Philosophical Issues, 19, Metaethics, 2009

EXPRESSIVISM AND CONTRARY-FORMING NEGATION

Terry Horgan and Mark Timmons
University of Arizona

A well known challenge faced by any version of metaethical expressivism is to provide an adequate treatment of the so-called Frege-Geach problem. The task is to make sense of logically complex sentences like 'If I insulted the host then I ought to apologize', to do so by making sense of the states of mind such sentences would be used to express, and to preserve the validity of intuitively valid arguments, e.g., "If I insulted the host then I ought to apologize; so, since I did insult the host, I ought to apologize." The overall Frege-Geach problem includes, as a special case, the problem of making sense of negative sentences like 'Murder is not wrong', making sense of the states of mind expressed by such sentences, and explaining why the states of mind respectively expressible by 'Murder is wrong' and 'Murder is not wrong' are logically inconsistent with one another. This special case raises certain specific difficulties of its own for expressivism, over and above the familiar difficulties raised by the generic Frege-Geach problem (Unwin 1999, 2001, Dreier 2006a, 2006b, Schroeder 2008). We will call this particular package of difficulties the *negation problem*.

The plan of this paper is as follows. In section 1 we describe the negation problem. In section 2 we propose a solution that looks potentially available to various different versions of expressivism, including the version that we ourselves espouse (Horgan and Timmons 2006). In section 3 we argue that our solution is theoretically preferable to two alternative solutions currently on offer in the literature on this topic. In the appendix we briefly describe how to modify and improve our own version of expressivism ("cognitivist expressivism"), including its treatment of the generic Frege-Geach problem, in order to incorporate the treatment of the negation problem proposed in the present paper.

1. The Negation Problem

We will set forth the negation problem in a way that largely draws upon the especially clear presentation of it in Schroeder (2008). Consider, say, the two sentences 'Murder is wrong' and 'Murder is not wrong'. For the expressivist, the former expresses a certain state of mind—one that does not purport to represent murder as instantiating a putative in-the-world property of *wrongness*, and does not purport to represent a putative, potential, in-the-world state of affairs *murder being wrong*. Rather, this state of mind is a certain non-property-attributing attitude toward murder. Call this attitude *disapproval* of murder—letting the term 'disapproval' function here as a free parameter, a schematic dummy-expression for whatever state of mind a given version of expressivism might associate with the moral use of the word 'wrong'. Different expressivist theories could gloss disapproval in different ways, and might indeed deploy different terminology that better suits the state under its preferred gloss.

What state of mind is expressed by 'Murder is not wrong'? How is this state of mind related to the one expressible by 'Murder is wrong'? And why exactly are the two sentences, and the states of mind they express, logically inconsistent—so that a rational agent could not simultaneously have both states of mind at once? A particular worry that arises in this connection was pointed out by Unwin (1999, 2001). Consider these four sentences:

w Jon thinks that murdering is wrong.
n1 Jon does not think that murdering is wrong.
n2 Jon thinks that murdering is not wrong.
n3 Jon thinks that not murdering is wrong.

If we take thinking that murdering is wrong to be disapproval of murdering, then how are we to make sense of each of n1-n3, in a manner that accommodates the fact that none is equivalent to any other? In particular, how are we to make sense of n2, and in a way that makes clear why the states of mind expressed by w and n2 are logically inconsistent with one another? Schroeder puts the worry as follows. (This and other quotations from Schroeder are labeled, for purposes of subsequent cross reference.)

(A) But there are simply not enough places to insert a negation in 'Jon disapproves of murdering' to go around:
w* Jon disapproves of murdering.
n1* Jon does not disapprove of murdering.
n2* ???
n3* Jon disapproves of not murdering.

> There is simply one place not enough for the negations to go around. There is no way to account for the meaning of n2 by applying 'not' somewhere to the meaning of w. And that makes it look very much like expressivists are not going to be able to offer a satisfactory explanation of why 'murdering is wrong' and 'murdering is not wrong' are inconsistent. (pp. 578–579)

One initially natural-looking move for the expressivist to make, at this juncture, would be to appeal to another logically primitive attitude alongside disapproval, in order to capture in expressivist terms the familiar distinction between the notions of moral impermissibility and moral permissibility. We can use 'toleration' as a generic dummy-expression here.[1] The idea then would be that the slot for n2* gets filled with this: 'Jon tolerates murdering'. The worry, though, is whether this approach can provide an explanation of why tolerating murdering and disapproving of murdering are *logically incompatible* states of mind. Thus Schroeder:

> (B) [T]he problem is not that expressivists have no answer as to what n2 means. The answer is that it means that Jon tolerates murdering. But once we do things this way, it should be very clear that we have left completely unexplained and apparently inexplicable why 'murdering is wrong' and 'murdering is not wrong' are inconsistent.
>
> Suppose, for example, that someone tells you that when she uses the word 'not' or the prefix 'im-' immediately before 'permissible', they are not to be understood as meaning what 'not' normally does. Instead, she says, she believes in distinct, unanalyzable, and non-interdefinable properties of permissibility and impermissibility. And then suppose that she tells you that she also believes that it is impossible—*logically* impossible—for something to be both permissible and impermissible. Finally, she tells you, by 'not permissible' she means 'impermissible' and by 'not impermissible' she means 'permissible'. That is why, she tells you, 'murdering is permissible' and 'murdering is not permissible' are logically inconsistent sentences. It is because the latter means 'murdering is impermissible', and permissibility and impermissibility are assumed to be logically incompatible.
>
> Surely this account leaves something to be explained! Obviously her view will be a bad view about permissibility and impermissibility unless they *do* turn out to be incompatible. But that does not mean that she is entitled to assume it! On the contrary, her view seems to have written out of existence everything that could be used to explain why permissibility and impermissibility are incompatible, and given us an account of why this sentence and its negation are inconsistent that appears to have nothing to do with the meaning of 'not', into the bargain. Expressivists are in the same position with respect to

> disapproval and tolerance. The negation problem shows that they can't simply be interdefined, which leads to the conclusion that they are distinct and unanalyzable attitudes. But if they are, then why on earth is it inconsistent to hold them toward the same thing? (pp. 580–581)

Faced with this problem about construing disapproval and tolerance as logically primitive attitudes, neither of which is definable in terms of the other, the expressivist might try securing their logical incompatibility by seeking to define one in terms of the other. For instance, one might try to define the state *tolerating p* this way: *not disapproving of p*. But the trouble is that the state of mind so characterized is the one attributable by statements like n1 above, whereas what was wanted was to characterize states of mind attributable by statements like n2.

So the negation problem, in a nutshell, is this: In order to capture the state of mind expressed by 'murder is not wrong', the expressivist apparently needs to invoke a second attitude—toleration—that is logically primitive rather than being definable via disapproval. But that maneuver apparently leaves the expressivist without the resources to explain why—and in what sense—the attitudes expressed by 'murder is wrong' and 'murder is not wrong' are logically incompatible with one another.

2. A Proposed Solution

Perhaps you find yourself, as we ourselves do, with an intuitive sense that the negation problem for expressivism is not really as hard as passage (B) makes it out to be—and that this is because the hypothetical interlocuter described in passage (B) is not really being sophistical and explanatorily vacuous, but rather is on to something important and fundamentally sound. We will propose a solution to the negation problem that is inspired by this thought, and that honors it.

The closing sentences of passage (B) pose a putative dilemma: either find a way to define one or both of the attitudes toleration and disapproval so that the definition(s) explain why it is logically inconsistent to simultaneously hold both attitudes toward the same thing, or repudiate definability and admit that one has no satisfactory explanation of why such attitudes are logically inconsistent. The first alternative is what we will call the *definition tactic* for solving the negation problem. In principle, it might be successfully implementable despite the fact that neither attitude is directly definable in terms of the other one (plus negation). For, perhaps one or both of them can be defined some other way—and in a manner that provides the needed explanation of inconsistency.

We will not attempt to implement the definition tactic. Instead, we will propose a solution that goes between the horns of the putative dilemma just described. The idea that there is a genuine dilemma here rests, we

suspect, on the presupposition that there is only one available form of negation to appeal to—viz., the one expressible by the standard negation-symbol of formal logic. Challenging that assumption will be central to our approach.[2]

Many familiar concepts exhibit a feature we will call *trivalence*: they pick out a feature F that has an associated "anti-feature" that is logically contrary to F, thereby effecting a tripartite categorization of things—those instantiating the feature F, those instantiating F's anti-feature, and those instantiating neither F nor the anti-feature. For instance: some experiences are *pleasant*, some are *unpleasant*, and some are neither; some arguments are *persuasive*, some are *unpersuasive*, and some lie in between; some ways of addressing a task are *effective*, some are *ineffective*, and some fall under neither category. Trivalent concepts are legion, and such examples are easily multiplied.

Many natural languages, including English, have syntactic devices that are applicable to a trivalent predicative expression E this way: if E picks out a feature F, then the device operates on E to yield a more complex predicative expression that picks out the anti-feature of F. In English, for instance, a commonly used device of this kind—as illustrated by the examples just given—is an appended adverbial prefix such as 'un-', 'in-', or the like. We will call this syntactic phenomenon *contrary-forming negation*.[3] As the label suggests, this syntactic operation is indeed a form of negation: it picks out a feature that is logically contrary to (although in general not logically *contradictory* to) the feature picked out by E.[4]

Standard formal logic does not model the phenomenon of trivalence, and does not include any device of contrary-forming negation. (Likewise for familiar extensions of standard formal logic, such as modal logic and deontic logic.) But, in the present context, this fact is best viewed as a *limitation* of standard formal logic. Trivalent concepts and predicates, and contrary-forming negation, are common and genuine features of thought and language. Moreover, a feature F picked out by a trivalent predicate E is related *logically* to the anti-feature of F picked out by the contrary-forming negation of E: the two features are logically contrary to one another. In order to come to terms with the negation problem for expressivism, we suggest, a requisite first step is to acknowledge these facts, and to embrace a conception of logic that is rich enough to incorporate them.

A natural and useful way to do this would be to enrich standard formal logic itself. Here is one simple way of doing so (although others could be envisioned too). Syntactically, let '↓' be an operator that attaches to a predicate E to form a logically complex predicate ↓E; the symbol '↓' is to be the formal device of contrary-forming negation. Semantically, for each predicate E in the formal language, let a semantic interpretation I of the language assign to E both an *extension* and an *anti-extension*. The extension is a set of (ordered n-tuples of) objects in the domain of I; so is the

anti-extension; also, the extension and the anti-extension must be disjoint from one another. Each complex predicate ↓E is also assigned an extension and an anti-extension by I: the extension of ↓E is the anti-extension of E, and the anti-extension of ↓E is the extension of E.[5]

Return now to the negation problem. Schroeder offers the following piece of methodological advice about how to approach this problem:

> (C) To see the general shape of the strategy that is required, all that we need to do is to employ the *basic expressivist maneuver*. The basic expressivist maneuver is simple. Whenever you are encountered with a problem, what you do is ask yourself what it would take to reconstruct the same problem for ordinary descriptive language. Since there is obviously no such problem for descriptive language, you use this in order to isolate the feature of our view that is creating the problem. And then you construct an answer to the problem that is based on your understanding of why ordinary descriptive language avoids the problem. (p. 587)

This is good strategic advice, and we propose to implement it—albeit without deploying the definition tactic. Here is a way of trying to reconstruct a putative negation problem for descriptive discourse, modeled very closely on passage (B) from Schroeder:

> Suppose, for example, that someone tells you that when she uses the word 'not' or the prefix 'im-' immediately before 'potent', they are not to be understood as meaning what 'not' often does, viz., contradictory-forming sentential negation. Instead, she says, she believes in distinct, non-interdefinable properties of potence and impotence. And then suppose that she tells you that she also believes that it is impossible—*logically* impossible—for something to be both potent and impotent, because these two features are logical contraries of one another. Finally, she tells you, by 'not potent' she means 'impotent' and by 'not impotent' she means 'potent'. That is why, she tells you, 'Prayer is potent' and Prayer is not potent' are logically inconsistent sentences. It is because the latter means 'Prayer is impotent', and potence and impotence are logical contraries of one another.
>
> Surely this account leaves something to be explained! Obviously her view will be a bad view about potence and impotence unless they *do* turn out to be incompatible. But that does not mean that she is entitled to assume it! On the contrary, her view seems to have written out of existence everything that could be used to explain why potence and impotence are incompatible, and given us an account of why this sentence and its negation are inconsistent that appears to have nothing to do with the meaning of 'not', into the bargain. They can't simply be interdefined, because some things—moderately efficacious ones—are neither potent nor impotent. But, then, why on earth is it inconsistent to attribute both features to the same thing?

The putative problem posed in this passage is no *real* problem, of course. The reason why not is that the predicates 'potent' and 'impotent' are logically related as contraries (but not as contradictories), a fact that is syntactically encoded in their grammatical structure: 'impotent' is constructed from 'potent' by means of the logical operation of contrary-forming predicate-negation (in this case, via a negative adverbial prefix).[6] And the free-standing word 'not' can be used this way too—although of course it also can be used in the manner depicted in standard symbolic logic, viz., as a device of sentential negation. (The meaning of 'not' is not thereby abused, because 'not' expresses negation and there are these two distinct kinds of negation.) Thus, one way to implement the generic strategy that Schroeder calls "the basic expressivist maneuver"—the way that we ourselves recommend—is to invoke contrary-forming negation, and treat expressivist attitudes like *approval* and *disapproval* as logical contraries that are not logical contradictories.

On this view, Schroeder is mistaken to say concerning statements w, n1, n2, and n3 above, as he does in passage (A), "There is simply one place not enough for negations to go around." On the contrary, in addition to the two available places for (contradictory-forming) *sentential* negation, there is also an available place for contrary-forming *predicate* negation. Consider, for instance, these four sentences:

p Jane thinks that murdering is permissible.
np1 Jane does not think that murdering is permissible.
np2 Jane thinks that murdering is not permissible.
np3 Jane thinks that not murdering is permissible.

An expressivist can characterize the attitudes here attributed to Jane this way:

p* Jane is tolerant of murdering.
np1* It is not the case that Jane is tolerant of murdering.
np2* Jane is intolerant of murdering.
np3* Jane is tolerant of not murdering.

Sentence np2* deploys contrary-forming predicate negation, whereas sentences np1* and np3* deploy sentential negation in two different available positions.[7] Likewise, the expressivist can characterize as follows the attitudes attributed to Jon in statements w and n1-n3 above, now taking 'opposition' as a dummy-term characterizing the state of mind that, according to expressivists, is expressed by moral-evaluative uses of the word 'wrong'.[8]

w* Jon is opposed to murdering.
n1* It is not the case that Jon is opposed to murdering.
n2* Jon is unopposed to murdering
n3* Jon is opposed to not murdering.

Contrary-forming predicate-negation to the rescue!

Are toleration and opposition both logically *primitive*, on this approach? No. Rather, the pair toleration/intoleration is definable in terms of the pair opposition/unopposition, and vice versa—which means that either pair can be taken as primitive and then deployed to define the other. If the pair opposition/unopposition is treated as primitive, then being tolerant of Φ-ing is definable as being unopposed to Φ-ing, and being intolerant of Φ-ing is definable as being opposed to Φ-ing. Conversely, if the pair toleration/intoleration is treated as primitive, then being opposed to Φ-ing is definable as being intolerant of Φ-ing, and being unopposed to Φ-ing is definable as being tolerant of Φ-ing.

Suppose, say, that one chooses to introduce the pair toleration/intoleration by definition, thereby treating the pair opposition/unopposition as undefined. Are opposition and unopposition both logically primitive *themselves*, on this approach? Well, the rubric 'primitive' needs careful handling, once one countenances contrary-forming negation within logic. Unopposition is not *definable* in terms of opposition (or vice versa), and in that sense the two locutions are both being taken as logically primitive.[9] On the other hand, the word 'unopposed' is *syntactically* complex, being built from the word 'opposed' by appending a contrary-forming negation-operator. Furthermore, opposition and unopposition are not logically independent of one another, because it is a matter of logic that a contrary-forming negation-operator (in this case, the prefix 'un-') attaches to a predicate to form a logically complex predicate that picks out a feature that is related as anti-feature to the feature picked out by the original predicate. Thus, it is a matter of logic that the attitudes picked out respectively by the expressions 'being opposed to Φ-ing' and 'being unopposed to Φ-ing' are related to one another as feature and anti-feature—which makes these two features logically contrary to one another. Opposition is a trivalent notion, and this explains why it is logically impossible for a single individual, at a single time, to be both opposed to Φ-ing and unopposed to Φ-ing. So, although opposition and unopposition are being treated as primitive in one sense—i.e., as not interdefinable—this kind of "primitiveness" is no obstacle to explaining why these two attitudes are logically incompatible with one another. Thus, this approach goes between the horns of the putative dilemma concerning the negation problem: on one hand, it refrains from attempting to *define* either opposition and/or unopposition; but on the other hand, it nonetheless provides resources to explain why it is logically inconsistent to be simultaneously both opposed and unopposed to the same thing.

Does the discussion so far in this section amount to a *complete* expressivist solution to the negation problem? Not quite, because a residual burden is to articulate, in a way that conforms with expressivism, what it is to hold an attitude toward something that is related as anti-feature to the attitude of opposition—i.e., what it is to hold an attitude of unopposition.

But, as theoretical tasks go in philosophy, this one looks fairly easy to discharge. We turn now to some suggestions about how to discharge it. (Note, however, that our basic proposal for how expressivism can address the negation problem could be correct in any case, whether or not one accepts as adequate what we will now say about how to understand the nature of unopposition.)

We will consider first what is arguably the simplest case, and will then generalize from it. The simplest case involves an act-type Φ that one is contemplating performing (i.e., one is considering performing an act that would be a token of type Φ). Also, Φ is a *de se* act-type, in this sense: it is built into the nature of Φ that the performer is oneself. Suppose that you judge that it would be wrong to Φ—which, according to the expressivist, is a matter of opposing Φ-ing.[10] Then this attitude will dispose you to refrain from Φ-ing, and to so refrain *because* such an act would be of type Φ. I.e., the attitude will constitutively involve a disposition to behave-in-a-specific-way-for-a-specific-reason—a "motivated disposition," as we will call it. We will use brackets as a device for characterizing the content of a motivated disposition—the behavior toward which one is motivated, and the consideration(s) one would regard as a reason for so behaving. Thus, your attitude of opposing Φ-ing constitutively involves a disposition toward this: [refraining from Φ-ing because an act of Φ-ing would have the feature of being a Φ-act]. For example, imagine that you find yourself contemplating the act of lying to the IRS about your total annual income, and you form the judgment that this would be morally wrong. Your attitude of opposition toward so acting constitutively involves a disposition toward this: [refraining from lying to the IRS about your total annual income, because such an act would be a lie].

Suppose, on the other hand, that you judge that Φ-ing is not wrong, i.e., is permissible—which, according to the expressivist who adopts the treatment of negation currently on offer, is a matter of unopposition toward Φ-ing. Then this attitude will constitutively involve a disposition that is related as anti-feature to the motivated disposition lately mentioned. I.e., the attitude will constitutively dispose you toward *not* doing this: [refraining from lying to the IRS about your total annual income, because such an act would be a lie]. We will call this negative disposition the *anti-feature counterpart* of the motivated disposition described in the preceding paragraph.

Two aspects of this negative disposition should be noted. First, it is not a flat-out disposition to behave in a way that is contrary to refraining from lying to the IRS—i.e., it is not a flat-out disposition to go ahead and lie to the IRS. Rather, it is a disposition against a certain kind of *motivated* behavior, i.e., a disposition against this: refraining from lying to the IRS *because* such an act would be a lie. Even if you are morally unopposed to lying to the IRS, you might still have other motives that dispose you against doing so—e.g., fear of getting caught and punished.

Second, it is important to distinguish these two features: (1) *not* being disposed toward [refraining from Φ-ing because an act of Φ-ing would have the feature of being a Φ-act], vs. (2) being disposed toward *not* [refraining from Φ-ing because an act of Φ-ing would have the feature of being a Φ-act]. Feature (1) is logically weaker than feature (2); i.e., (2) entails (1), but not conversely. Feature (1) is the mere *absence* of the motivated disposition with content [refraining from Φ-ing because an act of Φ-ing would have the feature of being a Φ-act]. Thus, one could possess feature (1) merely by lacking any determinate dispositions at all with respect to Φ-ing. Feature (2), on the other hand, is a determinate disposition regarding Φ-ing, and is the anti-feature counterpart of the motivated-disposition feature whose content is [refraining from Φ-ing because an act of Φ-ing would have the feature of being a Φ-act].

Return now to the question of why the attitude of being unopposed to Φ-ing bears the anti-feature relation to the attitude of being opposed to Φ-ing, in the case where Φ-ing is a *de se* act that one is contemplating performing. The explanation for this logical relation between these two attitudes is that they respectively involve constitutive dispositions that themselves are related to each other as feature and anti-feature. Opposition to Φ-ing constitutively involves the motivated disposition whose content is [refraining from Φ-ing because an act of Φ-ing would have the feature of being a Φ-act]. Nonopposition toward Φ-ing, on the other hand, constitutively involves the anti-feature counterpart-disposition, viz., the disposition toward *not* [refraining from Φ-ing because an act of Φ-ing would have the feature of being a Φ-act]. In short, the two attitudes are logical contraries of one another because they respectively involve, constitutively, two dispositions that themselves are logical contraries of one another.

This account of why opposition and unopposition are related as feature and anti-feature can be generalized. In the general case, if one has an attitude of opposition or unopposition toward some item Ψ, that item need not be an act that one is contemplating performing oneself—and also need not be a constitutively *de se* act. Indeed, Ψ need not be an act at all (either an act-type or an act-token), but might be some other kind of object of moral evaluation, such as a state of affairs. The main point to bear in mind, in seeking a generalized account, is that attitudes of opposition and unopposition, like most other mental states, typically involve constitutive dispositions only *in combination with one another*, rather than singly. (This fact was the undoing of analytical behaviorism in philosophy of mind.)

To begin with, one can construe in a generalized way the notion of a motivated disposition. Let Φ again be a *de se* act-type, i.e., an act-type that has built into it the feature of being done *by me* (where 'me' is the essential first-person indexical). Suppose an agent believes that in circumstances C, an act of Φ-ing would have feature F. Let a *circumstance-specific motivated disposition toward* Φ*-ing*, D(Φ,C,F), be a disposition to

perform, when one believes oneself to be in circumstances C, an action with the following content: [Φ-ing because one believes that Φ-ing has feature F in circumstances C]. And let the *anti-feature counterpart* of D(Φ,C,F) be a disposition toward *not* performing an action with that content when one believes oneself to be in circumstances C.

Next, there is a need to accommodate the dispositional holism of the mental—the fact that in general, constitutive behavior-dispositions arise only from mental states in combination with one another. The natural way to do this is in terms of the idea of *total* mental state instantiable by an agent, where a total mental state can include an attitude of opposition, or an attitude of unopposition, as a part. Total mental states often will constitutively involve certain behavioral dispositions—whether or not those dispositions accrue to the component mental states considered singly.

Consider, now, the attitude of opposition toward some item Ψ, and the attitude of unopposition toward Ψ. What is it about any such pair of attitudes that makes them constitutively related as feature and anti-feature? The natural-looking answer comes in two stages. First, there is no total mental state Ω that contains as parts both the attitude of opposition toward Ψ and the attitude of non-opposition toward Ψ. Second, this is so because of the following fact about behavioral dispositions that are constitutively associated with total mental states:

> for any two total mental states Ω and Π such that Ω includes as a part the attitude of opposition toward Ψ, and Π includes as a part the attitude of unopposition toward Ψ,
>
> there is some *de se* act-type Φ, and there is some circumstance-specific motivated disposition D(Φ,C,F) toward Φ, such that
>
> (1) possessing the disposition D(Φ,C,F) is partially constitutive of Ω, and
> (2) possessing the anti-feature counterpart of the disposition D(Φ,C,F) is partially constitutive of Π.

The idea is that opposition toward Ψ and unopposition toward Ψ inevitably involve logically contrary behavioral dispositions, and do so in a way that guarantees that these two attitudes are logical contraries themselves—even though in general, mental states constitutively involve behavioral dispositions only holistically and not singly. This underlying idea is the key thing, too—whether or not the principle lately articulated constitutes a fully adequate explication of it. The attitudes of opposition and unopposition play an inherently action-guiding role in the psychological economy of a morally judging agent.[11] Given this fact, it is no great mystery that the attitudes of opposition toward Ψ and unopposition toward Ψ are logical contraries of one another.

3. Why the Proposed Solution is the Best Bet

We will now briefly consider two other approaches to the problem of negation that have been proposed recently on behalf of expressivism, and we will argue that ours fares better in terms of comparative benefits and costs.

Gibbard

Allan Gibbard (2003) sets forth a version of expressivism that differs to some extent from his earlier version in Gibbard (1990). One important change is his introduction of the notion of *disagreement with an attitude*, to which he gives a central role as a means for addressing various problems—including the problem of negation, as originally articulated in Unwin (1999, 2001).[12] The idea is that the attitude of thinking that promise-keeping is not obligatory, for example, is *disagreement with requiring promise-keeping* (where 'requiring' is being used as a dummy expression for whatever attitude an expressivist might claim is expressed by assertions of moral obligation).

Talk of disagreement with a moral attitude can be construed in more than one way, however, and Gibbard's own discussion does not point determinately to any one construal in particular. Is disagreement with a moral attitude supposed to be a moral attitude itself, or might it be an attribution of error that need not be a moral evaluation? (Jane might think that Dick is *mistaken* to think that premarital sex is morally wrong, without thinking that Dick's attitude is *itself* morally wrong.) Is disagreement with a moral attitude directed primarily toward that first-order moral attitude itself, or is it really directed primarily toward the same item toward which the first-order attitude is directed?

If Gibbard's idea is to be at all plausible as a way of understanding, in an expressivism-friendly manner, what state of mind one is attributing to Jane by saying "She thinks that premarital sex is morally permissible," then talk of disagreement with a moral attitude evidently needs to be understood as picking out a state of mind that (1) is a moral attitude itself, and (2) is directed primarily toward the same item as is the disagreed-with attitude. Jane's thinking that premarital sex is morally permissible is a *moral* attitude, and it is primarily directed not toward some first-order moral attitude regarding premarital sex, but rather toward premarital sex *itself*.

The upshot, it appears, is that an expressivist's talk of "disagreement with an attitude" is most charitably construed as a suggestive, but somewhat inaccurate and misleading, way of capturing the same idea that we ourselves are putting this way: notions like moral opposition are trivalent, and such an attitude therefore has an associated anti-feature that is its logical contrary (in this case, moral unopposition). Describing the anti-attitude as "disagreement with the attitude" is indeed suggestive, because such talk makes salient two

facts: first, that the anti-attitude somehow *logically conflicts* with the attitude, and second, that this conflict is somehow ultimately grounded, at least in part, by some kind of conflict in the *behavioral dispositions* that are associated respectively with the attitude and the anti-attitude. However, describing an attitude like unopposition to Φ-ing as "disagreement with opposition to Φ-ing" is also rather loose and off-target, because the attitude so described is a moral attitude directed primarily not toward the first-order attitude of opposition to Φ-ing, but instead toward Φ-ing itself. All things considered, therefore, not only is Gibbard most charitably interpreted as advocating the same position we have described here, but expressivists would be well advised to adopt our own lingo (or something similar) in place of Gibbard's talk of disagreement with an attitude.

Schroeder

Mark Schroeder (2008) argues that no prior expressivist treatments of the Frege-Geach problem generically, or of negation specifically, provide an adequate solution to the negation problem; he then offers a proposed solution of his own, on behalf of expressivism. Among the targets of his critique are the treatments of negation in Gibbard (2003) and Dreier (2006b), and our own treatment of the Frege-Geach problem in Horgan and Timmons (2006). We will not attempt to summarize or assess his criticisms of others. (We do agree with him, though, that our own earlier approach to the Frege-Geach problem did not adequately handle the negation problem—a matter we take up in the Appendix.) Instead we will focus on Schroeder's positive proposal.

Schroeder's recommendation, which he presents as the appropriate implementation of the basic expressivist maneuver described in passage (C) above, is this:

> (D) [T]he expressivist accounts we were considering did not have enough *structure*. So if the problem arises from a lack of structure, there is only one solution. It is to introduce more structure. That is my solution .... It doesn't really matter how things go, from here, but just to make things concrete, let's work with the attitude of *being for* .... Disapproval of murder, we can say, inspired by Gibbard [1990], is being for blaming for. (Alternatively, we could say that it is being for avoiding, or any number of other things, but we only need one example to see how the view works.) .... How does this work? Didn't we have an argument ... that tolerance and disapproval had to be logically unrelated? No; actually all that we had was an argument that disapproval and tolerance can't be interdefined using negation. Nothing showed that they can't both be defined in terms of some third attitude, and that is what I have done .... This solution

> works by creating an extra place in which to insert the negation needed in n2, and does so *in the very same way* as this works for descriptive sentences. Compare:
>
> w^+ Jon is for blaming for murdering.
> $n1^+$ Jon is not for blaming for murdering.
> $n2^+$ Jon is for not blaming for murdering.
> $n3^+$ Jon is for blaming for not murdering.
>
> . . . . If the problem arises because the expressivist account has insufficient structure, there is *only one* solution: to give the expressivist account sufficient structure. And that is my solution. *Ipso facto*, my solution is the only one that works. (pp. 589–590)

The approach described in this passage evidently includes three components: (1) the generic idea articulated in the first four sentences and the final three, involving the need for more structure, (2) the proposal to define both tolerance and disapproval in terms of some third attitude, and (3) an illustration of such a definition and how it solves the negation problem.

We ourselves are fine with the first component. Indeed, our own proposal implements it. The additional structure that we advocate introducing into an expressivist account is this: logically contrary pairs of attitudes, where a predicate attributing one of the two attitudes is constructable by the logical operation of contrary-forming negation from a predicate attributing the other attitude—as in 'opposed to' and 'unopposed to'.

The second component in passage (D) is what we earlier called the *definition tactic*. The form of definition Schroeder has in mind deploys some third attitude—e.g., *being for blaming for*—plus contradictory-forming negation. Schroeder does not explicitly distinguish this component from the first component, and he treats them as inextricably intertwined. And indeed, if one assumes that there is only one pertinent form of negation available for expressivists to appeal to, then it becomes very hard to see how component (1) could be implemented otherwise than in the manner of component (2). But the assumption is false; and, as just noted, our own proposal constitutes a non-definitional way of implementing component (1).

Even granting these last observations, however, this question remains: Why prefer our approach to Schroeder's? We will offer two reasons, which mutually support one another. First, even though there are presumably some significant conceptual constraints on what could count as a viable candidate-definition of 'being for', nonetheless it is prima facie very plausible that there are any number of different, non-equivalent, potential ways of defining disapproval in terms of being-for, each of which meets the conceptual constraints and each of which is just as good a candidate-definition as any of the others (e.g., being for blaming for, being for avoiding, etc.). There is no apparent reason, other than sheer unsubstantiated optimism, to think

that any single candidate-definition is superior to all others. But if this is so, then *none of these potential definitions accurately analyzes the notion of disapproval itself*. I.e., since there are too many equally eligible ways to fill in Schroeder's definition-schema, all different in meaning, none of them accurately captures the state of mind that constitutes full-fledged normative or moral disapproval itself. This threatened version of what Peter Unger calls "the problem of the many" would be a very significant theoretical cost—one that our own proposal does not face.

An additional serious cost is this. Candidate-definitions that fit Schroeder's suggested format all seem too "second-order" in their focus, as compared to the notion of disapproval itself. Disapproval of Φ-ing is a negative attitude directed primarily at Φ-ing *itself*—rather than being a positive "being-for" attitude directed primarily at something else, something such as *blaming* for Φ-ing or *avoiding* Φ-ing. Further underscoring this fact is the way disapproval of Φ-ing can, and sometimes does, coexist with an absence of one or another of such second-order attitudes. For instance, middle-aged Mom might disapprove of getting drunk, soliciting prostitutes, and getting into bar fights; yet she might not be in favor of blaming her son and his mates for doing these things, and she even might not be in favor of their avoiding such behavior. For, she might well think to herself, "Boys will be boys, and these boys who are serving their country as soldiers in Iraq deserve a bit of fun—especially since any one of them might get killed any day."[13]

The two liabilities lately mentioned reinforce one another, thereby compounding the net theoretical cost of Schroeder's proposal. If there were a variety of equally eligible candidate-definitions of disapproval, all of which decently reconstructed the ordinary notion, then perhaps a tolerable bullet-biting strategy for expressivists would be to just pick any one of them as the official characterization of disapproval—and maybe claim, in addition, that all there is to disapproval is its constitutive psychological role, a role allegedly played well enough by any of the eligible being-for attitudes.[14] But it appears that none of the candidate-definitions that fit Schroeder's recommended format really does sufficiently well at reconstructing the pre-theoretic idea of disapproval, because they are all too higher-order. Disapproval of Φ-ing is, first and foremost, a *negative* attitude toward Φ-ing itself—not a positive attitude toward something else, such as blaming for Φ-ing, avoidance of Φ-ing, or whatever. Our own treatment of the negation problem accommodates this fact, whereas Schroeder's does not.[15]

4. Conclusion

Contrary-forming negation is an important feature of language and thought, although it has been insufficiently appreciated in philosophy and is

not accommodated in standard formal logic. Once acknowledged, however, it provides the basis for a plausible and attractive way for expressivists to solve their problem with negation. The key idea is to construe as trivalent whatever attitude-ascribing concepts figure centrally in any specific version of expressivism.[16]

Appendix

In Horgan and Timmons (2006) we described and defended a position we call *cognitivist expressivism*. Here we will make some brief remarks about how to modify that discussion in some respects, and how to elaborate it in others, to incorporate the treatment of the negation problem we have proposed here. We will proceed staccato-style, and without attempting any overall summary of the earlier paper. The reader is urged to consult that paper too.

1. We claimed then that there are two logically fundamental kinds of commitment-states with respect to descriptive contents, is-commitment and ought-commitment. We introduced a formal language in which these two kinds of commitment are expressible by means of the operators I[ ] and O[ ] respectively, with a sentence resulting when a descriptive-content formula is inserted into the "slot" of one of these operators. (Descriptive-content formulas correspond to non-normative 'that'-clauses in ordinary language, e.g., 'that Cheney is imprisoned'.)
2. We would now add that the notion of an ought-commitment is trivalent, and thus that the associated, contrariwise negative, commitment-states are themselves logically fundamental in a certain way too. Call these states *not-ought* commitments. (The present usage of 'not' is to express contrary-forming negation, of course.) The sense in which both ought-commitments and not-ought-commitments are logically fundamental is this: both kinds of state typically play constitutive *behavior-motivating* roles in the psychological economy of morally judging agents.
3. We claimed then that there is a whole recursive hierarchy of logically complex commitment-types expressible via logical constructions out of I[ ] and O[ ], and that such a commitment-type can obtain with respect to a sequence of descriptive contents. For instance, (I[ ] → O[ ]) expresses a commitment-type of the kind "If it is that case that . . . , then it ought to be the case that . . . ." A morally judging agent can be in a commitment-state of this kind with respect to a two-part sequence of descriptive contents—for example, the sequence <*that Cheney is found guilty of treason*, *that Cheney is imprisoned*>.

That is what it would be for the agent to be in the state *thinking that if Cheney is found guilty of treason, then he ought to be imprisoned*.

4. We also claimed then that for any state-type in the recursive hierarchy other than simple is-commitment and simple ought-commitment, the full constitutive functional role of states of that type, in the psychological economy of a morally judging agent, is *inferential mediation*. I.e., the constitutive functional role of such a state is entirely a matter of that state's being poised to inferentially generate, in combination with other potential mental states, logically simpler commitment-states. (This leaves open the possibility that there are other, non-functional, aspects of the constitutive essence of a logically complex commitment-state—e.g., phenomenal aspects.) For instance, if you are in the lately-mentioned commitment-state concerning Cheney, and you also come to believe that Cheney has been found guilty of treason, then the first state should mediate an inferential transition from the newly acquired belief to a new ought-commitment state, viz., *thinking that Cheney ought to be imprisoned*. This general approach solves the Frege-Geach problem by turning it on its head: the *full constitutive functional essence* of any logically complex commitment-state is no more and no less than its inferential-mediation role in an agent's psychological economy, insofar as the agent's mental state-transitions are fully rational. Sameness of meaning for moral language, across unembedded and embedded occurrences, is a matter of the constitutive logical connections among the mental states that are fully or partially expressible via that moral language—a natural way of extending the familiar expressivist idea that the way to explain the meaning of an unembedded moral sentence is to say what mental state is expressible by that sentence.
5. We would now qualify the constitutive-essence claim mentioned in item 4, in light of our remarks in item 2. Not-ought commitments typically have a constitutive functional essence that goes beyond inferential mediation: they also typically play constitutive behavior-motivating roles too. (Often the latter roles are played holistically, in combination with other mental states—as we emphasized at the end of section 2 above.) This qualification overcomes one respect in which our earlier discussion failed to deal adequately with the negation problem.
6. We would now also add the following observations, concerning the logically complex commitment-type ¬O[ ]. The negation symbol in this construction is to be construed as contrary-forming, rather than as contradictory-forming. The construction ¬O[ ] is the canonical format for expressing not-ought commitments in the formal language, and not-ought commitments are the logical contraries of ought-commitments. (Our earlier discussion did not distinguish

contradictory-forming and contrary-forming negation.) The state-type *thinking p permissible* is now definable as having a not-ought commitment with respect to ∼p; in symbols, this attitude is expressible as ¬O[∼p].

7. In the Appendix of Horgan and Timmons (2006) we specified the syntax of a proposed formal language deploying the operators I[ ] and O[ ], and we gave a formal semantics for that language that allows some sentences to be assigned no truth value. We said that on one natural construal, the sentences assigned the truth value T by a given truth-value assignment would correspond to all and only the sentences of the language that either (i) express actual psychological commitments of a given agent, or (ii) express psychological commitment-states that are rationally mandated by the agent's actual psychological commitments.
8. We would now elaborate, as follows, on the just-described way of construing truth-value assignments. A truth-value assignment that reflects a given person's commitment-states will assign F to a sentence O[p] if and only if the agent has a not-ought commitment vis-à-vis the descriptive content p. (Under the recursion clauses of the formal semantics, the sentence expressing the not-ought commitment, viz., ¬O[p], will then be assigned T.) And the assignment will withhold a truth value from O[p] if and only if the person lacks both an ought-commitment and a not-ought commitment toward p, i.e., the person is undecided ought-wise concerning p. (In that case, under the recursion clauses, ¬O[p] also will receive no truth value.) Our earlier discussion was not in error on these matters, but was silent about them. They deserve explicit emphasis.
9. We would now update the formal semantics we gave earlier, by adding the following clause to the definition of a valuation: If A is a closed nonsentential formula, then **S** assigns T to **O[A]** only if **S** assigns F to **O[∼A]**. Our earlier discussion, by failing to include this clause, failed to accommodate the fact that the sentences **O[A]** and **O[∼A]** express logically contrary attitudes. (Under the recursion clauses, this new clause guarantees that **O[A]** logically entails **¬O[∼A]**, which says in effect that obligatoriness logically entails permissibility.)
10. The converse of the lately-mentioned clause is not wanted, however, because a morally judging agent could have a not-ought commitment vis-à-vis **∼**A without thereby having an ought-commitment vis-à-vis A. I.e., the agent could think that A is permissible but not obligatory.

Notes

1. It bears emphasis that everyday usage of 'tolerate' need not coincide with judgments of moral permissibility, because one sometimess tolerates (in the ordinary sense) behavior by others that one considers morally impermissible.

2. We return to the definition tactic in section 3.2, where we also explain why Schroeder's putative dilemma appears to rest on the assumption that there is only one pertinent kind of negation.
3. For two enlightening discussions of contrary-forming negation and related linguistic phenomena, see Lehrer and Lehrer (1982) and Lehrer (1985).
4. The free-standing word 'not' can be used as a device of contrary-forming negation too, especially when positioned immediately prior to a trivalent predicative expression—although of course 'not' also can be a device of sentential negation, even when so positioned. The contrary-forming usage can be signaled, for instance, by stress on 'not'—as in "Going to the dentist is *not* fun."
5. On this very simple way of doing things, there is no formal-syntactic distinction between predicates that are trivalent and those that are not. In effect, a predicate E is non-trivalent, under an interpretation I, just in case I assigns to ↓E an extension that is the set-theoretic complement of the extension that I assigns to E. I.e., the categories delimited by E and ↓E, under I, are jointly exhaustive.
6. A disanalogy between Schroeder's passage (B) and the present passage is that (B) uses the words 'permissible' and 'impermissible'—which are contradictories rather than mere contraries, since 'permissible' is not trivalent. (See note 5.) But although the choice of a non-trivalent term does enhance the rhetorical force of passage (B), it does so by making it easier not to notice the distinctive logical role played by negative adverbial prefixes when these are appended to trivalent verbs. (Note too that the *attitudes* ascribable by the expressions 'thinks Φ-ing permissible' and 'thinks Φ-ing impermissible' are contraries without being contradictories, because one can fail to think Φ-ing permissible and also fail to think Φ-ing impermissible—viz., by being normatively undecided about Φ-ing.)
7. Note that 'Jane is not tolerant of murdering' is grammatically ambiguous as between np1* and np2*. This form of negation-ambiguity is common, although often it is clear enough in context which reading is intended.
8. We use 'opposition' rather than 'disapproval' because the word 'disapprove' is logically complex itself, being constructed from 'approve' by contrary-forming negation.
9. At any rate, neither attitude is definable in terms of the other under the usual standards of adequate definition in philosophy and in logic. Unopposition is characterizable this way, however: the unique anti-feature of opposition.
10. Here and throughout, we are discussing only "all in" moral judgments rather than "prima facie" or "pro tanto" moral judgments. Doing adequate justice to the latter is one important item on the agenda of tasks that must be accomplished by a fully adequate version of expressivism, but this is a matter beyond the scope of the present paper.
11. At any rate, it is constitutive of the attitudes of opposition and unopposition that they *typically* play action-guiding roles, and logically contrary ones. This does not necessarily preclude the conceptual possibility of the occasional "amoralist" who makes genuine moral judgments but is utterly unmotivated by them. For, perhaps it suffices, in order for attitudes of opposition and unopposition to be correctly attributable to the members of a social group, that the attributed attitudes play their constitutive motivational roles in *most* members of the group. Moreover, even in the case of immoralists (if they are a genuine possibility), there would have to be *some* logically contrary dispositions associated with opposition

and unopposition respectively—e.g., dispositions toward distinct sincere moral assertions which, in normal members of the group, would express attitudes that constitutively involve logically contrary motivated-behavior dispositions.

12. Another important change is his reliance on contingency planning by hypothetical ideally-rational agents, as a model for understanding normative judgments. His overall treatment of the Frege-Geach problem is framed in terms of this planning model, and also relies heavily on the notion of a "hyperplan"—a plan which, for every conceivable contingency C and every available alternative action A in C, either allows or forbids A in C (and allows at least one alternative in C). We ourselves doubt, for a number of reasons, that the framework of idealized contingency planning can ground an adequate version of expressivism—but we cannot pursue that matter here. It bears emphasis that our own recommended approach to the Frege-Geach problem, as modified in the Appendix to incorporate the approach to the negation problem that we recommend in this paper, has no need of anything like Gibbard's hyperplan-based semantic machinery. For discussions of Gibbard's treatment of the negation problem and the overall Frege-Geach problem that address the ways his account relies on the model of ideal continency planning, see Dreier (2006a, 2006b), Gibbard (2006), and Schroeder (2008).
13. Horgan thanks his wife Dianne for this example. Dianne also reminded him that he himself is not allowed shore leave.
14. This would be something like arbitrarily picking one specific way of defining numbers in terms of sets. There are numerous equally eligible ways of doing so, and it is sometimes argued that any eligible candidate-definition may be adopted—the idea being that essentially all there is to the notion of a number is what is expressed by principles like the Peano axioms.
15. We do acknowledge that our approach is similar to Schroeder's insofar as ours too has a certain kind of higher-order aspect: on our account the state of being opposed to murder constitutively involves certain motivated dispositions (cf. section 2); and a motivated disposition is directed toward doing-something-for-a-reason, an act-type with a certain built-in higher-order structure. But important differences from Schroeder remain, because on our account the state *being opposed to murder* is a negative attitude toward murder itself rather than a positive attitude toward something else like punishing for murder.
16. We thank Matt Bedke, Jamie Dreier, Dianne Horgan, Adrienne Lehrer, Keith Lehrer, and Mark Schroeder for helpful feedback and discussion.

References

Dreier, Jamie (2006a). "Disagreement (about) What to Do: Negation and Completeness in Gibbard's Norm-Expressivism," *Philosophy and Phenomenological Research* 72, 715–722.

Drier, James (2006b). "Negation for Expressivists: A Collection of Problems with a Suggestion for their Solution," *Oxford Studies in Metaethics*, volume 1. Oxford: Oxford University Press, 217–233.

Gibbard, Allan (1990). *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press.

Gibbard, Allan (2003). *Thinking How to Live*. Cambridge: Harvard University Press.

Gibbard, Allan (2006). "Reply to Critics," *Philosophy and Phenomenological Research* 72, 730–745.

Horgan, Terry and Timmons, Mark (2006). "Cognitivist Expressivism." In T. Horgan and M. Timmons, eds., *Metaethics after Moore*. Oxford: Oxford University Press, 255–298.

Lehrer, Adrienne (1985). "Markedness and Antonymy," *Journal of Linguistics* 21, 397–429.

Lehrer, Adrienne and Lehrer, Keith (1982). "Antonymy," *Linguistics and Philosophy* 5, 483–501.

Schroeder, Mark (2008). "How Expressivists Can and Should Solve Their Problem with Negation," *Noûs* 42, 573–599.

Unwin, Nicholas (1999). "Quasi-Realism, Negation and the Frege-Geach Problem," *The Philosophical Quarterly* 49, 337–352.

Unwin, Nicholas (2001). "Norms and Negation: A Problem for Gibbard's Logic," *The Philosophical Quarterly* 51, 60–75.