Published in *The Philosophical Quarterly* 63 (253),
pp. 633-659, doi: 10.1111/1467-9213.12073

Compositional Semantics for Expressivists

By Arvid Båve, the Department of Philosophy,

Linguistics and Theory of Science, Gothenburg University

ABSTRACT: I here propose a hitherto unnoticed possibility of solving embedding problems for noncognitivist expressivists in metaethics by appeal to Conceptual Role Semantics. I show that claims from the latter as to what constitutes various concepts can be used to define functions from states expressed by atomic sentences to states expressed by complex sentences, thereby allowing an expressivist semantics that satisfies a rather strict compositionality constraint (as well as a further, substantial explanatory constraint). The proposal can be coupled with several different types of concept individuation claim (e.g., normative or causal-functional), and is shown to pave the way to novel accounts for, e.g., negation.

Introduction

Embedding problems for noncognitivist expressivists in metaethics are roughly problems of explaining what the meaning of logical compounds containing normative predicates are, given that the meanings of those predicates consist in their expressing certain conative states. Although embedding problems have troubled expressivists (for short) for decades, and although many have seen a tight affiliation between expressivism and Conceptual Role Semantics (CRS), a possibility of solving the former by drawing resources from the latter has so far gone unnoticed. In this paper, I will present a *form* of solution of this kind, which is neutral vis-à-vis specific types of CRS, and can thus be filled in in accordance with one's favoured individuations of content or

meaning. In a nutshell, I present ways to transform *concept individuation claims* within CRS into *definitions of functions from states expressed by atomic sentences to states expressed by complex sentences*.

I begin, in section 1, by presenting expressivism and a general compositionality constraint, as well as a further, substantial "explanatory constraint". In section 2, I use the example of conditionals to present a form of semantics for expressivists meeting the compositionality constraint. Next, I compare my solution to others in the literature (section 3), as well as to CRS generally (section 4). Finally, I make a few tentative suggestions for dealing with negation (section 5). An appendix deals with a technical matter of obtaining meaning-to-meaning functions from the proposed state-to-state functions.

I. Adequacy conditions on expressivist semantics

Although there are several worries about expressivists' prospects of giving an adequate semantics, many of them are usually thought to be ones that would be put to rest if a reasonably demanding *compositionality constraint* could be shown satisfied by an expressivist semantics. An "expressivist semantics" is here a semantics for a language or language fragment, i.e., a theory saying, for each expression therein, what its meaning consists in, which is compatible with expressivism about simple normative sentences. The latter, I will define as a claim about a specific meaning-property (that of "Stealing is wrong"), but it should be clear that this is merely an example (for the purposes of this paper, I could even have exemplified with the meaning of "John ought to pray").

(Expressivism) The property of meaning that stealing is wrong is (constituted by) the property of being conventionally used to express disapproval of stealing (and so on).

I will for brevity speak as if the property of meaning that stealing is wrong is the meaning of "Stealing is wrong". This is of course not strictly true: the meaning of "Stealing is wrong" is that stealing is wrong and the property of meaning that stealing is wrong is something else, but this point will not be relevant for the claims I will be making. I will also for brevity speak of sentences' *expressing* states, rather than *being conventionally used to express* them, which, in my opinion, would have been more accurate.

Further, I choose to formulate Expressivism by identifying the meaning-property of an *example* of a subject-predicate sentence featuring "wrong", and affixing "and so on", instead of saying in general what the meaning of such sentences featuring "wrong" consists in. This is mainly because it is controversial how exactly to generalise the particular fact. A generalisation would arguably identify the property of meaning *wrong*, i.e., the meaning-property of the predicate "wrong", but I wish to remain open as to how this generalisation should go. (What is controversial, I think, is mainly whether the meaning-property should be taken as a property of expressing *disapproval of the referent of the subject term* or, rather, of expressing *disapproval with the (singular) content expressed by the subject term*. But since this paper concerns the question of how the meanings of sentences help determine the meanings of compounds containing them, and not how the meanings of atomic sentences are determined by the meanings of their parts, I think we can safely let this matter be.)

A final point, related to the foregoing one, is that it might be thought odd, at the very least, that my definition of Expressivism itself seems to violate some principle of compositionality. To wit, the definition seems to entail that the meaning of the sentence "Stealing is wrong" is constituted by its being conventionally used to express a certain state. The meanings of the primitive parts may of course be constituted by their being conventionally used to express

something, but the whole sentence cannot, since if there were *a* convention to the effect that the sentence be so used, the meaning of each sentence would have to be learnt separately and so English would be unlearnable.[1] But the last part of this objection shows how to respond to it: "conventionally" need not be interpreted as, "there is a convention to the effect that". To make (Expressivism) come out compatible with the relevant principle of compositionality, then, we must interpret this adverb rather as, "there are conventions with the (joint) effect that".

Now, the compositionality constraint I am about to state assumes, as is commonplace, that on an adequate semantics, the meanings of complex expressions come out as *functions* of the meanings of their immediate constituents. Of course, for any pair of objects, there is trivially a function from the one to the other. So it is not enough, to give a compositional semantics for a language simply to associate meanings with all of its expressions and then note that it follows that there is a function taking the meanings of the primitive expressions to the various complex expressions containing them. Rather, we must ensure that there are *definable* functions that can be seen to have the right values for each set of arguments, and which are also subject to certain further constraints (soon to be described). A definable function is one picked out by some identity-sentence of the form, "$f(x)$ = the object y such that …x…y…", where the blanks are filled in with already understood expressions (and similarly for many-place functions).

One important constraint on such definitions is that the definition not contain different clauses for different types of argument. The different types will, if the semantics is expressivistic, be different types of attitude or mental state, e.g., belief as opposed to desire, disapproval, or

[1] Cf. D. Davidson, "Theories of Meaning and Learnable Languages", in Y. Bar-Hillel (ed.), *Proceedings of the International Congress for Logic, Methodology, and Philosophy of Science* (North-Holland, 1965).

plan. For instance, if the function from states expressed by sentences to states expressed by their negations is defined along the lines of, "*neg*(*x*) = the *y* such that if *x* is a belief, then *F*(*xy*) and if *x* is a desire, then *G*(*xy*)", i.e., if the function has different types of values depending on the type of argument, the semantics would be inadequate. It would in effect take negation to mean different things depending on the nature of the negated sentence, though it would *mask* this fact by associating it with a single function. Let us call this the *uniformity constraint*.

What will be the arguments and values of these functions depends on the semantics. A species of truth-theoretic semantics might take the meaning of a particular name to be the property of referring to a particular individual and the meaning of a predicate to be the property of having a particular extension. Expressivists usually take the meanings of a sentence to be its property of expressing (or: being conventionally used to express) a given mental state and the meanings of subsentential expressions to be, roughly speaking, "contributions" to the sentences' having these properties. Let us say that the properties of expressions that a semantic theory T identifies with their meanings are their T-type properties. The T-type property of a particular name *n* is thus the property that theory T says is the meaning of *n*. This property belongs to the type of properties such that T says that all names have meanings that are properties of this type. For a given theory T, the T-type properties of names and the T-type properties of predicates may of course (arguably, *should*) be of different types (e.g., having an extension contra having a referent). Now that the notion of a T-type property has been explained, we can state the Compositionality Constraint as follows:

(CC) A semantic theory T for a language L is adequate only if it contains the means for defining, for each well-formed expression *e* in L, a function satisfying the uniformity constraint taking T-type properties of the immediate constituents of *e* to the T-type property of *e*.

An instance of this requirement is that each binary connective is associated with an adequately defined function from pairs of T-type properties of sentences to a T-type property of the complex sentence immediately embedding these sentences. I have here left out many intricate questions about "modes of composition" and "unarticulated constituents" which, some might say, must be arguments of the relevant functions, and much else. But I think it best to address such niceties only to the extent that they become relevant for the discussion at hand.

I will leave somewhat vague the phrase "contain the means for". The point is that we do not wish to disqualify a semantics only because it does not explicitly state the function definitions, as long as there is a way of providing them, given what the semantics says. Finally, note that (CC) is merely a necessary condition for adequacy. Clearly, this constraint is far from sufficient. There are plainly absurd semantic theories that satisfy (CC), e.g., the "theory" on which meaning is spelling (for it is trivial to define functions from the spelling of primitive expressions to the spelling of complex expressions containing them). I will address the question of what more substantial constraint should be set on semantics toward the end of this section.

Again, I believe that many of the more formal requirements that have been (more or less explicitly) assumed in discussions about expressivism and embeddings are captured by (CC). For instance, although the famous "Frege-Geach argument" has been interpreted several ways, it is often taken to have a solution if and only if one can give a compositional semantics consistent with Expressivism.[2] Of course, it cannot be true that if we can find a semantics satisfying (CC), then the Frege-Geach problem (and others) are solved, since, as we have seen, there are

[2] See, for instance, M. Schroeder, "What is the Frege-Geach Problem?", Philosophy Compass 3/4 (2008), pp. 703-720, at pp. 708 and 717f.

obviously absurd semantic theories that satisfy (CC). So, there must be some further, tacitly presupposed, substantial constraint that must be satisfied jointly with (CC). (And, indeed, I will take the substantial constraint to be described below, in conjunction with (CC), to be both necessary and *sufficient* for giving a correct semantics).

Though it would not be worthwhile examining how (CC) relates to each significant worry about expressivists' prospects of giving a workable semantics, we might at least briefly consider how it relates to three important constraints recently proposed by Neil Sinclair in his "Moral Expressivism and Sentential Negation" (henceforth, "Negation").[3] Firstly, (CC) plausibly covers his "Fregean condition", according to which a semantics "must explain how the meaning of sentences, both simple and complex, remains constant across negated and unnegated contexts" (as Sinclair says, this condition is really just a "particular instance of the general idea of compositionality" (*ibid.*, p. 387). (By saying that (CC) "covers" this condition, I mean that any theory satisfying (CC) *ipso facto* satisfies the condition.) I am not really happy with its formulation, however, since the fact that the meaning of a sentence "remains constant across negated and unnegated contexts" should hardly be *explained*—rather, the condition, in my view, should be that the semantics is guaranteed to be consistent with this fact. Also, we will see below that it is uncontroversial what exactly this condition requires of a semantics (more precisely, I argue there that a more specific constraint that Sinclair takes to follow from this condition is too restrictive and does not in fact follow).

His "Generality condition", further, requires that "an account of negation must be generalizable regardless of the topic of the sentence embedded" (*ibid.*–cf. also Matthew

[3] Philosophical Studies 152 (2011), pp. 385-411.

Chrisman's "Expressivism, Inferentialism and the Theory of Meaning"[4]). What he means is just that an account of negation that makes sense for negations of normative sentences must also make sense for non-normative sentences (and *vice versa*). Even without explaining exactly what "making sense" amounts to, there are clear examples of semantic accounts of negation that violate this principle. Since (CC) is already general in that it requires definitions of functions corresponding to embedding operators independently of the "topic" of the embedded sentences, it seems that this condition, too, is covered by (CC).

Still, it is important to note that even if (CC) is general in this way, it might still turn out that a function definition that makes sense for one type of argument (one type of mental state, say) does not make sense for a different kind of argument (perhaps due to a presupposition failure relating to the article "the" in the function definition). So, it must be ensured that the functions defined have values for both the kind of state expressed by normative sentences as arguments, as well as for the kind of state expressed by descriptive sentences. Note also that the uniformity constraint above is necessary to avoid trivialising Sinclair's Generality condition. For an account of negation that makes sense for normative sentences but not descriptive ones could easily transformed into one that makes sense for both (by the kind of logical trick mentioned above, of conditionalising on the kind of state expressed).

Finally, Sinclair's "Semantic condition" requires that an account of negation explain why sentences are inconsistent with their negations. Quite generally, it is often assumed that *logical facts* about embedding operators or connectives, i.e., facts about validity, consequence, and inconsistency, must be explicable on the basis of the semantics. I am not sure the various kinds of

[4] In M. Brady (ed.), *New Waves in Metaethics* (New York: Palgrave Macmillan, 2010), pp. 103-125.

semantic theories proposed below satisfy this condition. On the other hand, it is not certain that this semantic condition is a real adequacy condition on semantic theories at all.[5]

As we have seen, clearly inadequate ("crazy") theories can easily be shown to satisfy (CC). Merely showing that the kind of semantics I will propose below satisfies (CC) would thus not be very interesting. Showing that it also helps explain logical facts would be non-trivial, but since I am sceptical of this constraint, and since it is not certain that the proposals I present below satisfy it, I want instead to assume another substantial constraint, which I will call the *explanatory constraint*. This is a constraint Paul Horwich has developed in several works[6] and the satisfaction of which he in fact also takes to be *sufficient* for a semantics to be adequate. In brief, it says that a meaning theory for a language is adequate if and only if, for each expression *e* in the language, the theory's claim as to what constitutes the meaning of *e* can explain, together with auxiliary claims, every fact about the use of *e*. These facts include particular ones, like the fact that someone accepts a particular sentence containing *e* on a given occasion, as well as more general ones, e.g., that competent speakers of the language tend to accept sentences containing *e* in such and such circumstances, or accept such and such sentences categorically.

Though I am unaware of any other philosopher having stated this constraint, I think many philosophers implicitly assume something like it. And it seems rather plausible: surely, a theory about the meanings of natural language expressions must enter explanations of facts about language use. It is not, of course, required to explain such facts all by itself, but the constraint allows that other (presumably, mainly psychological yet non-semantic) theories and facts will

[5] In Chapter 3 of his *Reflections on Meaning* (Oxford: Clarendon Press, 2005), Paul Horwich argues that it is not.

[6] See, in particular, his *Meaning* (Oxford: Clarendon Press, 1998) and *Reflections on Meaning*.

also enter these explanations. In fact, it seems that it is much more certain that this is a real constraint on semantics than the constraint of explaining logical facts. However, I will not here argue further in favour of setting this constraint, but simply assume it.

I will also not be arguing, for any particular instance of the kind of proposal below, that it will satisfy the explanatory constraint (which, too, would be far beyond the scope of this paper). But I want to note that they are all similar to proposals that have been taken, for independent reasons, to satisfy it (by Horwich and presumably others who have implicitly been assuming it). Thus, there are independent reasons to think that some such proposal in fact satisfy the constraint. What I aim to make plausible, then, or at least reasonably promising (which is perhaps all we can hope for, given the magnitude of the project) is that there is a kind of semantics that satisfies both (CC) and the explanatory constraint. This of course goes far beyond the trivial task of formulating a semantics that merely satisfies (CC).

It may seem that (CC) is otiose in view of the explanatory constraint, in the sense that any semantics satisfying the latter will satisfy the former, if language is compositional at all. In other words, if the explanatory constraint can be satisfied by a theory not satisfying (CC), then (CC) is simply not a real constraint on meaning-theories. Still, there is a point to showing that the type of semantics proposed below satisfies both, because Horwich himself has actually argued that a strong constraint like (CC) need not be satisfied by an adequate semantics (*Reflections on Meaning*, Chapter 8). By showing that the proposals below satisfy both, we show that Expressivism is not hostage to this controversial view about compositionality.

II. (D)-semantics: the example of conditionals

In this section, I will propose a *form* of compositional, expressivist semantics and give examples of how it may be filled in in accordance with well-known proposals within Conceptual Role

Semantics (understood broadly enough to include also Horwich's "Use Theory of meaning", "inferentialist" or "functionalist" theories, etc.). The functions I will define take mental states expressed by sentences to the mental states expressed by various logical compounds containing the former sentences.

Let us use "$S(s)$" to denote the state expressed by a sentence s, and let us say that a function f *corresponds to* a binary connective c just in case the meaning of c consists in the fact that a sentence "A c B" expresses $f(S(\text{"A"}), S(\text{"B"}))$. So, for instance, a function f corresponds to "and" just in case the meaning of "and" consists in the fact that a sentence "A and B" expresses the state $f(a, b)$, where "A" expresses a and "B" expresses b. In order to define, say, the function corresponding to the indicative conditional, we must produce some definition of the form,

(DC) $\quad \text{CON}(x, y) =$ the state S such that $F(x, y, S)$.

Note that definitions of this form are stipulative, so their truth is not in question. The substantial claims rather concern which such functions, defined in this way, correspond to various natural language connectives (and other embedding operators).

(DC) of course only concerns a particular connective with two places, but it is obvious how to generalise this form for connectives of any arity:

(D) $\quad f(x_1, \ldots, x_n) =$ the state S such that $F(x_1, \ldots, x_n, S)$,

where the arity of F is that of $f + 1$. Let us say, finally, that to give

(1) (D)-form definitions (as I will call them) of state-to-state functions,

(2) statements of what functions, defined by these definitions, correspond to the connectives or other propositional operators of the language,

(3) adding a theory of unembedded normative sentences on the lines of Expressivism

(4) adding a semantics for unembedded descriptive predicates

is to give a *(D)-semantics*.

The question is now what should replace "*F*" in a (D)-form definition. I will not defend any specific proposal here, but rather produce a number of examples, all taken from some form of Conceptual Role Semantics. A major division of candidate (D)-form definitions is that between descriptive and normative, corresponding to descriptive and normative proposals within CRS. For simplicity, we may take CRS to be a theory about what individuates given concepts, where a concept is a constituent of thought contents. (Some self-labelled conceptual role theorists rather take their theory as concerned with linguistic meaning, but the difference is not important for present purposes.) CRS, I will say, contains *concept individuation claims*, and these, we will see, can be transposed into (D)-form definitions of state-to-state functions. Here is an example of a descriptive concept individuation claim, concerning the concept *if*, which follows Christopher Peacocke's so-called *A(C)* form[7]:

(PC) The concept *if* is the unique concept *c* such that to possess *c* is to find *primitively compelling* all transitions from a belief that *A* and a belief that *A c B* to a belief that *B*.

[7] See esp. Chapter 1 of his *A Study of Concepts* (Cambridge, MA.: MIT Press, 1992).

To find something primitively compelling, further, is spelt out in purely descriptive, psychological terms. An example of a *normative* version, following Ralph Wedgwood[8], reads as follows:

(WC) The concept *if* is the unique concept *c* such that to undergo a transition from a belief that *A* and a belief that *A c B* to a belief that *B* is to follow a *basic rule of rationality*.

A basic rule of rationality is one it is rational (a normative notion) to follow and whose rationality is not derived/explained by the rationality of anything else. Obviously, these are but two kinds of CRS-style concept individuation claims. Such claims also vary in which and how many inference rules or transitions they take to be constitutive, and in which conditions must obtain for the constitutive inferences to be rational, or found primitively compelling, etc.

Now, to extract a (D)-form definition that suits Expressivism from a concept individuation claim, we must, first and foremost, do away with the reference to beliefs, since the functions need to be defined also for the type of state taken to be expressed by normative atomic sentences. We must also ensure that the notions used in the CRS concept individuation claim is applicable to non-beliefs as well as beliefs (cf. the Generality Condition above). For instance, if it doesn't make sense to speak of mental transitions involving non-beliefs as primitively compelling, then the (D)-form definition extracted from (PC) violates the Generality Condition.

Secondly, in order to extract a (D)-form definition from a concept individuation claim, we must have a way of transforming the individuating condition on concepts into an individuating condition on states. However, since different concept individuation claims operate with different

[8] Chapter 4 of *The Nature of Normativity* (Oxford: Clarendon Press, 2007).

notions, this transformation may not be uniform. But we will see that there are often obvious transformations for specific concept individuation claims. For example, two (D)-form definitions extracted from (PC) and (WC), respectively, are:

(DPC) CON(x, y) = the state S such that, necessarily, for all z, z is in S iff z finds the transition from S and x to y primitively compelling, and

(DWC) CON(x, y) = the state S such that to undergo a transition from S and x to y is to follow a basic rule of rationality.

What relations these (D)-definitions take to hold between the three items CON(x, y), x, and y are the same as those that their ancestor concept individuation claims take to hold between the belief that if p then q, the belief that p, and the belief that q. Note also that (DPC) and (DWC) are identical to (DC) except "F" has been instantiated with substantial predicates. There might of course be more ways of extracting (D)-form definitions from (PC) and (WC) above, but none, I think, that will result in definitions very different from (DPC) and (DWC). There is also the question whether the notions used in (DPC) and (DWC) can be intelligibly and non-trivially applied to non-beliefs. I will not venture to argue that this is so, however, since these definitions are in any case merely examples of a general strategy of specifying the meanings of logical compounds consistently with Expressivism.

Although I am not out to defend any specific definition of "CON", I would like to give one final example, partly because it uses notions I personally think are the right ones to use in individuating contents and meanings, and partly because it is clear that these notions can be intelligibly and non-trivially applied to any mental state (hence, both beliefs and conative states).

This is because the notions are strictly causal-functional or dispositional. A (D)-form definition of CON phrased in such terms might be:

(DD) CON(x, y) = the state S such that, for all z, z is in S iff z is defeasibly disposed, upon consideration of S, x and y, to undergo the transition from S and x to y.

This places a heavy burden on the notion of a defeasible disposition, of course, but I agree with authors like Georges Rey and Paul Pietroski[9] that such notions can be defined in purely descriptive terms and so as to allow for adequate meaning individuations claims.

Something should be said, also, about the "consideration" involved in (DD). Although we are familiar with the notion of considering a *proposition*, i.e., considering whether p, it may be thought odd to speak of considering a mental state. But I think we can make sense of this simply by taking it as a matter of considering the *content* of the state. Perhaps we also need to qualify the "consideration" as specific to a given attitude, so that the same content can be "considered for belief", "for desire", and presumably also "for disapproval". Considering whether so and so will then naturally be understood as a consideration "for belief", but it seems reasonable that one can also consider a content "for desire", so that, given appropriate preconditions, someone who considers a certain propositional content will come to form a desire with this content. Likewise

[9] See "When Other Things Aren't Equal: Saving Ceteris Paribus Laws from Vacuity", The British Journal for the Philosophy of Science, 46, 1995, pp. 81-110 and G. Rey, "Saving Psychology from Normativism", in McLaughlin, B. and Cohen, J. (eds.), *Contemporary Debates in Philosophy of Mind* (Oxford: Blackwell, 2007), pp. 69-84.

for disapproval, except here the content is not propositional. But this is not the place to defend (DD) or any other specific (D)-form definition at length.

Rather, my point is much more general: the programme of individuating concepts (like *if*) in normative, psychological and/or causal-functional terms goes on and has ample motivation independently of Expressivism. Given how such individuation claims can be transposed into state-to-state function definitions, expressivists' most promising tactic, both with respect to (CC) and to other constraints on semantics, seems to be to follow the developments within CRS and adopt the (D)-form definitions extracted from whichever proposals seems most plausible.

Concerning (CC) specifically, any (D)-semantics will of course automatically satisfy it, since giving a (D)-semantics simply consists in giving definitions of functions that are the semantic values of sentences. Since (CC) is a rather strong compositionality constraint, there is thus a strong case for claiming that any (D)-semantics is adequately compositional.

III. Comparison with other solutions to embedding problems

Simon Blackburn's *higher-order attitude account*[10] (*Spreading the Word*, Ch. 6) and his *commitment semantics*[11] can both be rather straightforwardly formatted into the (D)-form. According to the former, the state expressed by a conditional is the state of disapproving of being in the state expressed by the antecedent while not being in the state expressed by the consequent. Put in the (D)-form, this suggestion would read:

[10] In Chapter 6 of his *Spreading the Word*, Blackburn, S. (Oxford: Clarendon Press, 1984).

[11] "Attitudes and Contents", Ethics, 98 (1988), pp. 501-517 and Section 3.4 of his *Ruling Passions* (Oxford: Clarendon Press, 1998).

(DBC) CON(x, y) = the state S such that, for all z, z is in S iff z disapproves of being in x and not being in y.

This account is widely considered refuted by Mark van Roojen.[12] Suppose "wrong" is used to express disapproval and that "A" expresses S1 and "B" expresses S2. Now, if *modus ponens* is valid, then, on Blackburn's account, the argument form A and "It is wrong to accept S1 while rejecting S2" to B must be valid too, since the second premise here expresses what "If A then B" expresses. But clearly it isn't valid.

But the real problem, in my view, is more general: if "If A then B" and "It is wrong to accept S1 while rejecting S2" come out as expressing the same state, then, given the usual expressivist assumptions about meaning, they come out as synonymous, which they are surely not. So even if a semantics need not explain validity facts (as argued by Horwich—see note 5 above), the higher-order attitude account must still be either false or in need of some very different view about meaning.

Blackburn's "commitment semantics", featuring his famous notion of being "tied to a tree", is more difficult to compare with, for it is rather unclear what the account is supposed to be. In the cited works, he proposes many different, often ambiguous accounts, mainly of disjunctions and conditionals. One passage that may seem to suggest a dispositionalist variety of (D)-semantics is: "Suppose I hold that either John is to blame, or he didn't do the deed. Then I am in a state in which *if* one side is closed off to me, I am to switch to the other—or withdraw the commitment. And this is what I express by saying 'Either John is to blame, or he didn't do the deed'" (*Ruling Passions*, 71), or the more general, "By advancing disjunctions and conditionals

[12] See his "Expressivism and Irrationality", Philosophical Review, 105 (1996), pp. 311-335.

we avow these more complex dispositional states" (*ibid.*, 72). However, there is little by way of full generalisation, and no mention either of compositionality or functions. Still, Blackburn's commitment semantics can certainly be seen as waving toward a dispositionalist version of (D)-semantics. But although the former is recognisably in the spirit of (D)-semantics, the more general form of the latter, in conjunction with the method of transforming concept individuation claims within CRS to (D)-form definitions, makes for a more systematic way of devising new, possible characterisations of the states expressed by logical compounds.

In connection to Blackburn's commitment semantics, I would like to comment also on a restriction on expressivist semantic theories, proposed by Neil Sinclair, which he attributes to Blackburn. This is the restriction that the state expressed by a compound sentence has the states expressed by the subsentences as "functional parts", in a specific sense. This is meant to be stronger than what I have claimed above, i.e., that the states expressed by compounds should be the values of definable functions for the states expressed by the subsentences as arguments. The restriction is meant to rule out such expressivist semantic hypotheses as Schroeder's "dominant attitude" account[13], on which a sentence "Killing is wrong" expresses an attitude of the form $\alpha!\beta!x$ while its negation, "¬(x is M)" expresses an attitude of the form $\alpha!\neg\beta!x$. Here, $\alpha!$ could be the attitude of *being for*, and $\beta!$ could be *blaming for* (so that "Killing is wrong" would express the state of being for blaming for killing, and its negation expresses the state of being for not blaming for killing). Sinclair holds that this violates the relevant constraint because, he says, the latter attitude does not contain the former as a functional part, because the "whole" of the former is not "reflected, invoked, inferentially embedded, involved or otherwise 'in the offing' in the

[13] In *Being For: Evaluating the Semantic Program of Expressivism* (Oxford: Clarendon Press, 2008).

mental state expressed by [the negation] that embeds 'p'" ("Negation", p. 396). This, in turn, is unacceptable, he says, because it violates the "Fregean Condition", saying that the meaning of a sentence must remain constant across embedded and unembedded contexts (see above).

Note that nothing of what we have said rules out the (D)-form definition corresponding to Schroeder's proposal, which would read,

(DS) $f(x)$ = the state which is the same general attitude type as x but which has as its content/object the negation of the content/object of x,

where f is supposed to correspond to negation. But it is hard to gather from the general motivations for compositional semantics why any such stricter condition should be set on a semantics. Sinclair's claim that a definition like (DS) would entail that a sentence would differ in meaning depending on whether it is free-standing or negated seems to me in need of further support, for it has not been made clear why we could not simply assert that both occurrences of the sentence means the same because their meaning consists in expressing a specific state *when assertorically uttered*. It is just that when a sentence occurs embedded under negation, *it* is not assertorically uttered (but the negative sentence might be).

Furthermore, we could easily make (DS) square with Sinclair's criterion simply by redefining "in the offing", e.g., as follows: the state expressed by a sentence s is "in the offing" relative to (for instance) $\neg s =_{df}$ there is a definable function f satisfying constraints C and taking any state expressible by a sentence to the state expressed by its negation, and (hence) taking the state expressed by s to that expressed by $\neg s$ (where C will include the uniformity constraint

above, and perhaps more). It is not clear to me why this definition of a state being "in the offing" should be found unacceptable from the perspective of giving a compositional semantics.

A possibly even more serious problem for Sinclair's restriction, interpreted literally, is the following: it seems reasonable that if one state *A* is literally part of another, *B*, it is not possible for something to be in state *B* without being in state *A*. But it is obviously possible to be in the state expressed by a disjunction or conditional without being in any of the states expressed by its subsentences. If this condition holds, then, then the restriction is clearly wrong. Whether this condition indeed holds is difficult to tell, since it is not obvious what it means to say that a state is part of another. Sinclair claims that one state may be a functional part of another and yet be unactual even when although the "more complex" state is actual, but this is to say what this relation must be like, if the restriction is reasonable, rather than showing that it is that way.

The following consideration, further, might indicate that the condition indeed holds: the most obvious examples of states being parts of states we can think of are cases in which a state of an object consists in several things "in" the object being a certain way, where a subset of those things' being that same way is part of the bigger state of all those things being that way. If a room contains two light bulbs, then the state (of the room) of both light bulbs being alight might, for instance, contain as a proper part the state (of the room) of the smaller light bulb being alight. This had better not be a good model of parthood among states, lest Sinclair's criterion be clearly at fault. I am not sure this is an example of parthood among states at all, but this, it seems to me, only goes to sustain my initial worry that the notion of a state-part might be too obscure to be workable.

Finally, it seems to me that what is driving the intuitions about parthood among states is the fact that the descriptions used to denote the various attitudes themselves have certain mereological relationships. For instance, the description "$\alpha!\neg\beta!x$" does not contain "$\alpha!\beta!x$" as a

part. However, obviously "$f(\alpha!\beta!x)$" *does* have "$\alpha!\beta!x$" as a part. Why is this not enough to show that (DS) satisfies the criterion? Presumably, the answer would be that the structure of "$f(\alpha!\beta!x)$" is misleading as to the mereological structure of what it refers to, and that this is revealed once we look at the reading of (DS). I agree that "$f(\alpha!\beta!x)$" is probably not very revealing as to the mereological structure of what it refers to, but I don't think "$\alpha!\beta!x$" is any different on this score. Sinclair, I suspect, presupposes that the attitude $\alpha!\beta!x$ contains its object, i.e., $\beta!x$ as a part, but I see no reason to accept this view of attitudes. (Oddly, Sinclair considers the objection that the belief that ¬(x is F) does not seem to contain the belief that x is F as a functional part in the relevant sense, but replies that "the problem can be avoided so long as the belief that ¬(x is F) can be understood as some function of the belief that x is F" ("Negation", note 19). But this seems to be to give up the stricter restriction in favour of the more relaxed condition I am recommending!)

Although Blackburn's higher-order attitude account, as we have seen, faces a devastating objection, I agree fully with Mark Schroeder's assessment that this theory at least has the virtue of actually saying what the relevant type of mental state is supposed to be (it is "constructive", in Schroeder's terminology).[14] This, says Schroeder, is in contrast to theories like that of Allan Gibbard,[15] which use undefined notions of "ruling out", "disagreement" and "allowing", and identifies the states expressed by compound sentences by reference to the sets of states which

[14] *Noncognitivism in Ethics* (London: Routledge, 2010), at p. 116.

[15] See his Chapter 3 of his *Thinking How To Live* (Cambridge, MA.: Harvard University Press, 2003) and his "Reply to Blackburn, Carson, Hill, and Railton", Philosophy and Phenomenological Research, 52, pp. 969-980, at pp. 972f.

they "disagree with" or "allow" (allowing us to use ordinary set-theoretic operations like union, intersection, etc., to compositionally characterise the states expressed by compounds).

The problem with Gibbard's proposal, according to Schroeder, is that it only explains what the states expressed by compounds must be like, if they exist, but the account has no way of guaranteeing that they exist (*Noncognitivism in Ethics*, p. 132). Also, although we must surely agree that Gibbard's claims using "disagreement" are true in *some* sense of "disagreement", it may be that this holds only in a sense in which the claims are useless to, or incompatible with, Expressivism. This would be the case, for instance, if "*x* disagrees with *y*" can only be appropriately defined as, "the contents of *x* and *y* cannot both be true". Schroeder concludes that Gibbard's semantics is little more than an empty and unexplanatory formalism.[16] I agree with these criticisms, which is why I have here tried instead to devise proposals for a *constructive* semantics, one that actually says what states are expressed by complex sentences. While I have not opted for any one type of definition of CON above, it should be clear that there are many options for (D)-definitions which are both constructive and avoid crucial undefined logical terms such as "disagree", "inconsistent", etc.

One might be tempted to think, however, that although perhaps the *descriptive* (D)-form definitions above really give us constructive accounts of the states expressed by compounds, the *normative* ones do not, and thus conclude that we can escape Schroeder's objection against Gibbard only if we adopt descriptive (D)-form definitions. However, to object this way to normative (D)-form definitions is to ignore the individuating intent of the definition. It has been argued, independently of Expressivism, that content and/or meaning can only be individuated in normative terms; that is, the only way of saying what the concept *if* is, is by saying what

[16] *Being For*, Section 3.5 and *Noncognitivism in Ethics*, Section 7.3.

normative principles govern it (e.g., by Ralph Wedgwood, in *The Nature of Normativity*). An adherent of a normative (D)-semantics will naturally say the same about the states that are the values of the functions there defined. Thus, they will say that the (D)-form definitions *do* say what the states are, contrary to the objection.

Could the same reply be given on behalf of Gibbard's semantics? That is, could Gibbard say that the disagreement patterns of various states are what make them what they are, and that no more "canonical" or "direct" description is to be had? Perhaps, but we cannot begin to answer this question until we know what "disagreement" is. It is the lack of such a definition that is the most serious problem with Gibbard's semantics, and which makes it hostage to the possibility that "disagreement" can be defined only in such a way as to make the semantics unsuitable for Expressivism. No such problem arises for the (D)-form definitions I have produced, since they do not trade in any undefined logical notion like "disagreement".

Suppose, though, that we have a definition of "disagreement" that makes it possible to say that conative states disagree. There is still a different problem with Gibbard's semantics, a problem it shares with possible-world semantics (which it mimics). To wit, this semantics does not distinguish between logically equivalent, but non-synonymous sentences. The main benefit of CRS is precisely its ability to provide *fine-grained* accounts of content, distinguishing logically equivalent but distinct contents. So even with a definition of "disagreement" congenial to Expressivism, there is reason to be sceptical of Gibbard's semantics, and this goes for any solution to embedding problems mimicking possible-world semantics. (D)-semantics faces no such objection. (D)-semantics is thus superior to Gibbard-style semantics in at least three respects: definitiveness, constructiveness, and grain.

Paul Horwich's expressivism[17] also comes close to certain proposals I made above. To wit, his "use-theoretic" accounts of the meanings of connectives are purely causal-dispositional, like (DD), and avoid the three aforementioned problems with Gibbard's account. However, Horwich expressly rejects the need to respect any such constraint as (CC), and argues on the contrary that compositionality does not place any non-trivial constraint on meaning-theories (*Reflections on Meaning*, Chapter 8). Also, he does not uniformly explain the meaning of connectives in terms of states expressed. Moreover, his "expressivism" is special in that he takes normative sentences to primarily express beliefs just like descriptive sentences, and is a non-cognitivist only in that he takes the contents of these beliefs to be individuated by the fact that they stand in certain relations to non-cognitive states (in fact, aside from the issue about whether meanings can be descriptively individuated, Horwich's view seems to come closest to that of Ralph Wedgwood's *The Nature of Normativity*, and their differences seem mainly verbal). (D)-semantics is thus clearly different from what Horwich proposes, since (1) it uniformly explains the meanings of expressions in terms of their contributing to sentences' expressing mental states, (2) it allows and is meant to accommodate the view that normative sentences do not express beliefs, but some type of conative state, and (3) it is designed to (and in fact does) satisfy (CC). The second point is perhaps the most important one, for on Horwich's account, there are normative contents, contrary to Expressivism, wherefore embedding problems arguably do not arise for his account at all! (More on this in section the next section.)

[17] See his "The Frege-Geach Point", Philosophical Issues, 15, pp. 79-93 and Chapter 9 of his *Truth-Meaning-Reality* (Oxford: Clarendon Press, 2010).

Finally, and for completeness, we should note also that (D)-semantics does not share the problematic consequences of Terry Horgan and Mark Timmons's "non-constructive" theory[18], noted by Schroeder.[19]

IV. Comparison with CRS

We have already seen that the commitments and problems of Expressivism and CRS are not the same. For expressivists need the functions in the (D)-form definitions to take conative states as arguments, whereas CRS does not by itself have any analogous commitment relating to their concept individuation claims.

Another important difference, I will now argue, is that while CRS may take connectives to correspond to functions from propositions to propositions (or from propositional attitudes to propositional attitudes), (D)-semantics may not. For if what is expressed by a complex sentence is a proposition, it would have to be complex. But a semantics incorporating Expressivism cannot provide any propositions that can be the subpropositions of the complex propositions allegedly expressed by compounds containing normative predicates. This is not to say that the conative states expressed by simple normative sentences cannot be propositional attitudes. But even if they are, the propositions they have as their contents cannot be the subpropositions of the complex propositions. To see this, suppose "*x* is wrong" is taken to express the desire that *x* does not occur (a propositional attitude). A sentence of the form, "If *x* is wrong, then *p*" clearly cannot express the proposition that if *x* does not occur then *p*. For this would rather be the proposition expressed

[18] "Cognitivist Expressivism", in T. Horgan and M. Timmons (eds.), *Metaethics after Moore* (Oxford University Press, 2006), pp. 255-298.

[19] See Schroeder's *Noncognitivism in Ethics*, Chapter 7.

by the non-normative sentence, "If *x* does not occur then *p*". And it is hard to see how else the complex proposition expressed by a complex sentence could be determined than by somehow taking its subpropositions from what is expressed by the embedded subsentences.

For essentially the same reason, expressivists cannot take the values of the functions to be propositional attitudes. For the propositional attitude supposedly expressed by a complex sentence would presumably have to have a complex proposition as its content. But we have already seen that, given Expressivism, (D)-semantics provides no proposition that can be the subproposition of the complex proposition of the attitude expressed by a complex sentence containing a normative predicate. And it is difficult to see how else the propositional attitude supposedly expressed by a complex sentence could be determined from what is expressed by its subsentences (whether propositions or propositional attitudes).

Since this is a rather subtle point, it is important to stress that none of the above entails that Expressivism (and hence, (D)-semantics) is incompatible with holding that complex sentences containing normative predicates express propositions or propositional attitudes. What Expressivism rules out, rather, is that the correct *semantics* couples complex normative sentences with propositions or propositional attitudes. In other words, expressivists must reject the claim that the meanings of complex normative sentences consist in their expressing propositions or propositional attitudes. But even if what constitutes the meaning of a sentence is not its having a given property, it does not follow that it doesn't have that property. Thus, for all I have said, Expressivism is compatible with saying that complex sentences express propositions and/or propositional attitudes. Expressivists often say that normative sentences might express

propositions or beliefs "in a minimal sense". While it is not always clear what this amounts to,[20] we might here at least conclude that this is true in the sense relating to meaning-constitution just explained. In any case, I submit that we have here found another important commitment of (D)-semantics, which it does not share with CRS.

One may think our above conclusion conflicts with the idea that complex descriptive sentences express (complex) beliefs. For, surely, a good semantics should not take complex normative and complex descriptive sentences (of the same form) express different kinds of states (cf. the uniformity constraint of Section 1). The reply is, as above, that even if we agree that complex descriptive sentences express beliefs, we need not therefore take the *semantics* to associate them with beliefs. Thus, the semantics can be perfectly uniform after all: all complex sentences–purely descriptive, purely normative, or mixed–are taken to express the kind of state specified by the (D)-form definitions, and the claim that they also express propositions or propositional attitudes is not part of the semantics.

Thus, expressivists are committed to rejecting any semantics that couples propositions or propositional attitudes with purely descriptive, complex sentences, i.e., committed to denying that the meaning of complex sentences consists in their expressing propositions or propositional attitudes. This has a special import for the account of negation. To wit, the view, common in CRS, that negation must be explained in terms of a special *propositional attitude of rejection*, is unavailable to (D)-semantics. (However, if the state of rejection is taken instead as a relation to a *state*, rather than to a proposition, then it is available.) We turn now to considering some proposals of (D)-form definitions relating to negation.

---

[20] For an overview of proposed explications, see Neil Sinclair's "Recent Work in Expressivism", Analysis, 69 (2009), pp. 136-147, at pp. 139ff.

V. Negation

Negation seems more difficult to individuate in informative terms than the conditional, and no very satisfactory account has been given within CRS. Still, I will give some examples of such proposals, if only to illustrate the constraints and problems involved. Christopher Peacocke[21] has proposed the following concept individuation of negation: negation is the unique concept *C* such that *Cp* is the weakest proposition inconsistent with *p*. The problem with this account is that it uses "inconsistent" but does not define it (cf. Gibbard's "disagreement"). This means that the truth of this claim makes it no more certain that negation can be individuated in CRS-friendly terms (also, one wonders what happened to Peacocke's possession conditions that were supposed to individuate concepts). Similarly, even if Peacocke's claim is easily transposed into a (D)-form definition, we have as yet no reason to believe that the claim that *x* and NEG(*x*) are inconsistent is true in a sense of "inconsistent" in which the claim could be used in a compositional semantics compatible with Expressivism. As with Gibbard's account, there is no guarantee that there is any other way of explaining the crucial notion than in terms of truth, which would be incompatible with Expressivism (on the assumption that the relevant conative state is not truth-apt).

Stephen Barker's causal-functional account of negation[22] may seem more promising. His account is rather complex and theoretically involved, however, and it is also not a case of concept

[21] "Proof and truth", in J. Haldane and C. Wright (eds.), *Reality: Representation and Projection* (Oxford University Press, 1993), pp. 165-190, at p. 176.

[22] See §23 of his *Global Expressivism: Language Agency without Semantics, Reality without Metaphysics*, published online in 2007 at http://eprints.nottingham.ac.uk/696/.

individuation (which he thinks cannot be had), so for simplicity I will just use the crucial notions Barker uses in a (D)-form definition (that he might well not accept):

(DBN) NEG(x) = the state S such that, for all y, y is in S iff y is in a state S' which grounds y's being constrained not to token x,

where NEG corresponds, in the sense defined above, to "not". An important constraint on expressivist accounts of negation is that the state expressed by negative sentences come out as distinct from both *indifference* (not having an opinion on the matter) and *agnosticism* (being determined/committed to not have an opinion—cf. Sinclair's "Negation", p. 388). Although Barker takes pains to show that his account properly distinguishes what is expressed by negated sentences from indifference, there is no mention of agnosticism. But the real problem with this account, it seems to me, is rather that the crucial notions, especially "being constrained not to …", remain obscure.

The just-mentioned problem, of distinguishing within CRS belief in a negation from mere indifference or agnosticism, has proved very difficult. A recent account of negation by Neil Sinclair seems to fall prey to this difficulty. He takes the state expressed by, e.g., "Killing is wrong" to be a "policy" (such as the policy of blaming anyone one thinks has killed someone), and the state expressed by its negation to be the state one is in just in case one "adopt[s] the policy that might be characterized as 'the way to respond to the object of evaluation is some way other than that given by the unnegated attitude'" ("Negation", p. 399). In another of his formulations, "To reject […] a policy is to have the policy of responding to the world in some way other than that given by the original policy." (*ibid.*, p. 407). Here, "rejection" of a state is by definition the state expressed by the negation of the sentence expressing the state rejected. He

also emphasises that rejecting a policy is not a higher-order attitude directed toward the original policy, but rather a policy directed to the same subject-matter as the original policy (*ibid.*, p. 400), where the "subject-matter" seems to involve (only?) the act-type. However, even granting that these *explanantia* of rejection will be satisfied by someone who is in the state expressed by a negative sentence, I will argue that, for all Sinclair has said, they might also be satisfied by someone who is *agnostic* about the matter. Or, more specifically, the only ways of avoiding this consequence are unacceptable to expressivists for independent reasons.

If what is expressed by "Killing is wrong" is a policy of responding by a certain reaction R to killing situations (to use another of Sinclair's phrases), then the person who is agnostic relative to this state is determined/committed to not have this policy. Agnosticism is thus a higher-order attitude, directed toward another state, namely, the policy in question (or perhaps the state of *having* this policy). Sinclair claims that rejection is distinguished from agnosticism because rejection is *not* a higher-order attitude. But consider Sinclair's favoured description of rejection, "the policy of reacting to killing situations some other way than by policy P" (where "P" refers to the original policy). This description is ambiguous, in that on one interpretation, it contains an expression referring to the original policy within the scope of the first occurrence of "policy", and on another interpretation, the expression referring to the original policy comes outside the scope of "policy". On the first interpretation, Sinclair's own theory entails that the state expressed by negative sentences is a higher-order attitude directed toward the policy expressed by the unnegated sentence. If you are uneasy about this talk of "policy" having a *scope*, consider the intuitive difference between having the policy of Φ-ing *F*s, on the one hand, and the policy of Φ-ing objects, which happen to be *F*s. The predicate "*F*" does not enter into the scope of "policy" in the latter case, whereas it does (at least on one reading) in the former. But on a reading on which

an expression referring to a policy comes within the scope of "policy", we must conclude that the policy referred to by the more complex expression is a higher-order policy directed toward the policy referred to by the embedded "policy-designator". The greater complexity of the content of this policy (as compared, e.g., with Schroeder's dominant attitudes) can hardly make any difference on this score.

One can also interpret the description, "the policy of reacting to killing situations some other way than by policy P", in such a way that "P" comes outside the scope of the first "policy". To make this come out clearer, we could rephrase the description thus: "the policy of (responding by reaction R to killing situations), where R is distinct from the reaction featuring in the original policy P", stipulating that the relative clause comes outside the scope of the policy (this scope is, in addition, indicated by parentheses). However, on this interpretation, Sinclair's theory has a false presupposition (or consequence). For since there are many reactions distinct from that figuring in the original policy, there are also many *policies* satisfying this description. Thus, depending on how the theory is taken, either it involves a false presupposition (of uniqueness) or it entails that "Killing is not wrong" expresses many different states, which is unacceptable to expressivists. So, it seems, only by manoeuvres that are independently objectionable can Sinclair's account of rejection be saved from the objection that it fails to rule out agnosticism as a way to reject something.

Here are three possible responses on Sinclair's behalf, responding to which I think will make the problem come out clearer. First, one might think, he could respond that to reject, on his view, is simply to have one of these many policies that are distinct from the original one. But this would obviously not work, since, for instance, I might have the policy of blaming murderers but also the policy of dissuading my children of befriending murderers, which is a distinct policy, but hardly a policy being in which entails that one rejects the original one.

Secondly, one might think that Sinclair could respond that it is not distinctness, but *opposition*, that defines rejection. But although Sinclair makes this claim repeatedly, it is not in itself an informative description of rejection, but something that needs to be defined (and I am arguing that Sinclair's attempts to define it fail). If we could simply appeal to a relation of opposition, we could just say that negative sentences express states that *oppose* the states expressed by the unnegated sentences, but this is clearly not informative enough.

Thirdly, one might argue that even if Sinclair's attempt to distinguish rejection from agnosticism (i.e., by arguing that the former is not, but the latter is, a higher-order attitude) fails, he could still argue that agnosticism is not a policy, but a "commitment" or "determination". However, neither of these three terms are defined, and thus, for all that has been said, policies to do something might as well be determinations or commitments to do it, and thus rejection has still not been properly distinguished from agnosticism.

In conclusion, I think Sinclair's account of negation is ambiguous and that once disambiguated, all of the precisifications turn out to be unacceptable. On one precisification, rejection comes out as a higher-order attitude, wherefore Sinclair's attempt to distinguish it from agnosticism fails on that reading. On a second one, on which it does not come out as a higher-order attitude, it carries a false presupposition of unicity (or, alternatively, it entails that each negative sentence expressed several states). On a third interpretation, which would avoid both problems identified above, the account would characterise certain states as states of rejection, which are clearly not (e.g., the policy of dissuading one's children from befriending people who have killed would come out as sufficient for rejecting the policy of blaming people who have killed). On yet another interpretation, Sinclair's *definiens* of rejection involves the undefined relation of opposing, which makes the definition uninformative. Finally, Sinclair might claim that rejection is different from agnosticism in that it is a policy rather than a determination or

commitment, but, as with opposing, these notions have not been defined so as to show how the states are supposed to differ. I conclude that his account does not properly distinguish rejection from agnosticism.

I also think it is unclear whether Sinclair's account of negation really meets the "Generality Condition", i.e., that negation means the same whether occurring in a normative or non-normative sentence. Sinclair assures us that it is:

> […] the belief that grass is green represents the world as being such that grass is green. To reject a belief is to reject a representation of the world; it is to think that the correct way to represent the world is some way other than the representation given by the rejected belief. Once again, this rejection is a first order commitment rather than second-order commitment directed at the original commitment. The rejection of the belief about the color of grass is itself a belief about the color of grass: it is the belief that the world is one in which grass is some color other than green.[23]

It is rather opaque exactly how the rejection of a belief and the rejection of a policy are supposed to be characterised by the same function. I can see only one way, namely, if the belief that *p* is *identified* with a policy: that of representing the world as being such that *p* (this is vaguely suggested by the passage quoted). Now, the account of negated moral sentences may seem applicable: they express a policy of responding differently. In the case of descriptive sentences, then, negations express the policy of representing the world as being some other way than the way they are represented by the belief expressed by the unnegated sentence. But to represent the

[23] "Negation", p. 401.

world as being such that snow is white is, at least on one reading, to represent it as being different from the way it is represented when represented as being such that grass is green. Presumably, the intended meaning is that the one way of representing should rather be *opposed to* (or *inconsistent with*, etc.) the other, which is not the case with the belief that grass is green and the belief that snow is white. But, again, since opposition, inconsistency, etc., are no better understood than negation or rejection, this is no proper solution.

The difficulty of ruling out indifference and agnosticism re-emerges in a rather serious way when we contemplate the possibility of transforming CRS accounts of negation to (D)-form definitions. To wit, there are two seemingly promising ways of individuating the concept *not* within CRS that, it seems, cannot be immediately transposed to (D)-form definitions, precisely because the latter would fail to distinguish the state expressed by a negation from indifference and agnosticism.

A very widely accepted claim about negative beliefs is that one *ought not* believe a proposition and its negation. More controversially, one might hold that one *cannot* believe both a proposition and its negation (at least not while simultaneously *considering* both beliefs).[24] One might now think that one of these properties of negative beliefs may be the individuating feature of the concept *not*. For instance, one might propose either:

(NN) The concept *not* is the unique concept *c* such that the belief that *c p* is the weakest belief such that one ought not believe both that *c p* and that *p*.

[24] See Richard Foley, "Is It Possible To Have Contradictory Beliefs?", Midwest Studies in Philosophy 10 (1986), pp. 327-355.

or

(ND) The concept *not* is the unique concept *c* such that the belief that *c p* is the weakest belief such that one cannot believe both that *c p* and that *p* (while simultaneously considering both beliefs).

As before, these individuation claims can of course be varied in many ways. Perhaps the simple "ought" claim in (NN) should be replaced with a claim to the effect that the norm is a "basic rule of rationality" in Wedgwood's sense (*The Nature of Normativity*, p. 84).

But it may now seem as if there is a principal problem with (D)-form definitions extracted from (N)-like claims. Consider:

(DNN) NEG(*x*) = the weakest state *S* such that one *ought not* be in *x* and *S*.

(DNN) has certain virtues lacking from some of the proposals considered above: it features no undefined logical notion like "disagreement" or "inconsistent" and does not seem to smuggle in any unexplained notion of rejection or any concealed negation within the scope of an attitude verb (thereby violating Peacocke's non-circularity constraint (*A Study of Concepts*, p. 9)). But it seems that it fails to rule out indifference. That one is indifferent relative to the state expressed by a sentence *s* means that one is not in that state. But assuming the Necessitation rule of standard deontic logic plus intersubstitution of synonyms, it follows that one ought not both be indifferent toward *s* and be in the state expressed by *s*. Thus, it may seem that indifference satisfies the condition on NEG, in which case NEG cannot be said to correspond (in the sense defined above) to negation.

The same, it would seem, holds of the (D)-form definition obtained from (ND):

(DND) NEG(x) = the weakest state S such that one *cannot* be in x and S (while considering both S and x).

Since, by definition, one cannot be indifferent relative to a state one is in, indifference seems to satisfy the condition on NEG also on (DND). Note that both (NN) and (ND) avoid this problem, since it is there already supposed that the state in question is a *belief*, and so indifference is ruled out as a way of being in the state in question.

I think the clue to solving this problem lies in the qualification that the state be the *weakest* state satisfying the condition—not exactly in what this qualification says, but in what it presupposes, namely, that the state is the kind of entity that can figure in inferences. Whether "weaker than" should be defined descriptively (dispositionistically), normatively, or truth-theoretically, it seems clear that the mere absence of an attitude cannot figure in inferences, i.e., as premise or conclusion of an inference. Hence, it cannot be said to be weaker or stronger than other states. (It is of course controversial whether other kinds of states than beliefs can figure in inferences at all, but expressivists are already committed to holding that this is so, so this worry is irrelevant for the present discussion.) This, then, seems like an attractively principled way of excluding indifference as what is expressed by negative sentences. So, since it is presupposed by the qualification that the state be weaker than any other state that the state is of the kind that can figure in inferences, (DNN) and (DND) can remain unchanged in the face of this worry.

What about agnosticism? It is perhaps not as clear that agnosticism satisfies the main condition on NEG that is set by (DND): perhaps it is possible (if irrational and/or unusual) to both be in a state and be determined/committed not to be in it, even while considering both states.

If so, then (DND) excludes agnosticism as a way of rejecting a state, and thus need not be revised. However, I do not know how one might establish this. Could we then instead show that agnosticism can be excluded by the prior qualification of "weakest state"? Unfortunately, this is not as clear as with indifference. For being determined or committed might well be a kind of state that expressivists must say is capable of figuring in inferences. Further, it is not clear whether agnosticism is stronger than rejection. For instance, it may seem that if one rejects a state, one ought to be determined not to be in it, but not *vice versa*. Thus, at least given certain normative definitions of "weaker", agnosticism is plausibly weaker than rejection.

A consoling fact is that a similar problem arises for (NN) or (ND). The fact that it is here claimed that the state expressed by a negative sentence is a belief, we have seen, rules out indifference. But it does not clearly rule out agnosticism. For being determined or committed to not be in a state might, for all we know, be to believe that one ought not be in it. But it is far from clear whether (and in what sense) rejection of a state *S* is weaker than the belief that one ought not to be in state *S* (where, again, rejection of *S* is by definition the state expressed by the negation of the sentence expressing *S*, whether a belief or other type of state). Perhaps, given the right kind of specification of the main condition on negation (the concept) in (NN) or (ND), this belief fails to satisfy the condition. But this can equally be said of (DNN) and (DND). Of course, there are many conceivable ways to revise these definitions so as to exclude agnosticism, but since accounting for negation thus reductively would in any case be a great accomplishment, this is not the place to venture further into such an enterprise.

Appendix: meaning-to-meaning functions

(D)-form definitions define state-to-state functions, but one might complain that a semantics consistent with Expressivism should rather define functions *from meanings to meanings*, and

Expressivism does not hold that the meaning of a sentence is the state it expresses, but is rather a claim as to the nature of the property of meaning that stealing is wrong, and so on. So, the semantics should really define functions from meaning-properties to meaning-properties, where the meaning-property of a sentence "*p*" is the property of meaning that *p*. Fair enough. But, as I will try to show in this Appendix, we can obtain the desired definitions from the (D)-form definitions above in a rather straightforward, if slightly technical, way.

Let us first define a function from states to properties, thus:

(DG) $g(x)$ = the property of expressing x.

The meaning-property of a sentence *s* ("$m(s)$", for short) is now identified with $g(S(s))$ (as before, $S(x)$ is the state expressed by x). In other words, $m(s) = g(S(s))$. Note that this is a substantial claim of (D)-semantics, not a definition.

Now recall that, on (D)-semantics, S("If A then B") = CON(S("A"), S("B")). Since, for all s, $m(s) = g(S(s))$, it follows that m("If A then B") = $g(S$("If A then B")) = g(CON(S("A"), S("B"))). But this merely shows that there is a definable function from the states expressed by "A" and "B" to the meaning-property of "If A then B", whereas what we wanted was a function from meaning-properties to meaning-properties. More specifically, what we want is to define a function h such that m("If A then B") = $h(m$("A"), m("B")). (Notice that h is specific to conditionals. The task of the semantics with respect to connectives in general, we may say, is to define a function, for each connective c, which stands to c as h stands to "if".) The trick is to define h by first defining a function j that takes meaning-properties to states expressed (unlike g, which takes states expressed to meaning-properties), as follows:

(DJ) $j(m(s)) = S(s)$.

This might be somewhat difficult to grasp, but I think it helps to notice that (DJ) entails that for any meaning-property M, M is the meaning-property of s just in case $j(M)$ is the state expressed by s. (Saying that there is such a function as j presupposes that there is a one-to-one correspondence between meanings and states expressed. But this is congenial to expressivism and so should cause no worry.) Now, we can finally define h, the desired function from meaning-properties to meaning-properties, thus:

(DH) $h(x, y) = g(\text{CON}(j(x), j(y)))$.

We now prove (assuming (D)-semantics) that $m(\text{“If A then B”}) = h(m(\text{“A”}), m(\text{“B”}))$, as follows:

1. $m(\text{“If A then B”}) = g(S(\text{“If A then B”}))$ (Assumption)
2. $g(S(\text{“If A then B”})) = g(\text{CON}(S(\text{“A”}), S(\text{“B”})))$ (Assumption)
3. $g(\text{CON}(S(\text{“A”}), S(\text{“B”}))) = g(\text{CON}[j(m(\text{“A”})), (j(m(\text{“B”}))])$ (by (DJ))
4. $g(\text{CON}[j(m(\text{“A”})), (j(m(\text{“B”}))]) = h(m(\text{“A”}), m(\text{“B”}))$ (by (DH))
5. $m(\text{“If A then B”}) = h(m(\text{“A”}), m(\text{“B”}))$ (1-4 and transitivity of =) ■