attitudes than he does. So something is definitely going wrong.

Now, it may be that the problem arises because I've incorrectly translated Ridge's actual proposal for the semantics of 'true'. What Ridge's actual proposal says is that 'q is true' expresses the ordinary descriptive belief with the following content:

> There is a proposition s, natural-kind-belief in which would (at least partly) constitute the causal-regulation belief that q for anybody who [is expressing the same desire-like attitude as I hereby do], and s is true.[48]

To get my version of Ridge's view, I've understood "would (at least partly) constitute the causal-regulation belief that q" to mean "would be part of thinking that q."[49] But since, officially, Ridge's "deflationist" notion of belief ranges over complements 'q', which can't be part of the contents of ordinary descriptive belief (the sense of 'belief' in which moral sentences express both beliefs and desires), 'q' shouldn't appear anywhere in the content of the ordinary descriptive belief. In order to get around that problem, I've substituted what I took the descriptive content of 'S thinks that q' to be, on Ridge's account—namely, that S is in the mental state expressed by 'q' relative to Al's context of utterance. Otherwise I've made no amendments to the account.

If my substitution is what created the problem, that must be because Ridge needs the descriptive content of 'S thinks that q' to work differently, depending on whether the context dependence in 'q' arises from moral terms or from ordinary context-dependent terms. Ridge wants 'q' to contribute its descriptive content relative to the subject's context of utterance when 'q' is a simple moral sentence like 'stealing is wrong'. But of course, in order to get attitude ascriptions for ordinary context-dependent sentences right, he needs 'q' to contribute its descriptive content relative to the context of the speaker, when 'q' is just an ordinary context-dependent sentence. And the way that I spelled out the descriptive content of 'S thinks that q' didn't manage to do this, which is why I was able to exploit it to create an argument using ordinary context-dependent sentences that his view also validated.

But it's no easy thing to get around this problem; for example, it

[48] Ibid., 26.

[49] Ibid., 20–24.

doesn't suffice to just state a disjunctive semantics for 'thinks that', depending on whether its complement is moral or not. For example, for 'S thinks that I've done something wrong', getting the results Ridge wants requires getting 'I' to have its descriptive content assigned relative to the context of the speaker but getting 'wrong' to have its descriptive content assigned relative to S's context. Now, if Ridge had an adequate alternative semantics for 'thinks that', it would have solved this problem, so he wouldn't get these unintuitive results. But Ridge hasn't in fact provided such an alternative semantics for 'thinks that', and the difficulty of doing so is precisely the original problem about compositionality and context dependence that faces the second fork of our dilemma. The problem I've been discussing in this section is therefore just an illustration of the complications involved in doing this.

My second observation about Ridge's solution is much simpler. It is that Ridge's account, and any similar account, requires that all sentences containing 'true' have to express every desire-like attitude that might be expressed by any moral predicate whatsoever. This is because it is not enough, to explain the validity of arguments like the following, to explain why the argument among their descriptive contents is truth preserving:

P15: Jeremy thinks that stealing is wrong.
P16: Everything Jeremy thinks is true.
C8: Stealing is wrong.

As we saw in Section II.C, in order to explain why the argument is inference-licensing, we need to explain why anyone who accepts both of the premises is committed not just to the belief expressed by the conclusion but to the desire-like attitude it expresses.

But since P15 is just an ordinary descriptive sentence, it presumably doesn't express any desire-like attitude. That means that if we are to explain why someone who accepts the premises already has the desire-like attitude expressed by the conclusion, as the general hybrid strategy for explaining the validity of moral arguments requires, we have to assume that the desire-like attitude expressed by 'stealing is wrong' is also expressed by 'everything Jeremy thinks is true', even though it is not obviously a moral sentence. That means that for hybrid theorists who think that 'wrong', 'good', 'bad', 'just', 'objectionable', 'rational', and 'correct' all express different desire-like attitudes, 'everything Jeremy thinks is true' must express all of these desire-like attitudes. Things are a little bit easier for Ridge, who holds that sentences containing each of these words all express the very same desire-like attitude. He only needs 'everything Al thinks is true' to express that one desire-like attitude, rather than many.

Is this a problem for an account of truth like this one? It does create some pressure for hybrid theorists to go for Ridge's version, on which all moral predicates are associated with the same desire-like attitude. I don't know that it is a problem in its own right, but I'll come back to consider it in Section IX, when we look in more detail at what desire-like attitudes contribute to the semantics of attitude ascriptions.

VII. A PROBLEM IF THE BELIEF EXPRESSED DEPENDS ON THE DESIRE EXPRESSED (YES TO Q4)

A. The Succession: Who Inherits the Mantle of the Expressivist Dynasty?

In Section IV, I argued that the answers to questions Q2 and Q3 must go hand in hand, and in Sections V and VI, I argued that yes answers to each, like those given by Barker and Ridge, lead to highly problematic commitments, even in the company of Ridge's creative semantics for 'true'. Given such views, it is not at all clear what 'expression' even means or whether there is even any sense of 'expression' at all that would fulfill these theorists' commitments. Moreover, we saw that views like these face a further, difficult problem with attitude ascriptions.

If this is right, then we could almost go without considering the differences between these views, except that Michael Ridge has so firmly suggested that these differences are of such fundamental importance that they entitle his view, and his alone, to its claim to be the true heir or successor of traditional, pure expressivist views. According to Ridge, his view counts as "ecumenical expressivism" while the other views count as "ecumenical cognitivist"—and the main structural difference which divides Ridge's view from Barker's is its positive answer to question Q4.[50] In what follows I'll argue that there is important truth in this claim of Ridge's but that precisely this feature of his view is also responsible for making him unable to account for even the inconsistency property of valid inferences—precisely the feature that motivated hybrid views in the first place. Far from being an advantage to his view, Ridge's yes answer to Q4 is a serious liability.

To see both why Ridge's view in fact has a better claim than Barker's to be the true heir to traditional expressivism and why it is in fact more problematic than Barker's, we will first need to compare these views side by side. To make the parallels visible, I will idealize slightly from

[50] At points Ridge seems to suggest ("Finessing Frege," 307, and "Truth in Ecumenical Expressivism," 8–9) that the difference lies in his semantics for 'true', discussed in the last two sections. But it is hard to see how that could be the case, since Barker's proposal doesn't include a semantics for the object-language word 'true' at all and hence doesn't include one that is inconsistent with Ridge's. So it could be supplemented with the same sort of semantics for 'true' as Ridge adopts, without any extra complications. In any case, even if his answer to Q4 does not turn out to be the main feature that Ridge thinks makes his view special, it does create a special problem for his view.

the details of the view Ridge describes, but the details that I idealize from are inessential to his fundamental picture of the mechanics of how sentences express desire-like attitudes and beliefs.[51]

Both Ridge and Barker hold that a sentence like 'that knife is good' expresses a certain very general sort of approval. On Ridge's favored view, it is approval of a certain sort of ideal observer, but on a simpler version of the view that he uses in order to illustrate how it works, the approval is of a kind of property that knives can have. On Barker's view, it is approval of a kind of property that knives can have. So I'll work with this version of the view. In both cases, which state of approval is expressed depends on what property of knives the speaker in fact approves of. Recall that on Barker's view, the belief expressed by 'that knife is good', relative to any given context of utterance C, is the one whose content is the proposition expressed by the following sentence, relative to C.

> Barker: That knife instantiates dthat(the property I approve of knives for having).

So on Barker's view, the descriptive content of a normative sentence like 'that knife is good' is quite independent of which desire-like attitude is expressed by the sentence, and in fact this could be the descriptive content of the sentence even if it did not express any desire-like attitude at all.

On Ridge's view, in contrast, normative sentences only express beliefs because they express desire-like attitudes. On his view, the descriptive content of sentences is derivative from the desire-like attitude expressed by each sentence and makes what Ridge calls "anaphoric reference" to that desire-like attitude. Now, I believe that it is not possible to make literal sense of Ridge's claim of 'anaphoric back-reference'. Literally, an anaphor is a syntactic constituent which requires a syntactic antecedent. So Ridge's view fails to make sense of how what he is talking about could be literally a case of anaphora on at least two different counts: first, his view does not actually provide a syntactic antecedent for the anaphor, and second, his view does not explain which syntactic constituent of the sentence 'that knife is good' is the anaphor. So if we are to make sense of Ridge's view, we need to read him loosely, as insisting on the view that 'that knife is good' has a descriptive content which makes reference to something that is determined by the matter of what desire-like attitude that sentence actually expresses. It is easy to

[51] The idealization is that Ridge's preferred version of his view appeals to an 'ideal observer', which I'm leaving out here and which Ridge sometimes leaves out himself, e.g., in "Finessing Frege," 315–16. Holding fixed the difference in their answer to question Q4, Barker's answer could adopt the ideal observer idea as well, in any case.

construct such a view; the belief it expresses, relative to a context of utterance C, will be the one whose content is the proposition expressed by the following sentence, relative to C:

> Ridge: That knife instantiates dthat(the property this sentence expresses approval of).

So understood, Barker's and Ridge's views agree about both the belief and the desire-like attitude expressed by any given utterance of 'that knife is good'. If the speaker approves of serratedness in knives, for example, then both views hold that the desire-like attitude expressed is approval of serratedness, and the belief expressed is the belief that the knife is serrated. If the speaker approves of bluntness, in contrast, then both views hold that the desire-like attitude expressed is approval of bluntness, and the belief expressed is the belief that the knife is blunt. The difference between these two views is not over what attitudes are expressed by the sentence, but in the mechanism by which the descriptive content of the belief is determined. On Barker's view, this mechanism is independent of the desire-like attitude expressed by the sentence, while on Ridge's view it is dependent on the desire-like attitude expressed.

I'm prepared to grant that this difference entitles Ridge to his claim to be offering the only hybrid view that would be a true heir to traditional versions of expressivism, both because it makes his view the only one on which the semantic role of normative sentences is primarily expressive rather than descriptive and because it enables him to claim that normative sentences literally couldn't work as they do if they did not express desire-like attitudes. So I believe that we should grant Ridge his claim to the succession.[52] The interesting question is whether a hybrid theorist should want this or whether, once we have cast our lot with the aspirations of the hybrid account, it is better to install a new regime than to inherit the mantle of the old one.

B. Ridge's Problem with Validity

If Ridge's view is better poised to inherit the mantle passed down by traditional expressivism, it also creates problems that Barker's view doesn't have.[53] It follows from Barker's view that anyone who under-

[52] Although, Jake Ross has justly pointed out to me that it isn't crazy to think that we might do better to reserve the claim to the 'succession' for pure expressivists rather than for hybridists at all.

[53] The problem in this section can be thought of as a generalization of the problem raised in van Roojen, "Expressivism, Supervenience, and Logic."

stands the rule by which moral sentences express beliefs will be in a position to appreciate that the following argument is a valid one:[54]

P17: This knife is good.
P18: If this knife is good, then that knife is good.
C9: That knife is good.

After all, on Barker's view, the sentences express beliefs with contents determined by the following rules:

P17B: This knife instantiates dthat(the property I approve of knives for having).
P18B: If this knife instantiates dthat(the property I approve of knives for having), then that knife instantiates dthat(the property I approve of knives for having).
C9B: That knife instantiates dthat(the property I approve of knives for having).

So anyone who understands that the descriptive contents of these sentences are determined in this way will be in a position to see that this descriptive argument is truth preserving and hence in a position to rationally draw the conclusion and to be justly accused of irrationality if she accepts the premises and denies the conclusion.

On Ridge's view, in contrast, the sentences express beliefs with contents determined by the following rules:

P17R: This knife instantiates dthat(the property this sentence expresses approval of).
P18R: If this knife instantiates dthat(the property this sentence expresses approval of), then that knife instantiates dthat(the property this sentence expresses approval of).
C9R: That knife instantiates dthat(the property this sentence expresses approval of).

Since Ridge holds that all of these sentences express approval of the same property for a given speaker at a given time, his view of course predicts that this descriptive argument is truth preserving. But in contrast to Barker's view, someone who understands that the descriptive contents of moral sentences are determined by these rules is not, in fact, in a position to see that the conclusion follows from the premises. It is necessary, further, to realize that each sentence expresses approval of the same property. If you do not realize this, or if you believe that different sentences express different states of approval, then you will

[54] Ignore possible effects of context on the reference of 'this knife' and 'that knife'; I would give the knives names to avoid this complication, but that would make the sentences look awkward since most knives don't have names.

not be in a position to see that accepting the premises commits you to the conclusion, and so it will be perfectly rational for you to accept the premises and deny the conclusion.

An analogy, at this point, may be instructive. The difference between the descriptive arguments that Barker's and Ridge's views take to be doing the work in the case of our moral argument is like the difference between the following two arguments:

P19: Superman is strong.
P20.1: If Superman is strong, then I'm a walrus.
C10: I'm a walrus.

P19: Superman is strong.
P20.2: If Clark Kent is strong, then I'm a walrus.
C10: I'm a walrus.

These two arguments differ only in their second premises, which both say (within the relevant fiction, a qualification I'll henceforth ignore) of Superman that if he is strong, then I'm a walrus. But while Lois Lane cannot rationally accept the premises of the first argument and deny its conclusion, it is perfectly rational for her to accept the premises of the second argument and deny its conclusion. She can do this because she doesn't realize that Clark Kent is Superman. Ridge's descriptive argument differs from Barker's in the same way—grasping the rules by which its sentences express their descriptive contents is not enough to see that the property referred to by each sentence is the same.

This is a serious problem for Ridge. There is something rationally inconsistent about accepting the premises of a moral *modus ponens* argument and denying its conclusion: that is the inconsistency property that hybrid expressivist originally set out to explain. But Ridge's view can only explain why this is irrational for people who happen to have the background knowledge that every sentence containing the word 'wrong' in fact expresses the very same desire-like attitude—that the answer to question Q1 is no.

Of course, Ridge could claim that this is background knowledge that every competent speaker is required to possess—in contrast to the Superman case, in which competent speakers may fail to have this information. But recall that before hybrid theories came along, the metaethical terrain was supposed to be divided between ordinary cognitivists, who don't believe that moral sentences express desire-like attitudes at all, let alone the same one, and traditional, pure expressivists like Blackburn and Gibbard, who quite explicitly answer yes to question Q1. It is a steep cost that Ridge can explain why it would be inconsistent of Gibbard or Nick Sturgeon to accept both 'stealing is wrong' and 'stealing is not wrong' by attributing to them views that they have spent their professional careers denying.

The moral is that inheriting the mantle of the expressivist tradition is no advantage at all for Ridge's view. It is, in fact, a very steep cost. Functionally speaking, the only difference between Ridge's view and Barker's is that Barker's view can actually use the descriptive contents of normative sentences to do the work that hybrid theorists would like them to do, while Ridge's view is not even able to do that, except for the special case of people like himself who happen to believe in his theory. On the plausible assumption that even pure expressivists like Gibbard and ordinary cognitivists like me would in fact be irrational if we accepted the premises of a moral *modus ponens* argument and denied its conclusion, something seems to be going seriously wrong with Ridge's view. Much better to stick with a proposal like Barker's, even if it means losing the mantle of the expressivist tradition. At least that view obtains the advantages advertised for hybrid expressivism in the first place. And still better, as I suggested in Sections IV–VI, to answer no to questions Q2 and Q3 in the first place.

VIII. WHAT'S LEFT IF WE ANSWER NO TO ALL FOUR QUESTIONS?

A. The Lower Right-Hand Corner

So far, I have given reasons why hybrid theorists should answer no to each of questions Q1–Q4. The problem with a yes answer to Q1 is that it makes the inference-licensing property no more explicable than it would be on a traditional expressivist account. So even though such views are possible, they don't offer any clear progress. The problem with answering yes to Q2 but not to Q3, or conversely, is that it robs new moral conclusions of the ability to independently motivate. Of course, perhaps new moral conclusions really don't have the ability to motivate, but it is precisely this idea that is at the heart of the theoretical grounds for the sort of internalism that motivates hybrid theories in the first place.[55]

The problem with answering yes to Q2 is that, together with a no answer to Q1, it makes the hybrid theorist's commitments about the expression relation so different from those of the ordinary, pure expressivist that it is far from clear what the hybrid theorist is to mean by 'express', and some of the existing hybrid theories were shown to run full bore into this obstacle. This is, again, not to say that such views are obviously false, but it makes it extremely unclear what the expressive

[55] For example, compare esp. the argument for internalism in Michael Smith, *Moral Problem* and "The Argument for Internalism: Reply to Miller," *Analysis* 56 (1996): 175–84.

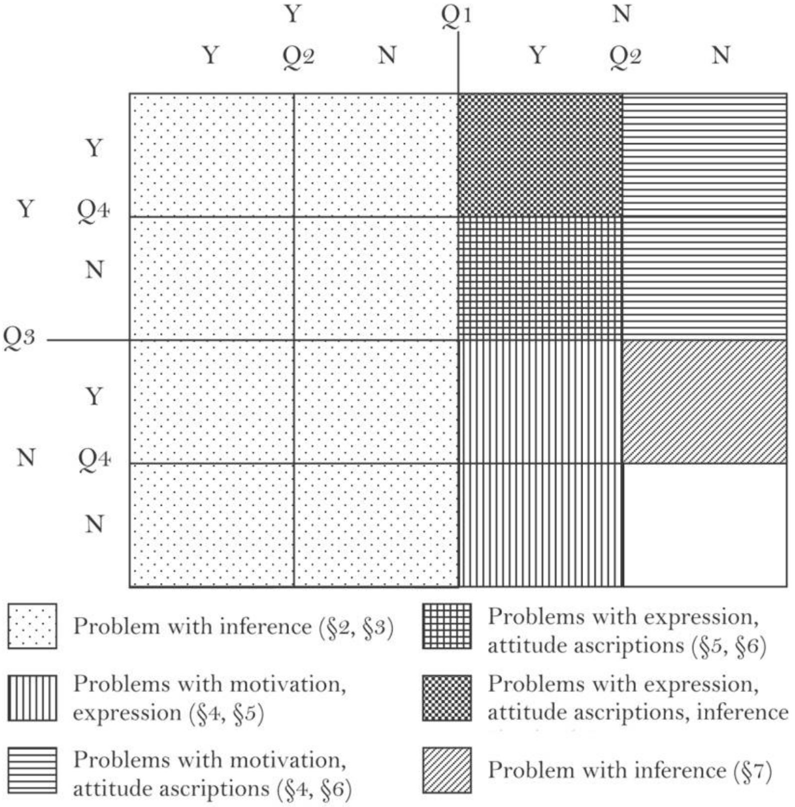
Y
Q1
N
Y
Q2
N
Y
Q2
N
Y
Y
Q4
N
Q3
Y
N
Q4
N
Problem with inference (§2, §3)
Problems with expression, attitude ascriptions (§5, §6)
Problems with motivation, expression (§4, §5)
Problems with expression, attitude ascriptions, inference
Problems with motivation, attitude ascriptions (§4, §6)
Problem with inference (§7)

Fig. 2

component adds that the descriptive component alone can't do. Answering yes to Q3 leads to our difficult problem about attitude ascriptions. And the problem with answering yes to Q4 is that if the descriptive content of a sentence is determined by the desire-like attitude expressed by that very sentence, then only speakers who realize that every sentence expresses the same desire-like attitude will be rationally inconsistent for accepting the premises of a valid argument and denying its conclusion. Figure 2 summarizes the results of these arguments, and I repeat the four questions here, just as a reminder.

Q1: Do different sentences containing the word 'wrong' express different desire-like states?

Q2: Do different speakers express different desire-like states with the same sentence?

Q3: Does a given sentence have a different descriptive content for different speakers?

Q4: Does the descriptive content of a sentence depend on the desire-like state it expresses?

This makes me think that the most promising hybrid theories are going to have to occupy the lower right-hand corner of figure 1. They will have to join Daniel Boisvert in answering no to all four questions. I don't think that this should be a surprising result; there are reasons why the examples in natural languages of terms which are plausible candidates for bearing both descriptive and expressive meaning are terms like pejoratives, which carry a single descriptive meaning that is invariant across speakers. It is because the purposes of language could not be served without these kinds of constraints. The problems we've encountered along the way about inference, expression, and attitude ascriptions are illustrations of how the pressures of the purposes of language play out.

But now it is important to take stock and see what sorts of proposed advantages might remain for hybrid views in Boisvert's family. If we accept a view like Boisvert's, then we are agreeing that there is a single, invariant, descriptive content of a term like 'wrong'—a property that actions can have or lack. But once we concede this much, we must still answer questions like, What property is it? Is it reducible or *sui generis*? How do our words and thoughts manage to be about such a thing? How do we manage to find out what is right or wrong? Why does wrongness supervene on nonnormative properties and relations? These are most of the main challenges facing standard versions of cognitivist realism—in fact, they are some of the problems that ordinary, pure expressivism is motivated precisely in order to avoid having to answer—and hybrid theories in the lower right-hand corner do not help us to answer them.

That's okay—but it is important to understand. Hybrid views which say yes to question Q3, whatever their problems, do have the potential payoff that they could make many of the metaphysical and epistemological questions of metaethics go away, in much the same way that they do on ordinary, pure expressivism. So what advantages do views like Boisvert's have over ordinary cognitivism? In the next two sections I want to explain exactly what advantages hybrid theories like Boisvert's do, at least potentially, have over ordinary cognitivism and to show how those potential advantages turn on an important hypothesis about attitude ascriptions. In Section IX we'll consider the evidence for that hypothesis.

B. Open Questions

Since G. E. Moore first advanced his famous Open Question argument, many philosophers have taken it to establish some important or compelling difference between moral and nonmoral language. On the face of it, the Open Question argument was merely a test of cognitive significance—noting that it is apparently possible, for any ordinary descriptive predicate, ‘K’, to think that stealing is K without thinking that it is wrong. Some philosophers, including myself on other occasions, have not been persuaded that there is anything more exciting about what is revealed by Open Question considerations than goes into instances of Frege’s Puzzle in any other domain. But other philosophers have held that there is “something more” to the Open Question argument.[56]

A natural thought for the hybrid theorist to have is this: I can explain what this “something more” is. According to my account, after all, ‘wrong’ has the same descriptive content as the ordinary descriptive term ‘K’, but someone can easily think that stealing is K without thinking that stealing is wrong—and not merely because they entertain these thoughts under different guises, or modes of presentation, or because they are associated with different Fregean senses. No, on my theory there is an important something more—the practicality of the term ‘wrong’—that doesn’t follow merely from thinking that stealing is K. To think that stealing is wrong, you have to think that stealing is K and desire not to do what is K.

This is an initially attractive thought, but we have to be careful with it. Any hybrid theory which says no to question Q1 is going to hold that there are an awful lot of sentences which express the very same desire-like attitude as ‘stealing is wrong’—for example, ‘lying is wrong’, ‘stealing is not wrong’, ‘if lying is wrong, then stealing is wrong’, and so on. So anyone who accepts any of these sentences whatsoever must already have this desire-like attitude. This leads to the conclusion that anyone who has any thoughts at all about what is wrong—or not—must already have the desire-like attitude expressed by ‘stealing is wrong’. So on the assumption that you have some other views about what is wrong, the only gap between thinking that stealing is K and thinking that it is wrong is just the gap associated with their respective modes of presentation—because you do have everything else that it takes.

So on the assumption that you have some other views about what is wrong, the ‘openness’ that you feel when you contemplate the Open Question argument can’t be the something extra that the hybrid view allows us to diagnose, stemming from the additional desire-like attitude that you need to have. You already have that attitude, after all. So if the hybrid theory gives us any diagnosis of Open Question phenomena that goes beyond what we get from any solution to Frege’s Puzzle, it is going to be somewhat more subtle than we have so far articulated.

[56] See Mark Schroeder, *Slaves of the Passions* (Oxford: Oxford University Press, 2007), 65; and compare Stephen Darwall, Allan Gibbard, and Peter Railton, “Toward *Fin de Siècle* Ethics: Some Trends,” *Philosophical Review* 101 (1992): 115–89.

To see how this more subtle diagnosis might work, consider the following three sentences:

1. Max believes that stealing is K, but does Max believe that stealing is wrong?
2. Max believes that stealing is wrong, but does Max believe that stealing is wrong?
3. Max believes that stealing is wrong, but does Max believe that stealing is K?

Now, one aspect of the Open Question phenomena is the idea that question 2 "answers itself" in a way that question 1 does not. There is, in fact, some evidence that this is not merely an ordinary Frege's Puzzle case, because it seems asymmetric. That is, question 3 doesn't seem "unanswered" in the way that question 1 does.

The hybrid theory can explain these observations on a certain hypothesis about attitude ascriptions, which I'll call the Big Hypothesis and which Boisvert explicitly endorses to just this kind of end:

> Big Hypothesis: If 'P' is a sentence expressing mental states $M_1 \ldots M_n$, then the descriptive content of 'S believes that P' is that S is in each of mental states $M_1 \ldots M_n$.

What the Big Hypothesis says is that attitude ascriptions (at least, 'believes that' ascriptions) attribute not only belief in the descriptive content of the complement clause but the desire-like attitudes that it expresses as well. The Big Hypothesis makes it easy to see the difference between these questions:

1. Max believes that stealing is K, but does Max believe that stealing is K and desire not to do what is K?
2. Max believes that stealing is K and desires not to do what is K, but does Max believe that stealing is K and desire not to do what is K?
3. Max believes that stealing is K and desires not to do what is K, but does Max believe that stealing is K?

Clearly, when so understood, the second question "answers itself" in a way that the former question does not—even once we are clearheaded about the descriptive content of 'wrong'. The same reasoning also predicts the intuitive asymmetry between questions 1 and 3 because, unlike question 1, question 3 does "answer itself" on this reading—though this is something that someone may fail to recognize, of course, for ordinary Frege's Puzzle reasons.

So though hybridism doesn't give us a diagnosis of every interesting Open Question phenomenon—in particular, on the assumption that you really do have at least some views about what is right or wrong, it can't give any different explanation of why questions like 'stealing is K, but is it wrong?' feel "open" to you that goes over and above what any solution to Frege's Puzzle will tell you. But it does give us a nice diagnosis of the differences between questions 1, 2, and 3—and similar phenomena arising through other uses of attitude ascriptions. However, this diagnosis works only on the assumption of the Big Hypothesis.

C. Judgment Internalism

It is also widely thought that hybrid views will facilitate elegant and noncoincidental explanations of moral motivation. After all, according to hybrid theories, accepting a moral sentence involves more than just having an ordinary belief. It involves that plus having some desire-like attitude, and we've seen that most hybrid theories systematically associate the belief and the desire-like attitude expressed by moral sentences, so that together they will be the right kind of pair to motivate someone to act.

But just as the initial, attractive thought about the hybrid diagnosis of the Open Question argument was too hasty, it is also too hasty to jump to the conclusion that hybrid theories really have the resources to give us an exciting explanation of moral motivation in cases of new moral judgments, which is unavailable to ordinary externalist cognitivists. Every hybrid theory that answers no to question Q1 accepts that for each speaker and each time, there is some desire-like attitude such that if that speaker has any views at all about what is wrong—or not—she must already have that desire-like attitude. This is a major background assumption. Moreover, it is not hard to see that it is this background assumption, and not the actual hybrid features of these views, which does the work in explaining moral motivation.

To see that this is so, notice that for any given hybrid view which says no to question Q1, we can construct an ordinary externalist, cognitivist theory, which assigns moral sentences the same descriptive contents as the hybrid view does and makes the same background assumption about desires but does not claim that these desires are expressed by the sentences. So, for example, if the hybrid theory says that the descriptive content of 'wrong' is 'K', the externalist cognitivist imitator says the same thing and assumes, along with the hybrid theorist, that pretty much everyone who has ever lived (since pretty much everyone who has ever lived has had at least some view about what is wrong or not) desires not to do what is K. Then whenever anyone forms a new moral belief about what is wrong, the externalist cognitivist imitator explains that she will be motivated by this background desire not to do that thing.

So hybrid theories which say no to question Q1 do not, after all, help us to explain cases of new moral motivation by appeal to resources that couldn't be adopted by any old ordinary externalist, cognitivist theory. So just as with the Open Question phenomena, if they are to get any leverage at all that ordinary cognitivists cannot, it is going to have to be more subtle than this.

Again, the Big Hypothesis comes to the rescue. If we assume the Big Hypothesis, then the truth of the following sentence turns out to follow from the meanings of the terms involved:

> Judgment internalism: Necessarily, for all *x*, if *x* believes that stealing is wrong, then *x* will be motivated, other things equal, not to steal.

This is because the hybrid theory and the Big Hypothesis together tell us that '*x* believes that stealing is wrong' has the descriptive content, '*x* believes that stealing is K and desires not to do what is K'. So if we assume further that the nature of beliefs and desires is to motivate us, other things being equal, to do what, given the truth of the beliefs, would attain the object of the desire, then we note that any possible individual who satisfies '*x* believes that stealing is wrong' will also satisfy '*x* believes that stealing is K and desires not to do what is K'. But since—given the truth of the belief that stealing is K—not stealing is how to attain the object of the desire not to do what is K, it then follows from the nature of belief and desire that this possible individual will be motivated, other things equal, not to steal.

So although hybrid theories can't really help to explain actual cases of moral motivation by appeal to resources unavailable to ordinary externalist cognitivists, they can obtain something that many theorists have wanted: they can explain how moral motivation could be both necessary and in some sense a "conceptual truth"—that the truth of judgment internalism is somehow guaranteed by the meanings of the words involved. As with the diagnosis of the Open Question phenomena in the last section, they can obtain this result given the Big Hypothesis.

I am not myself so certain that we should want to obtain either the diagnosis of the Open Question phenomena in the last section or the explanation of judgment internalism in this one. I am especially unsure that we should want to explain the truth of judgment internalism—particularly in such a strong form. The diagnosis of the Open Question phenomena seemed nice, in its way, but it leaves us with the same old Frege's Puzzle diagnosis of what is going on in some of the other more salient Open Question phenomena. Moreover, the assumptions that we need in order to get this diagnosis of the Open Question phenomena also predict a very strong (I would say too strong) form of judgment internalism. So given the costs, I'm inclined to think we can get on just fine by doing without the diagnosis of the Open Question phenomena, anyway.

Still, I can definitely see why for some theorists these consequences would be well worth going in for a hybrid view along the lines of Boisvert's, to supplement their existing cognitivist, realist commitments. From the fact that there are many problems that such a view would not solve, it does not follow that the advantages it has are not well worth having. Still, in order to have them, we need the Big Hypothesis. So let's look a bit at what grounds the hybrid theorists have for assuming it.

IX. DO PEJORATIVES SUPPORT THE BIG HYPOTHESIS?

A. The Analogy with Pejoratives

The overall trend in this article has been to encourage downward and rightward movement across figure 1—I've been pushing the view that the most promising hybrid theories are going to look broadly like Daniel Boisvert's account. Along the way, we've seen that pejoratives—words like 'kraut' and 'n——r'—can seem, at first glance, to have precisely these sorts of features. Pejoratives do seem to have dual aspects to their meaning, both to have a descriptive meaning and to endorse or at least manifest some sort of negative attitude—contempt of some kind, in most cases. Boisvert and Copp have taken the analogy to pejoratives very seriously in developing their views, and Copp has gone so far as to suggest that we model 'wrong' on however it turns out that pejoratives work.

Insofar as using pejoratives for our model suggests that things will be as in the lower right-hand corner of our table, I think that it has been fruitful—for, after all, I've been pushing that this is probably the most promising way to go. But the pejorative model is problematic, in the context of the Big Hypothesis. This is because it is far from clear whether pejoratives obey the Big Hypothesis, and there is much to suggest that they don't.

Here is a direct argument that pejoratives don't obey the Big Hypothesis: suppose that Nice Guy has no prejudicial attitudes about people with dark skin and would never use the word 'n——r', but that Bigot is (surprise) a bigot and uses it frequently with contempt. And furthermore suppose that Nice Guy and Bigot correctly believe that Loretta is an African American and know this about each other as well as each others' attitudes. Now consider the sentences, 'Nice Guy thinks that Loretta is a n——r' and 'Bigot thinks that Loretta is a n——r'. Boisvert argues that, in these sentences, the negative attitude associated with 'n——r' is attributed to Nice Guy and to Bigot, respectively, rather than expressed by the sentence as a whole. Since he thinks that 'wrong' is on a par with pejoratives, that is why he thinks that 'wrong' obeys the Big Hypothesis.

But let's evaluate this claim. If 'n——r' obeys the Big Hypothesis, that predicts that if Nice Guy says, ‘Bigot thinks that Loretta is a n——r’, he speaks truly, and if Bigot says, ‘Nice Guy thinks that Loretta is a n——r’, he speaks falsely. But this is peculiar. I would expect Nice Guy to resist characterizing Bigot’s views in this way—to avoid using the word ‘n——r’ at all. And I would expect Bigot to have no hesitation whatsoever in characterizing Nice Guy’s views in this way. If Bigot really uses ‘n——r’ as his word for people of African descent, pejoratively or not, he’s not going to substitute something else like ‘African American’ when describing Nice Guy’s views—unless perhaps he is mocking him.

This leads me to suspect that with pejoratives like ‘n——r’, attitude ascriptions don’t ascribe the associated negative attitude to their subject and do express the associated negative attitude themselves. So this leads me to suspect that if ‘wrong’ really works like a pejorative, then we should not expect it to obey the Big Hypothesis either. This is bad for the hybrid theory if it is true. Not only does it yield the counterintuitive conclusion that sentences like ‘Al thinks that stealing is wrong’ themselves express a desire-like attitude, but it also fails to make good on the Big Hypothesis and, hence, on the key potential attractions for the hybrid theory.

B. Back to Our Observation about ‘True’

I believe that my conjecture that pejoratives do not obey the Big Hypothesis is also indirectly supported by consideration of the following sort of argument, familiar from Section VI.C:

P21: Bigot thinks that Loretta is a n——r.
P22: Everything Bigot thinks is true.
C11: Loretta is a n——r.

Asserting C11 evinces a contemptuous attitude toward people of African descent. And this conclusion follows validly from the two premises. Since valid arguments have the inference-licensing property, someone who accepts the premises is committed to going on to accept the conclusion (or else giving up one of the premises). If we follow the hybrid idea from Section III that a desire-like attitude associated with the conclusion must already be associated with one of the premises, that means that either P21 or P22 must express this same contemptuous attitude.

Given the choice between concluding that ‘Everything Bigot thinks is true’ expresses a contemptuous attitude toward people of African descent, however, and concluding that ‘Bigot thinks that Loretta is a n——r’ does so, the choice is easy. Ridge’s account of truth solved this problem by claiming that ‘true’ expresses the desire-like attitudes expressed by other sorts of sentences. But it is clearly possible to think that everything Bigot thinks is true without having any contemptuous attitude toward people of African descent (though not, of course, if you recognize that Bigot is a bigot). So for pejoratives like ‘n——r’, we should reject this solution and say instead that attitude ascriptions with pejoratives in their complements are associated with negative attitudes in the very same way as those complements are themselves. That makes sense of why, intuitively, Nice Guy should resist characterizing Bigot’s views by saying, ‘Bigot thinks that Loretta is a n——r’.

This reasoning supports the conclusion that attitude ascriptions with pejorative complements associate the negative attitude with the speaker—and hence, if ‘wrong’ is really like a pejorative, the conclusion that ‘Al thinks that stealing is wrong’ itself expresses the same desire-like attitude as ‘stealing is wrong’ does. It doesn’t directly support, however, the conclusion that attitude ascriptions with pejorative complements don’t also have as part of their descriptive content that their subject has the negative attitude. It is compatible with this observation, that is, that ‘Bigot thinks that Loretta is a n——r’ both has a descriptive content that is true only if Bigot has a certain contemptuous attitude and expresses that very contemptuous attitude itself. So it is only indirect evidence, on the assumption that the attitude will be associated either with the speaker or with the subject but not with both. Still, if ‘wrong’ works like this, then even reports to the effect that someone does not think that something is wrong are going to express the same attitude as ‘wrong’ sentences themselves express.

C. Further Evidence and Directions for Progress

In *The Logic of Conventional Implicature*, Christopher Potts surveys a wide range of “expressive” constructions in natural languages, including epithets like ‘damn’, infixes like ‘fucking’, and the Japanese honorific. (He doesn’t discuss racial slurs directly.)[57] One of his primary theses is that the Big Hypothesis fails for all of these constructions. The special expressive character of each of these kinds of construction travels all of the way up to the speaker, even when buried under attitude ascriptions. Moreover, Potts argues, none of these constructions makes any contribution to the truth conditions of any of the constructions under which it is embedded—including attitude ascriptions. For example, to borrow just one example from Potts:[58]

> Clinton: The damn Republicans should be less partisan.
>
> Bush: Clinton says the damn Republicans should be less partisan.

It is clear that the second sentence is not how we would expect the real-

[57] Christopher Potts, *The Logic of Conventional Implicature* (Oxford: Oxford University Press, 2005).

[58] Ibid., 160.

world Bush to report Clinton's utterance. Rather than affecting the truth conditions of the sentence, 'damn' travels all of the way up to express the attitude of the speaker. And many other examples are like this as well.

The fact that not only pejoratives but so many other "expressives" fail to respect the Big Hypothesis is somewhat discouraging, if hybrid theorists are looking for any close analogy. One might suppose that a better candidate might be to focus on presuppositional constructions like 'knows':

1. Al thought he knew the answer to the question, but he got it wrong.

Here we see that the factiveness of 'knows' sticks to Al rather than becoming a commitment of the speaker—that is why the speaker can go on to note that Al got the answer wrong. This seems like a promising start, insofar as we want, along the lines of the Big Hypothesis, to get the expressive content of 'wrong' to stick to the subject of attitude ascriptions rather than to be expressed by the speaker.

But presuppositional phenomena like the factiveness of 'knows' don't transmit through all of the constructions other than attitude verbs for which the hybrid theorist who answers no to question Q1 needs them to. For example, when 'knows' appears in the consequent of a conditional, the speaker is not committed to the truth of its complement:

2. If it is true that the answer is 17, then Al knows that the answer is 17.

So that is no good. What the hybrid theorist really needs is a family of constructions which always commit the speaker to the same thing even when they are embedded under negation and in conditionals and disjunctions, but for which that commitment is transferred to the subject of an attitude report.

A potentially helpful example is 'but':[59]

3. It is not the case that Shaq is huge but agile.
4. If Shaq is huge, then he is huge but agile.
5. Marv said that Shaq is huge but that he is agile.

If someone who asserts sentence 3 or 4 is committed to the contrast between size and agility but sentence 5 reports that Marv is committed to the contrast, then 'but' is the kind of model that the hybrid theorist is looking for. This is not a total surprise; hybridist advocates of the idea

[59] These examples are adapted from Kent Bach, "The Myth of Conventional Implicature," *Linguistics and Philosophy* 22 (1999): 367–421.

that expression is conventional implicature, including both Copp and Barker, have specifically drawn attention to examples like sentence 5 in order to support their case. But if hybrid theorists are to make good on the potential advantages mentioned in Section VIII, they are going to need to pursue this idea more aggressively and consequently depart from fully embracing the model of pejoratives.

D. Where We Are

In this article I've tried to impose some structure on the discussion of the wide and interesting class of hybrid metaethical theories, by forcing reflection on the theoretical considerations which come into play at a relatively small group of important choice points in the development of such theories. It hasn't been my objective to argue against these theories, although I think it is clear that some are more promising than others, and I've tried along the way to indicate some of the reasons why. It has instead been my objective to try to create understanding of just which potential advantages come easily from such views and from which ones they arise and how, as well as of which advantages are more elusive. And of course I've tried to show at least some of the places in which existing hybrid views need substantial filling out and to offer some suggestions about where they should look for progress.

I'll close with a question I find hard to articulate, but to which I nevertheless feel that a successful hybrid view ought to owe us some satisfactory answer. When we notice that pejoratives like 'n——r' and 'kraut' are not purely descriptive, we look for different ways of talking about the same thing—ways which don't involve us in commitment to those kinds of attitudes. Once we find out that moral terms have an ordinary descriptive content but are also associated with further attitudes, why isn't the same a natural response? Stevenson anticipated something much like this question, and answered:

> Those who cherish altruism, and look forward to a time when a stable society will be governed by farsighted men, will serve these ideals poorly by turning from present troubles to fancied realms. For these ideals, like all other attitudes, are not imposed upon human nature by esoteric forces; they are a part of human nature itself. If they are to become a more integral part of it, they must be fought for. They must be fought for with the words "right" and "wrong," else these attitude-molding weapons will be left to the use of opponents.[60]

The sentiment here is clear and admirable, but the difference from the contemporary wave of hybrid theories is also visible in the phrase

[60] C. L. Stevenson, *Ethics and Language* (Oxford: Oxford University Press, 1944), 110.

"attitude-molding weapons." For Stevenson, the noncognitive component of moral words was not their expressive character, a matter of conveying or expressing the attitudes of the speaker, but rather their causal power. Moral words are powerful, Stevenson thought, because they are more effective at inducing emotions and attitudes in their audience. They contrast with descriptive words, as he put it on another occasion, as irrigating a desert contrasts with describing it.[61]

For Stevenson, I can see why it is worth going on with moral language—because there is something to be won with it. For contemporary hybrid theories, I'm not so sure. After all, we already have the associated desire-like attitudes, according to hybrid theorists. At least, anyone who has any views about what is wrong or not does. So holding that background assumption fixed, why can't we do just as well with ordinary descriptive terms?

[61] C. L. Stevenson, "The Emotive Meaning of Ethical Terms," reprinted in his *Facts and Values* (Westport, CT: Greenwood, 1963), 10–31, 16.