PROPOSITIONAL ATTITUDES: SOME LOGICAL ISSUES (SELECTED PAPERS 1982–1997)

Lloyd Humberstone

Note: This is an incomplete draft, some 'Updates and Afterthoughts' material still needing to be supplied... I ran into other commitments, as a result of which it was never submitted as a PhD Thesis after all. (LH)

CONTENTS

Declaration. The work included here is my own. This thesis is an annotated compilation of published papers, none of which was co-authored. Acknowledgement of assistance received will be found in each paper, and these acknowledgements are also collected together at the end of Chapter 0.

PROPOSITIONAL ATTITUDES: SOME LOGICAL ISSUES (SELECTED PAPERS 1982–1997)

Lloyd Humberstone

ABSTRACT

We discuss issues of logical interest connected with propositional attitudes, especially wanting, believing, and knowing. One respect in which the first two of these attitudes resemble each other more than they resemble the third is that while a person can want or believe something to be the case without that thing's actually being the case, a person cannot know something to be the case without its actually being the case. This logical difference makes it possible in principle to give an informative account a concept's application in terms of its believed or desired extension, in a way in which it is not possible to do so in terms of its known extension. Although in one sense either type of account would be called 'circular', there is a narrower sense – we call it inferential circularity in Chapter 7 where these matters are discussed – only the less informative style of account counts as ('viciously') circular. A second division amongst the attitudes mentioned would draw the line between belief and knowledge on the one hand, as cognitive attitudes, and desire on the other, as a non-cognitive attitude. In Chapter 6 we take up this theme, in particular as it relates to the difference in what is called 'direction of fit' for belief and desire, and scrutinize several proposals as to what this difference consists in. In passing, we point out that one can agree that desire and belief have different directions of fit in respect of their propositional objects, and at the same time maintain – as is urged in Chapter 2 – that every desire is itself a belief. The latter proposal makes possible a natural representation of certain otherwise intractable sentences, and seems connected with the special affinity of desire as opposed to (mere) belief with the subjunctive mood in languages marking a subjunctive/indicative contrast in verb morphology: see Chapter 1. We consider desire further in Chapter 3, in which, independently of Chapter 2's proposal, two different dimensions of strength of desire are compared (*inter alia*). The remaining two Chapters – 4 and 5 – are concerned with knowledge. The former is concerned with the idea if you can always tell when you are presented with something having a certain property and you know that you have this ability, then you have the further ability to tell, when presented with an object lacking the property, that it does indeed lack the property. (This turns out to be connected with certain issues in traditional epistemology.) The latter chapter shows that under surprisingly mild assumptions, it follows from the hypothesis that

whenever there are two individuals such that whatever is true is known by one or other of them to be true – in which case we call the pair *collectively omniscient* – that one or other of them is omniscient *tout court* in the traditional sense of knowing every truth. Though there are occasional cross-references the papers forming the main body of these chapters were all originally published as independent journal articles and can be read as such here.

CHAPTER 0. FOREWORD

This thesis collects together previously published papers in a single general area—that indicated by its title and somewhat more fully in the Abstract—in accordance with Section 34 ("Staff Candidates") of Monash University's PhD regulations (as published in Appendix A of the 1999 Doctoral Information Handbook). All chapters after this one consist mainly of papers published in the period 1982–1997, supplemented in each case by a postscript section (sub)titled 'Updates and Afterthoughts' touching briefly on a few germane points to have emerged since the original date of publication. Full publication details appear at the head of the first page of each of these chapters. Some changes were made in the interests of uniform presentation. The separate bibliographies and reference-citing styles have been merged, so that now all references are of the 'Segerberg [1973]' variety, directing the reader to the entry so headed in the main bibliography at the end of the thesis. This necessitated a certain amount of re-wording at places, but such references mainly occur in the notes—which here appear at the end of each chapter—and the number (as well as the numbering) of such notes has been kept as in the original publications. I have also changed symbolic notations (such as in the choice of a symbol for negation, which in some of the earlier-published papers appears as "~") so as to give a more uniform effect; negation, conjunction, disjunction, implication and equivalence are represented by $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$; the notation for quantifiers and modal operators is standard and does not need to be explained here. I have also taken the opportunity of re-instating certain features of the papers which were changed to meet the sometimes unfortunate 'house style' demands of the journals in which they were published. To give just one example, the phrase "*a priori*" is not allowed to appear as such in the journal *Mind*, whose editor insists on "a priori" instead; accordingly for the reproduction here of Chapter 6, published in that journal (as 'Direction of Fit'), the italics have been restored.

Two of the papers, 'Scope and Subjunctivity' and 'The Formalities of Collective Omniscience' (forming the body of Chapters 1 and 5, respectively) appear here without the technical appendices that were originally published as part of them. Although the intention is to display certain issues of logical interest, which does in every case except that of Chapter 3 ('Wanting, Having, Getting') indeed involve some formalization, this thesis is not meant to be primarily a work in formal logic, and the technicalities of the original appendices seemed to tip the balance in the wrong direction here; nor, in hindsight at least, do they make for especially edifying reading. (Chapter 7 does have one historical and one somewhat technical appendix, but in this case I have not performed a similar appendicectomy, the material in question being more integrally connected to the main body of the paper than in the earlier cases.) In addition to these excisions, a few clarificatory insertions have been made when this seemed necessary; these are enclosed in double brackets [[like this]]. The subdivision into sections of the original papers has been retained in their reproduction here, which results in one minor anomaly: Chapter 5 ('The Formalities of Collective Omniscience') has, unlike the other chapters—the present Introduction aside—no such internal subdivisions.

The common theme of all these papers is as indicated in our overall title: issues of logical interest arising in connection with propositional attitudes (belief, desire, knowledge,…) and their ascription. However, what this description would most immediately suggest to a philosophical reader is a cluster of much-discussed issues on which thee papers collected here have nothing to say, such as the issues of referential transparency and hyperintensionality: failures, respectively, of the replacement of a singular term by a co-referential term or of a sentence by a logically equivalent sentence in a propositional attitude ascription, to preserve the truth of the ascription. These topics rapidly involve one in what we might call the 'foundational' questions of what the objects of propositional attitudes are and of the logical form of propositional attitude ascriptions *in general*. Famous proposals on such matters are associated with the 'great names' in the philosophy of language, including Frege, Carnap, and Davidson, and there is a large secondary literature, of the type anthologized in Salmon and Soames [1998] and Anderson and Owens [1990]. (The closest we come to a reference in our bibliography pooled from the papers included below is Quine [1956], referred to in Chapters 1 and 2 mostly as the a *locus classicus* for certain worries there set aside. To be fair to Quine, we also mention the paper in Chapter 1 as having – more usefully from our perspective – made a concrete observation on the subjunctive mood in desire-ascriptions in Spanish.) All these issues are no doubt important and of interest to some, but I have had nothing to say about them and have found them to be somewhat orthogonal to – settleable independently of – the issues with which the papers collected here deal. (A brief indication of these issues may be gleaned from the Abstract.)

Since we deal with specific propositional attitudes and sometimes with pairs of such attitudes, another issue which does not receive treatment is what – in general – a propositional attitude is. For example, a general account should make the paradigm cases of knowledge, desire and belief come out as propositional attitudes. (In the case of desire this is perhaps less obvious: what about *wanting a cake*? In Chapter 3 we argue that any such desire-ascription implicitly specifies a proposition whose truth is desired by the wanter. The notion of a proposition, as we note there, has to be understood rather generously – so as to make room for *de se* desire-ascriptions like this one.) But what of *not believing*, for example? That is, not believing in the sense of its not being the case that one believes, that, e.g., the sun is shining, as opposed to not believing that the sun is shining in the sense of believing that the sun is not shining. In the former sense, the most natural answer is that this is *not* a propositional attitude, on the grounds that one can correctly describe someone as not believing a certain proposition not just when – as with the latter sense – the subject disbelieves the proposition, or when the subject grasps the proposition but withholds judgment, having no opinion either way as to the truth of the proposition, but also when the subject is not in a position to grasp the proposition, typically because it involves concepts the subject does not possess. In this last category of cases it would be unnatural to think of the subject as having the propositional attitude of non-belief, and more natural to say that only someone who grasps (= roughly, is in a position to entertain) a proposition can have any attitudes towards it. Fortunately, interesting though this question might be, we are not forced to take sides over it, precisely because of the relative specificity of our concerns in the papers collected here. Some promising thoughts on the matter may be found – sometimes cast in terms of mental states and sometimes in terms of propositional attitudes – scattered throughout Williamson [1995].

The chapters present the papers in very roughly their chronological order of appearance, with departures from this aimed at keeping closely related topics together. Thus the main body of Chapter 3 was published in 1990 whereas that of Chapter 4 appeared in 1988, but it seemed more appropriate for Chapter 3, on desire, to follow Chapter 2, on that same topic. Though Chapter 1 is something close to a prerequisite for Chapters 2 and 6, otherwise the papers may be read in any order. Incidentally, the material in Chapters 1 and 2 and much else originally formed part of a longer single paper presented to the Victorian Branch of the Australasian Association for Philosophy in the late 1970s, and from memory – the typescript having long since gone astray – was much better than the sum of the two papers various unfortunate decisions by journal editors forced to me to carve out of it. To illustrate the order-indifference of the rest of the papers: neither Chapter 2 nor Chapter 3, though both deal with desire (wanting), refers to the other. It would be nice if they were consistent with each other, but no special steps have been taken to argue that this is the case. We are not presenting a sustained argument over the course of several chapters, but simply assembling a collection of published papers on a common theme. This is a primarily a historical record of what actually appeared, which is why the 'Updates and Afterthoughts' have been kept brief, and also why no attempt has been made to reconstruct the full original version of the material in the first two chapters, mentioned above, or to improve the other chapters in any substantive way.

Acknowledgements. While working on the original papers I received considerable assistance from the following individuals: Allen Hazen and Martin Davies (Ch.1); Michael Smith (Ch.3); David Wiggins, David Lewis, David Bostock and John Collins (Ch.4); Frank Jackson and Pamela Tate (Ch.5); John A. Burgess, Rae Langton, Michael Smith, James Hopkins and David Velleman (Ch.6); John A. Burgess, Michael Smith, and Andrew Markus (Ch.7).

From *Philosophia* **12** (1982), 99–126. One proof-sketch from Section 2, and the Appendix to the original paper, have been omitted from this reproduction of the article; a turnstile symbol with three horizontal lines in the original printing has been replaced here by '⊩'. Notes begin on p. 21 below.

CHAPTER 1. SCOPE AND SUBJUNCTIVITY

1. Introduction

Recently, several authors have noted, and attempted to correct, certain expressive inadequacies in the conventional language of modal logic. In Crossley and Humberstone [1977], for example, we observe that the sense of 'It is possible that everything which is red should be shiny' (or, 'It is possible for every red thing to be shiny') in which this sentence asserts the existence of a possible situation or world in which every object red in our world is shiny, is not representable in the usual language. In that paper, it is suggested the expressive resources of this language be enriched by the addition of a new sentence-forming operator on sentences, 'A', (for 'actually'), with the help of which the form of our previously unrepresentable sentence emerges as: $\Diamond \forall x(\mathrm{A}\varphi x \rightarrow \psi x)$. (In this formula, the scope of 'A' is 'φx'; we use '$\Diamond$' and '$\Box$' for possibility and necessity.) Similar considerations move Hazen [1976], [1978], to explore the logic of an 'actually' enriched language. As far as I am aware, the first place in which a relevantly similar expressive problem was noted and corrected in a relevantly similar way is Kamp [1971], in which the background is a Prior-style tense logical language and the required addition is a 'now' operator. The moral of this sort of work is that a certain perhaps *prima facie* plausible 'scope will cope' approach to sentences involving 'actually' ('now') has been shown to be inadequate. Its apparent plausibility is due, no doubt, to the fact that there is a sense—made clear in the papers mentioned—in which scope can be made to cope when only non-quantificational examples are considered.[1]

The trouble with the 'actually'-less modal language, when it comes to sentences like our 'It is possible for every red thing to be shiny', is that if we place '$\Diamond$' in the scope of $\forall$, we get what might be called any-scattering: the resulting formula enables us to pick, for every object red in our world, a different world in which it is shiny. On the other hand, if we place $\forall$ in the scope of '$\Diamond$', we are restricted to making the shininess of objects conditional on their redness in one or more possible worlds, with no way, once we come to consider such worlds, of referring back to our world and what is red in it. It is this function of back-reference that the 'actually' operator fulfills. Other instruments come to mind that might be employed to do the same job—to make good the same expressive weakness—in perhaps other ways. A rather trivial variation on the 'actually' operator idea would be to introduce instead a propositional constant, 'J', say, true only at the actual world (more precisely, using notation introduced in Section 2 below, we require that for any model $\langle W, w^*, V\rangle$ and any $w \in W$, $V(\mathrm{J}, w) = \mathrm{T}$ iff $w = w^*$). Then we could write '$\Box(\mathrm{J} \rightarrow \alpha)$' to express 'Actually α'.[2] Quite different ways of remedying the trouble

would include letting the conventional modal language go infinitary, and letting in '$\in$' and explicit quantification over sets. Just to illustrate briefly how this last idea might prove useful, suppose we use '*s*' as a set variable and '*x*' as an individual variable in a two sorted first order modal language, and consider this formula:

$$\exists s[\forall x(x \in s \leftrightarrow \varphi x) \wedge \Diamond \forall x(x \in s \rightarrow \psi x)]$$

Here we avoid *any*-scattering by parcelling up all (and only) the red objects into a set *s*, and then we say there's a world in which everything in *s* is shiny. Naturally, for this solution to work, we have to think of *s* as being the same set of objects, as defined by membership, from world to world, even though in some worlds those members do not satisfy the condition we use to specify *s* in the first place. None of these suggested alternatives to the use of 'actually' will concern us further in this paper. Instead, we shall be looking into the connections between the 'actually'-enriched language and another, rather unusual language, modelled (not very closely, but sufficiently so in crucial respects) on a deontic language evolved by H.-N. Castañeda to make good parallel expressive defects in the languages customarily used by workers in deontic logic. The concentration will be on the propositional fragments of those languages because it is at this level that the semantic novelties involved emerge, even though, as already suggested, the more interesting applications arise from the interplay between the semantic machinery set up at that level and the apparatus of quantification. This matter receives a brief discussion in Section 3. Finally, in Section 4, attention is directed to related expressibility problems arising in connection with certain propositional attitude ascriptions.

2. Languages, Models, Translations

We begin with some informal remarks about that aspect of Castañeda's motivation for developing his own deontic language that is of concern to us here. (For fuller discussion of these and related points, see Castañeda [1967], [1967*a*], [1975].) An appropriate point at which to start is Castañeda's observation[3] that each of the following sentences has a clear sense not representable in standard deontic languages (with monadic or dyadic obligation operators):

(1) It is permitted that everybody who did B do A.

(2) It is obligatory that someone who chaired a meeting last year preside over the entire committee this year.

The interesting thing about these sentences, taken in the sense at issue, is that – as Castañeda puts it – the reference in (1) to having done B, and in (2) to having chaired a meeting, occur *inessentially* but *ineliminably* within the scope of the deontic operators. What this comes down to is this: there is no equivalent formulation of either (1) or (2) in which these references appear outside the scope of the deontic operators, yet all the same, one who asserts either is not thereby saying anything about the permissibility of having done B or about the obligatoriness of having chaired a meeting. Taking (2) first, and using 'O' for the obligation operator, 'φx' for '*x* chaired a meeting last year' and 'ψx' for '*x* presides over the entire committee this year', we cannot write '$\exists x(\varphi x \wedge O\psi x)$', since this attributes too specific an obligation; there need be no person or persons who are such that they chaired a meeting last year and have an obligation to

preside over the entire committee this year. But re-scoping to get round this difficulty gives us '$O\exists x(\varphi x \wedge \psi x)$' and this implies that it is obligatory that someone chaired a meeting last year, which is not part of what is intended in (2). Similar points may be made in connection with (1). Castañeda speaks of it being only the 'ψx' part of the formal renderings provided that specifies an action 'prescriptively considered'; the 'φx' part, on the other hand, describes only the circumstances surrounding this prescription. Let us therefore introduce into our deontic language an operator which explicitly indicates that what appears in its scope is being prescriptively considered. I will write 'S' for this operator. The 'S' is for 'subjunctive', since, as Castañeda has observed, in languages, such as Spanish, in which a subjunctive/indicative contrast survives more robustly than in English, the 'prescriptively considered' clause characteristically has a subjunctive main verb.[4] Notice, incidentally, that even in English (2), we have 'preside' instead of 'presides'. Overexposure to the English renderings of deontic formulae given by many deontic logicians can blind one to such niceties. So the idea is that we write, for (1) and (2) respectively, (3) and (4); in (3), 'φx' is for '*x* did B', 'ψx' for '*x* does A', and 'P' is the permissibility operator:

(3) $P\forall x(\varphi x \rightarrow S\psi x)$

(4) $O\exists x(\varphi x \wedge S\psi x)$

It is only 'P' or 'O' *together with* 'S' that expresses permissibility or obligation; nevertheless, the initial 'P' or 'O' must appear precisely where it does appear in order that the permissions and obligations in question be correctly allocated.

It can hardly have escaped the reader's notice that Castañeda's (1) is nothing other than a deontic version of our modal sentence about everything red being shiny. Similarly, we could provide modal analogues to (2). Thus, using the resources of Crossley and Humberstone [1977], adapted to the deontic case, we can give 'actually' translations of (3) and (4):

(5) $P\forall x(A\varphi x \rightarrow \psi x)$

(6) $O\exists x(A\varphi x \wedge \psi x)$

(It is notable that the appropriate reading for (6)—appropriate in the sense of corresponding to what the formal semantics to be given presently for the 'actually' language delivers as the truth conditions for such formulae—is one using the relative clause construction ('It ought to be that someone who is actually/in fact φ, is ψ') rather than explicit conjunction.) Conspicuously, 'A' appears where 'S' doesn't appear in the Castañeda versions, and *vice versa.* This suggests that the two operators are performing complementary jobs, in some sense. This sense we must now attempt to make a little clearer, and, as has already been mentioned, the clarification can best proceed at the purely propositional level. (Lest anyone unfamiliar with Castañeda's own work should be misled, I ought to point out that Castañeda does not himself use an operator like 'S', instead classifying atomic formulae into indicative and non-indicative on the basis of their form. Further, he thinks of the non-indicative formulae as somehow imperatival, as is no doubt suggested by consideration of expressibility problems in deontic logic. Since parallel problems arise in non-deontic areas, however, we will not adopt that way of speaking, or the accompanying idea that it is somehow especially a failure to mark off acts prescriptively considered from other acts and states that is responsible for the expressive weakness of the

received languages, though this is undoubtedly how such problems most strikingly present themselves in the deontic case.)

Turning now to the interaction between 'S' and the truth-functional connectives, we find the suggestion that it is only in combination with 'S' that the deontic operators express permissibility and obligation borne out by Castañeda's acceptance of equivalences like (7) and (8), and their analogues with 'P' for 'O':

(7) $O(p \wedge Sq) \leftrightarrow (Op \wedge OSq)$

(8) $O(p \vee Sq) \leftrightarrow (Op \vee OSq)$

Further, since the work done by 'O' (or 'P) is exhausted by its role in combination with 'S', and 'S'-free formula, such as the plain propositional variable 'p', preceded by 'O' ('P') is equivalent to that variable standing alone, and we have:

(9) $O(p \wedge Sq) \leftrightarrow (p \wedge OSq)$

(10) $O(p \vee Sq) \leftrightarrow (p \vee OSq)$

In a deontic language with an added 'actually' operator but otherwise conventional, one would have, parallelling (7) and (8):

(11) $O(Ap \wedge q) \leftrightarrow (OAp \wedge Oq)$

(12) $O(Ap \vee q) \leftrightarrow (OAp \vee Oq)$

and parallelling (9) and (10):

(13) $O(Ap \wedge q) \leftrightarrow (Ap \wedge Oq)$

(14) $O(Ap \vee q) \leftrightarrow (Ap \vee Oq)$

In view of all these equivalences, I think it is useful to contrast 'S' and 'A' in the following terms. Relative to some operator (deontic, alethic, or whatever) whose semantics is givable by quantification over possible worlds, 'S' functions semantically to activate the possible worlds quantifier at the point where it appears. Loosely, it says "There's a free world-variable here"; what this comes down to exactly will be apparent from the semantics for a Castañeda-style language given below. If, in such a language, 'S' functions as an *activator*, then, again relative to some modal operator, in the 'actually' language, 'A' functions rather as an *inhibitor*; semantically, it protects what is in its scope from the influence of an outlying modal operator by saying "There's no free world-variable here".[5] To elucidate and substantiate these rather impressionistic remarks, we proceed to define well-formedness for the two languages we are interested in comparing, and then to give the model theory needed. For simplicity, we deal with modal rather than deontic languages, taking it that the 'S' operator is to behave with respect to necessity and possibility just as it is seen to behave, in (7)–(10), with respect to obligation and permissibility.

First, for the 'actually' language, which I shall call L_A. It has a familiar grammar: an expression is a (well-formed) formula of L_A iff it is so in virtue of (*i*) and (*ii*).

(*i*) A propositional variable (p, q, . . .) is a formula of L_A

(*ii*) If α and β are formulae of L_A, then so are $\neg\alpha$, $\Box\alpha$, $A\alpha$, and $\alpha \wedge \beta$.

We economize on the primitives in (*ii*) so as to shorten truth-definitions and inductive proofs, helping ourselves where convenient to the familiar abbreviations. And now to what I shall call—although it differs somewhat from any of the languages he has actually proposed—the Castañeda language, L_C. Well-formedness is not amenable to the simple inductive type of definition which suffices for L_A, and must be approached via the characterization of an 'interim' language, L_I:

(L_I) (*i*) A propositional variable is a formula of L_I

(*ii*) If α and β are formulae of L_I then so are $\neg\alpha$ and $\alpha \wedge \beta$.

(*iii*) Nothing else is a formula of L_I .

(L_C) (*i*) If α is a formula of L_I, then α and $S\alpha$ are formulae of L_C .

(*ii*) If α and β are formulae of L_C, then so are $\Box\alpha$, $\neg\alpha$, and $\alpha \wedge \beta$.

(*iii*) Nothing else is a formula of L_C.

Obviously, L_I contains just the purely truth-functional formulae; the way L_I is extended to L_C guarantees that 'S' has only truth functional formulae within its scope, though complete combinatorial freedom is given in this regard to the other connectives, including $\Box$. We shall have occasion to comment on this feature of L_C below, in Section 3. In the meantime, we turn to semantics.

We shall pursue the semantics of L_A and L_C from the point of view of a common notion of what it is to be a model. A model, $\mathtt{M}$, is to be a triple $\langle W, w^*, V\rangle$, where W is a set containing at least w^*, and V is a function taking propositional variables paired with elements of W to truth-values (T and F). (Of course, we are thinking of W as the set of possible worlds and of its distinguished element w^* as the actual world of the model.) We now extend V to define truth at an element w of W for an arbitrary formula α of L_A (symbolized '$\mathtt{M} \vdash_w \alpha$'). This extension is straightforward, and proceeds thus (as in Crossley and Humberstone [1977]). Given some model $\mathtt{M} = \langle W, w^*, V\rangle$:

A(*i*) If α is a propositional variable and $w \in W$, then $\mathtt{M} \vdash_w \alpha$ iff $V(\alpha,w) = \mathrm{T}$.

A(*ii*) $\mathtt{M} \vdash_w \neg\alpha$ iff not $\mathtt{M} \vdash_w \alpha$

A(*iii*) $\mathtt{M} \vdash_w \alpha \wedge \beta$ iff $\mathtt{M} \vdash_w \alpha$ and $\mathtt{M} \vdash_w \beta$

A(*iv*) $\mathtt{M} \vdash_w \Box\alpha$ iff for each $w' \in W$, $\mathtt{M} \vdash_{w'} \alpha$

A(*v*) $\mathtt{M} \vdash_w A\alpha$ iff $\mathtt{M} \vdash_{w^*} \alpha$

Similarly, we must show how to extend V to define truth for an arbitrary formula α of L_C at an element w of W (symbolized '$\mathtt{M} \models_w \alpha$'). Predictably, for L_C, matters are a little less simple and we must define the primary notion of truth we are interested in, $\models_w$, partly in terms of an ancillary notion, $\Vdash_w$, let's call it. We begin with $\Vdash_w$, which is defined only for truth-functional formulae of L_C. With its aid, from C(*iv*) onward, $\models_w$ is determined. Given some model $\mathtt{M} = \langle W, w^*, V\rangle$:

C(*i*) If α is a propositional variable, then $\mathtt{M} \Vdash_w \alpha$ iff $V(\alpha, w) = \mathrm{T}$.

C(*ii*) $\mathtt{M} \Vdash_w \neg\alpha$ iff not $\mathtt{M} \Vdash_w \alpha$

C(*iii*) $\mathtt{M} \Vdash_w \alpha \wedge \beta$ iff $\mathtt{M} \Vdash_w \alpha$ and $\mathtt{M} \Vdash_w \beta$

C(*iv*) If α is a propositional variable, then $\mathtt{M} \models_w \alpha$ iff $\mathtt{M} \Vdash_{w^*} \alpha$

C(*v*) $\mathtt{M} \models_w \mathrm{S}\alpha$ iff $\mathtt{M} \Vdash_w \alpha$

C(*vi*) $\mathtt{M} \models_w \neg\alpha$ iff not $\mathtt{M} \models_w \alpha$

C(*vii*) $\mathtt{M} \models_w \alpha \wedge \beta$ iff $\mathtt{M} \models_w \alpha$ and $\mathtt{M} \models_w \beta$

C(*viii*) $\mathtt{M} \models_w \Box\alpha$ iff for each $w' \in W$, $\mathtt{M} \models_{w'} \alpha$

An immediate consequence of our definitions of $\vdash$ and $\Vdash$ is that, relative to any model $\mathtt{M} = \langle W, w^*, V\rangle$, for any purely truth-functional formula α of L_C (which is accordingly also a formula of L_A) $\mathtt{M} \Vdash_w \alpha$ iff $\mathtt{M} \vdash_w \alpha$, for all $w \in W$. We shall make use of this fact shortly, in discussing a certain translation from L_C into L_A. In the meantime, notice that the clauses for $\Box$ in both languages are uncluttered by any restriction to accessible worlds. Thus we have in mind S5 as our underlying modal logic in both cases. As in Crossley and Humberstone [1977], we call a formula of L_A S5A-*valid* just in case it is false at no world in any model. Similarly, if for some formula α of L_C we can find no model $\mathtt{M} = \langle W, w^*, V\rangle$ with $w \in W$ such that not $\mathtt{M} \models_w \alpha$, then we call α S5C-*valid*. [[See the end of 'Updates and Afterthoughts' on this terminology.]] The notions of S5A-validity and S5C-validity by no means coincide for those formulae in $L_A \cap L_C$; for example, '$p \rightarrow \Box p$' is S5C-valid but not S5A-valid. This is because the force of some occurrence of a symbol, such as a propositional variable, is apt to be different depending on whether that symbol occurs in a formula of L_C or in a formula of L_A. In L_A a propositional variable is true at a world as V, the valuation component in the model, dictates for that world, whereas in L_C, a propositional variable, indeed any formula no subformula of which is prefixed by an 'S' is true (in the sense of $\models_w$) depending only on how V dictates things should be for the actual world in the model. It is only when we hit an 'S' that we become concerned about how things are (in the sense of $\Vdash_w$, which just records the consequences of V's allocation of truth values to variables in w) at other non-actual worlds, which is why it is only together with 'S' that '$\Box$' expresses necessity. Thus our semantic account of L_C casts the subjunctive mood, or our

formal generalization thereof, in the role of the mechanism which brings considerations about non-actual worlds into the determination of the truth of sentences at the actual world.

For the benefit of those who only begin to feel at home with a logic when they've got their hands around some axioms, a few comments about proof theory are in order here, before we go on to investigate, with the aid of some translations, the semantic differences between L_A and L_C. A sound and complete axiomatization of the S5A-valid formulae is got by subjoining to any basis utilizing necessitation sufficient for (propositional) S5, all instances of the following schemata:

(A1) $A\alpha \leftrightarrow \neg A \neg \alpha$

(A2) $A(\alpha \rightarrow \beta) \rightarrow (A\alpha \rightarrow A\beta)$

(A3) $\Box\alpha \rightarrow A\alpha$

(A4) $A\alpha \rightarrow \Box A\alpha$

This fact is proved in Crossley and Humberstone [1977].[6] For S5C, we add to a similar basis, though here insisting also that the basis is given by axiom-schemata rather than by particular axioms with a uniform substitution rule,[7] all (well-formed) instances of the following forms:

(C1) $S\alpha \leftrightarrow \neg S \neg \alpha$

(C2) $S(\alpha \rightarrow \beta) \rightarrow (S\alpha \rightarrow S\beta)$

(C3) $\Box S\alpha \rightarrow \alpha$

and, provided that α contains no occurrence of 'S'[8], all formulae of the form:

(C4) $\alpha \rightarrow \Box\alpha$

This axiom system is easily seen to be sound with respect to the semantics already outlined; completeness is proved in the Appendix [[omitted from this reproduction]] to this paper. For the moment, we content ourselves with two observations. The first is that, since we already have all of S5, (C4) assures us that any S-free formula is equivalent to its own necessitation; this should come as no surprise in the light of our earlier remarks to the effect that '$\Box$' only gets to involve other worlds (i.e., those other than *w**) through its interaction with 'S'. Secondly, (Cl) and (C2) conspicuously copy (Al) and (A2). This reflects the fact that both 'A' and 'S' are 'monocosmic' operators. These axioms play the vital role of guaranteeing that these operators, unlike the quantificationally defined modal operators, distribute over all the truth-functional connectives. Nevertheless, whereas for 'A' it is truth at the one world *w** that matters, for 'S' (as C(*v*) shows) there is no one world in the model such that truth at that world is what matters: rather, relative to any world, what matters is truth at that world itself.

We now turn to the promised translation from L_C into the more familiar language L_A. We define this translation, $(\ )^A$, in two stages. If α is any formula of L_C, then to get $(\alpha)^A$, *first* replace every occurrence of a propositional variable β not in the scope of some occurrence of 'S' in α, by $A\beta$, and *then*, delete all occurrences of 'S'. Clearly, $(\alpha)^A$ will always be a formula

of L_A; but to justify calling the function $(\ )^A$ a translation we must show that it meets the condition that for any model $\mathtt{M} = \langle W, w^*, V\rangle$ and any $w \in W$, $\mathtt{M} \models_w \alpha$ iff $\mathtt{M} \vdash_w \alpha$. We sketch an inductive argument to that effect. [[Argument omitted in this transcription of the original article, *q.v.* for details.]]

It is interesting to observe that our translation $(\ )^A$ does not satisfy a condition satisfied by more familiar examples of translations between formal languages—namely, that if β is a subformula of α, then the translation of β is a subformula of the translation of α. For instance, while 'p' is a subformula of 'Sp', $(p)^A$, i.e., 'Ap', is not a subformula of $(\text{'S}p\text{'})^A$, i.e., 'p'. It might be objected that this circumstance shows that something must be wrong either with L_A or with L_C, since one of them must have the relation 'is a semantic constituent of' (reflected syntactically by the subformula relation) all wrong. This would again be to forget that those expressions be longing to both languages have quite different features accorded them by the different semantic descriptions of the languages. A second point to notice about $(\ )^A$ is that it is many-one: there is no such thing as 'the' formula of L_C which has a given formula of L_A as its translation. For example, '$p \wedge q$' is the translation of the L_C formula 'S($p \wedge q$)' as well as of the L_C formula 'S$p \wedge$ Sq'. Of course, it is an immediate corollary to the proof given in the previous paragraph that whenever, for distinct α and β, $(\alpha)^A$ is the same formula as $(\beta)^A$, α and β are equivalent in S5C. Thirdly, $(\ )^A$ is *into* rather than *onto*: there are many formulae of L_A which are not the translations of any formulae of L_C under $(\ )^A$; for example, this would be true of any formula of L_A not all of whose occurrences of 'A' preceded propositional variables. Again, however, any such formula is equivalent to some formula which is the translation of some formula of L_C. It is perhaps worth closing this discussion of $(\ )^A$ by noting that a translation differing slightly from $(\ )^A$ could have been considered for which the result of the previous paragraph also holds, namely the following: given a formula α of L_C, *first* replace every maximal S-free subformula β, of α by Aβ, and *then*, delete all occurrences of 'S'. Here, by 'β is a maximal S-free subformula of α' we mean that β is a subformula of α containing no occurrences of 'S' and which is not a proper subformula of any other subformula of not containing any occurrences of 'S'.

These considerations suggest that L_A and L_C are expressively equivalent languages, in any sense that such a phrase might reasonably bear (though I shall not attempt to make the notion precise in full generality, i.e., for arbitrary languages). So it might occur to the reader that as well as considering translating from L_C to L_A, which is doubtless more important for clarificatory purposes in view of some of the unusual features of L_C, a translation for which a similar result holds ought to be exhibited from L_A to L_C. A natural suggestion is the following. Given a formula α of L_A, to get $(\alpha)^C$—let's call it—*first*, replace every variable, β, not in the scope of some occurrence of 'A' in α by Sβ; *then* delete all occurrences of 'A'. This translation is not, however, generally truth-preserving in the way that $(\ )^A$ is. (Consider, for example, $(\text{'A}\Box p\text{'})^C$.) One remedy might be to replace the phrase 'in the scope of some occurrence of 'A'', in step one of our proposed $(\ )^C$ by "in the direct scope of some occurrence of 'A'", where we say that a formula is in the *direct scope of* a particular occurrence of 'A' if it is in the scope of that occurrence but not in the scope of any occurrence of '$\Box$' which is itself within the scope of that occurrence of 'A'. A simpler solution would be to stick with $(\ )^C$ as it stands and regard it as a

translation of a proper fragment of L_A into L_C. The formulae in this fragment will be those in which the scope of any occurrence of 'A' is a purely truth-functional formula. Every formula of L_A is equivalent (in S5A) to such a formula and the proof of truth-preservingness of $(\ )^C$ for these formulae goes through without a hitch, much as for $(\ )^A$. (Of course, for formulae of this fragment, the possibilities of occurrence for 'A' are just those in L_C for 'S'.)

Finally, let us return to our initial characterization of 'A' and 'S' in their relation to '$\Box$', as, respectively, inhibitors and activators. I hope that the intervening discussion has helped to clarify the point of that characterization, originally prompted by perusal of the equivalences (9), (10), (13) and (14)—the analogues of which with '$\Box$' in place of 'O' are of course equally valid. My intention now is to forestall a possible misapprehension that could be caused by an overconcentration on those equivalences. Rather than complicate matters by dealing simultaneously with L_C and L_A, we make the point with reference to L_C, noting that an analogous point holds for L_A. Let us call a first-degree formula of L_C (one containing no embeddings of 'L') ordinary just in case it is equivalent to that formula we get from it by first replacing every 'S' in the scope of '$\Box$' by '$\Box$S', and then deleting all occurrences of '$\Box$' which do not immediately precede an occurrence of 'S'. Intuitively, we may think of this operation as a moving in of occurrences of '$\Box$' so that they attach directly to the occurrences of 'S' which activate them: the significance of ordinariness is then that means nothing is gained in expressive power in separating '$\Box$' from 'S', and so nothing is gained in what is really the distinctive feature of L_C. For example, (9)—reading 'O' as '$\Box$'— tells us that the formula '$\Box(p \wedge Sq)$' is ordinary since this formula is equivalent to '$p \wedge \Box Sq$'; similarly, as (10) says, '$\Box(p \vee Sq)$', being equivalent to '$p \vee \Box Sq$', is another ordinary formula. Now the misapprehension that seems worth averting is this: that every first-degree formula is ordinary. A simple example of a non-ordinary formula would be '$\Box(Sp \vee Sq)$', which is of course not equivalent to '$\Box Sp \vee \Box Sq$'. More interesting examples of non-ordinary formulae are those in which not every variable within the scope of '$\Box$' is in the scope of 'S', such as the formula '$\Box((p \wedge Sq) \vee (r \wedge Ss))$'. I say these formulae are more interesting because, in terms of the $\wedge$:$\forall$ ($\vee$: $\exists$) analogy, they are the propositional analogues to the formulae involving quantifiers which gave rise to the expressibility problem in the first place. The non ordinary formula just exhibited, for example, is a propositional version of Castañeda's problematic sentence (2) about the committee. An important fact about propositional L_C, however, is that although not every formula is ordinary, every formula is equivalent to some formula each of whose occurrences of '$\Box$' immediately precedes 'S'. For example, the formulae '$\Box(Sp \vee Sq)$' and '$\Box((p \wedge Sq) \vee (r \wedge Ss)$' already cited as not being ordinary formulae, are equivalent to '$\Box S(p \vee q)$' and '$(p \vee r) \wedge (p \vee \Box Sq) \wedge (r \vee \Box Sq) \wedge \Box S(q \vee s)$', respectively, both of which pass the test of containing no S-separated '$\Box$'s. A similar notion of ordinariness can be defined for L_A and the result that any formula is equivalent to some formula containing no occurrence of 'A' within the scope of '$\Box$' (the relevantly similar result) can be shown to hold. As the discussion of Section 1 (and the references cited there) shows, these results do *not* extend once quantifiers are added—of which more in Section 3. There should therefore be no tendency to feel that as a result of these facts about L_A and L_C at the propositional level, no increase in expressive power has been achieved by adding our new operators—whether 'A' or 'S'—to the customary stock.[9]

Before closing the present section, we should return briefly to our starting point—deontic logic. The most straightforward way to feed the machinery developed here back into deontic logic is to take a model to consist of a set W with distinguished element w^* and also (as well as the valuation V) a distinguished subset X which we require to be non-empty but which we require neither to include nor to exclude w^*. Formation rules for L_C and L_A are changed so that 'O' replaces '$\Box$'; and in the semantics '$\Box$' is interpreted by quantification over X instead of over W (thus we are thinking of X as the set of ideal worlds).

3. Two Extensions of the Machinery

In this section, we offer a few remarks on the subjects of iteration and of adding quantifiers. The question of iteration raises no special problems for L_A since its formation rules and semantics are of the standard sort. In L_C, '$\Box$' can appear within its own scope but not within the scope of 'S'. Thus only innermost occurrences of '$\Box$' get triggered (activated) by occurrences of 'S': outer occurrences are just idling. It is clear that we can permit ourselves this convenient simplicity only because we have taken our underlying modal logic to be S5;[10] one wants to know what the *general* situation is.

The syntax and semantics of L_C are right for each other: the former throws up as well-formed just what the latter can interpret. '$\Box$' doesn't occur within the scope of 'S' because '$\Box$' is a $\models$-level operator and 'S' takes us down to the $\Vdash$-level; similarly, 'S' does not occur within its own scope because, having been made, this descent cannot be repeated. One could of course liberalize the formation rules so as to allow '$\Box$' within the scope of 'S' (with or without further occurrences of 'S') and in the semantics give a $\Vdash$-clause for '$\Box$' just as, alongside C(*vi*) and C(*vii*), we have C(*ii*) and C(*iii*) — $\Vdash$-clauses for '$\neg$' and '$\wedge$'. This would go against the spirit of L_C however, undermining the leading idea that necessity (or whatever other modal notion) is expressed by '$\Box$' and 'S' in collaboration. There are, however, a number of legitimate modifications to the syntax that might be made, and which one chooses had probably better depend on the particular application one has in mind. For reasons of space, I shall mention only that which is most suited to the discussion of quantifiers to follow. The idea is to have an operator which reverses the effect of 'S': via it, in other words, we ascend from the $\Vdash$-level to the $\models$-level. Although such an operator has much in common with the 'A' of L_A. to avoid any risk of confusion, we had better symbolize it differently, say as 'Z', with the semantic clause:

(CZ) $\mathtt{M} \Vdash_w Z\alpha$ iff $\mathtt{M} \models_w \alpha$.

(There is thus *some* similarity between the relation between 'S' and 'Z' and that between Vlach's and Kamp's operators in two-dimensional tense logic.[11] It would, however, take several paragraphs to go into the connections between the present approach using two truth relations and the idea of using a single doubly-indexed such relation.) For the language for which C(*i*)–C(*viii*) together with (CZ) provide a semantic account, what is the matching account of well formedness? This account should match in the sense that we require as well-formed anything which could be assigned truth (-in-the-sense of-'$\models$') conditions by the clauses listed,

and nothing which could not (so that we must not, in the course of evaluating a formula, come across a subformula which we are required to evaluate with respect to a truth-relation for which it is not defined). Intuitively, we have to lace 'S' and 'Z' together in such a way that they alternate; an eye must be kept on the behaviour of '$\Box$' also. My suggestion is the following. Let L_{C0} be L_C and say: if α is a formula of L_{Ci} containing the propositional variables $\pi_1,\ldots,\pi_n$ within the scope of 'S', and $\beta_1,\ldots,\beta_n$ are any formulae of L_C, then the result of substituting (simultaneously) '$Z\beta_i$' for π_i in α is a formula of L_{Ci+1}. Finally, we define L_{CZ}, the language we are interested in, to be the union of all the L_{Ci}. We shall presently have occasion to observe that, when quantifiers are introduced, L_{CZ} is a more powerful language than L_C: the ability to shift back up to the $\models$-level is by no means redundant. For the moment, we note that, when, as is appropriate for the weaker modal logics, C(*viii*) is replaced by a clause involving an accessibility relation, L_{CZ} gives us what we want in the case of embeddings and iterations. For example, if the accessibility relation is required to be transitive, we have as a valid formula of L_{CZ} the following—admittedly rather clumsy—version of the S4 axiom: $\Box Sp \rightarrow \Box SZ\Box Sp$.

We now turn to the question of quantification, which after all prompted interest in L_A and L_C in the first place. We assume that both languages start from atomic formulae of subject-predicate structure and that quantifiers are added in the usual way (so that if v is a variable and Q is either '$\forall$', or '$\exists$', 'Qv' when attached to a formula, makes a formula); naturally we are primarily interested in those formulas containing no vacuous quantifiers and no free variables. For the semantics, we take our models $\mathtt{M}$ now to have the form $\langle W, w^*, D, Q, V\rangle$ where $w^* \in W$, as before, D is a nonempty set, subsets of which are assigned by Q to elements of W (so that $Q(w)$ is the domain of world w) and V assigns to each n-adic predicate paired with element w of W a subset of $Q(w)^n$. We want to interpret the quantifiers 'locally', i e., when evaluating a quantified subformula such as $\exists v.\alpha$ with respect to a world w, we take the quantifier as ranging over $Q(w)$.[12] This means adopting for L_A the following clause, in which s, s', etc. range over countable sequences of elements drawn from D and "$\vdash^s_w \alpha$" is read "s satisfies α in w", "$\cong_i$" expresses sequence-difference in at most i^{th} position, and we take A(*ii*)–A(v) of Section 2 to have an idle sequence parameter read into them, with A(i) suitably changed for predicate + terms atomic formulae:

(AE) $\mathtt{M} \vdash^s_w \exists v_i.\alpha$ iff for some $s' \cong_i s$ such that $s'_i \in Q(w)$, $\mathtt{M} \vdash^{s'}_w \alpha$.

We may take 'A' to be introduced by definition, and so to have a predictably corresponding sense. It is now clear why in Crossley and Humberstone [1977] to formalize in quantified L_A the sentence 'It is possible for every red thing to be shiny' (understood as in Section 1) we needed to assume that the domain did not vary from world to world. More particularly, the danger arises from loss of objects in the domain of w^* as we pass to the world in which we want to be shiny (if red in w^*): for our proffered representation, '$\Diamond \forall x(A\varphi x \rightarrow \psi x)$' ends up meaning that there is a possible world in which everything *which happens to exist there and* which is red in the actual world, is shiny. We cannot, once we admit variation in domain, write down a formula of quantified L_A which means what we have just said, minus the italicized restriction. (Not that this renders quantified L_A uninteresting; alter all we cannot express even the restricted thought in the conventional language of modal logic.)

Next, we consider the addition of quantifiers to L_C. To follow the pattern set in Section 2 for the truth-functional connectives, we shall need to give not one (such as (AE)) but two clauses for 'E', and the natural choices are the following:

(CE1) $\mathtt{M} \models^{s}_{w} \exists v_i.\alpha$ iff for some $s' \cong_i s$ such that $s'_i \in Q(w^*)$, $\mathtt{M} \models^{s'}_{w} \alpha$.

(CE2) $\mathtt{M} \Vdash^{s}_{w} \exists v_i.\alpha$ iff for some $s' \cong_i s$ such that $s'_i \in Q(w)$, $\mathtt{M} \Vdash^{s'}_{w} \alpha$.

The difference between these two clauses reflects our established policy of letting the facts of the actual world determine $\models$-truth until we are forced (by an occurrence of 'S') to consider $\Vdash$-truth, which does genuinely hinge on how things are at non-actual worlds. In view of the expressive equivalence of L_C and L_A at the propositional level, it is of some interest to note that the quantified versions of these languages are not even comparable: i.e., neither can express everything the other can. This was pointed out to me by Allen Hazen, who offered the following examples:

(15) It is possible for there to be something other than what there actually is.

(16) It is possible for everything which actually exists to exist and it be the case that *p*.

(15) is representable in quantified L_A as: $\Diamond \exists x A \sim \exists y(y = x)$ and (16) goes into quantified L_C as: $\Diamond \forall x S \exists y(y = x \wedge p)$, but, as the reader may verify by experiment, (15) will not go into L_C and (16) will not go into L_A. *Ad* (15)—this cannot be represented in L_C because to assert possible existence, we need a quantifier within the scope of an 'S' in the scope of a '$\Diamond$': but then to assert non-identity with any actual object we need to get back to the $\models$-level still within the scope of the quantifier just mentioned (and so still within the scope of 'S')—and this we simply cannot do in L_C. *Ad* (16), in L_A we can achieve what in Section I was called back reference to what holds of objects in the domain of some possible world in the actual world, but there is no way of quantifying over the whole domain of the actual world from the perspective of some other world because of the 'local' interpretation of the quantifiers. Interestingly, both of Hazen's examples are representable in quantified L_{CZ}, i.e., the language standing to L_{CZ} as quantified L_C stands to L_C, with a sequence-relativized version of (CZ) and the two clauses (CEI) and (CE2). Since any formula of L_C is a formula of L_{CZ}, we have already shown this for (16). For (l5) we write:

$$\Diamond S \exists x(Z \sim \exists y(y = x))$$

Thus L_{CZ} combines the expressive virtues of both L_A and L_C. As it happens, the original example about the shiny red things can be adequately captured, even with varying domains, within just quantified L_C, namely by: $\Diamond \forall x(\varphi x \rightarrow S\psi x)$. This may serve as our final example of the differences between the two languages of Section 2 that emerges when the apparatus of quantification is added.

4. Some Remarks on the Attribution of Desires

In the final section of the paper we note that the expressive weaknesses of conventional modal (and deontic) languages arise again in what is perhaps the most widely used sort of language for

the ascription of propositional attitudes. The discussion will be rather more in formal than that of the preceding sections because it is not obvious just what a suitable formal semantic treatment of such ascriptions ought to look like; it is, however, heuristically rather helpful to bear the possible world approach to such matters in mind here. Such issues as raise problems for that approach (e.g., hyperintensionality) will in any case be completely ignored. Further, I offer the following warning to the more delicate reader: I shall quantify unashamedly into these contexts, taking it that the intelligibility of the resulting formulae stands or falls with the intelligibility of the English sentences they are claimed to represent.

In what might be called the standard approach to propositional attitude ascriptions, it is assumed that we can treat, e.g., 'Tim believes that', 'Harry knows that', and 'Bill wants that', as (no doubt analysable) sentence-forming operators on sentences with the embedded sentence giving what is normally regarded as the content of the attitude in question (*what* is believed, known, or wanted.)[13] Now this approach works well enough for some propositional attitudes, but not at all well for others. Belief is an example of the first class, and desire, an example of the second. To see the difficulty about ascriptions of desire, we may consider the well-known example[14]

(17) John wants to catch (some) communist spies.

We are to imagine that John is an agent for, say, the F.B.I. and is accordingly interested in capturing at least some of the communist spies around. It is quite widely known that there is no way of representing (17) within what I have called the standard framework, which provides us with only the following three options (I abbreviate 'John wants that' to 'W'):

(18) $W\exists x(x$ is a communist spy $\wedge$ John catches $x)$

(19) $\exists x W(x$ is a communist spy $\wedge$ John catches $x)$

(20) $\exists x(x$ is a communist spy $\wedge$ W(John catches $x))$

(18) will not do because it represents John as wanting, *inter alia*, that there be communist spies, which is certainly no part of what he wants; (20) will not do because it attributes to John a specific desire with respect to one or more communist spies that he catch them, whereas his suspicions may not yet have settled on any particular individuals; and (19) manages to suffer from both faults at once. Our hopes for the representation of (17) thus seem destined to remain unfulfilled as long as we treat 'W' as all there is to the want-ascribing element available, forever vying with the existential quantifier for scope priority. To give either priority is to give up hope of rendering (17).

We must now make two observations about the *impasse* thus reached in the 'standard framework'. First: that no such difficulty arises if the sentence is not (17) but (21):

(21) John believes that he will catch (some) communist spies.

So that what is at issue does not affect all propositional attitudes (though it is not only desire that is involved; 'fears that' gives the same trouble, as do, quite generally, what are sometimes called the conative, as opposed to the cognitive, propositional attitudes). Secondly, observe that we are now right back where we started with Castañeda's example about the committee (sentence (2) in Section 2). In both cases, we have the feature that we must somehow sail between

conveying the wrong content (here, of desire, there, of obligation) and of conveying the right sort of content overspecifically. This suggests that we regard 'W', like 'O', as doing only part of the want-ascribing element required, the other part being supplied again by a subjunctivising element which 'activates' W at the appropriate point in the sentence, writing, as the reader will no doubt have anticipated:

(22) W$\exists x$(x is a communist spy $\wedge$ S(John catches x)).

Here, as in the deontic case examined above, 'x is a communist spy' is not said to be part of what is wanted, though it is part of what is required for the correct expression of that is wanted. That desire and subjunctivity have some sort of affinity has been noted by Castañeda and by Quine.[15] It is particularly evident when the predicate to which the subjunctivity attaches is expressed with the aid of a verb, for then it will show up in the surface structures of subjunctive-using languages as a subjunctive verbal inflection. A pair of examples which illustrate rather well the differences in role between subjunctive and indicative parts of a want-ascription are the Spanish examples (23) and (24):

(23) Me gustaría que Federico trabajara para un hombre que vive en esta casa.

[= I would like Fred to work for a man who lives ('vive'—*indicative*) in this house.]

(24) Me gustaría que Federico trabajara para un hombre que pague bien.

[= I would like Fred to work for a man who pays ('pague'—*subjunctive*) well.]

We are to imagine, for (23), that the speaker and his interlocutor are just passing a certain house, the house where the man and the speaker wants Fred to work for happens to live, whereas, for (24), that the speaker is expressing his concern that Fred should not be underpaid. It is not hard to see what is going on here: in (24), 'pague' is subjunctive, because that the man in question should pay well is part of the content of the desire, just as much as Fred's working for him ('trabajara'—subjunctive in both sentences), whereas in (23), 'vive' is indicative because it is no part of the speaker's desire that Fred's employer should live in the house he in fact does live in. ('In fact' here should remind us of the availability of an L_A-style alternative to the present Castañeda-style treatment: the reader will be able to construct these alternative representations.[16])

For the ascription of desire, then, the standard approach must be modified in that 'W' requires supplementation by 'S'. One could do this for belief also, since any notion for which a quantifying-over worlds semantics is possible can be treated after the manner of Section 2, but since the belief operator would never need to appear other that immediately before its activator there would arise the serious question as to whether there really were two operators involved here. Thus it would be misleading in the extreme to use the duplex notation so appropriate for desire.[17] The same combinatorial possibilities are just not open, as we have already seen in the non existence of an F.B.I. sort of problem for belief, and as we may confirm by the following considerations concerning ambiguity.

(25) Tom thinks is taller than he is.

(26) Tom wants to be taller than he is.

There are two senses for (25): we weed out the contradictory-belief reading by inserting 'actually' between the last two words (or introducing quantification over heights, and scope-shuffling). But (26) is three-ways ambiguous, there remaining two senses after the contradictory-desire reading is eliminated – again, for example, by insertion of 'actually', showing itself to be something of an all-round inhibitor here.[18] In one sense, we are saying there is some height which Tom wants to exceed; to see the other sense, suppose that Tom has merely overheard some girls saying that he's too short to be of interest to them, and that he knows neither how tall he is nor how tall he would have to be for them to find him attractive: he just wants to be taller than he is. It is not true, in this case, however, that there is some height he wants to exceed, since to say this is to fall into over specificity, so that here we have our third sense, illustrating that extra complexity desire-attributions have over ascriptions of belief. There is much more that can be said about the two sorts of propositional attitudes, and to say it, one would certainly need to have control of the points about expressive weaknesses and how to remedy them made in this paper; but one would also need to give a substantive account of belief, desire, and the relation between them – a relation, in my view, rather more intimate than has been suspected.

NOTES

1. For further discussions bearing on the logic of "actually", see Åqvist [1973] and Davies [1976]; for a suggestion with a rather different motivation, Woolhouse [1975], esp. pp. 217–220; more recently, various papers have been written which overlap the subject by E. Saarinen – see his [1977]. For more on "now", consult Prior [1968]. The example about the shiny red things used in this paragraph was originally suggested to me by a similarly problematic tense-logical example I heard from Christopher Peacocke some years ago. I have made a number of revisions to earlier drafts of this paper as a result of helpful comments from A. P. Hazen and from M. K. Davies.

2. Or equivalently, $\Diamond(J \wedge \alpha)$. The idea of using such a propositional constant is due to C. A. Meredith (see Meredith and Prior [1965]). Adoption of Meredith's axioms, however, would lead to the propositional constant formulation of the modal analogue of Kamp's treatment of "now" (in Kamp [1971]) rather than the treatment presented in Crossley and Humberstone [1977]. In the former 'Actually α' and α are provably equivalent and the rule of necessitation cannot apply unrestrictedly, while in the latter case this equivalence does not hold and necessitation is unconstrained.

3. See Castañeda [1967*a*], p. 18, and [1967], p. 28.

4. Of course nobody would maintain that contexts like these are the sole, or even the central, occasions for the use of the subjunctive in Spanish, even if attention is restricted to subordinate clause subjunctives.

5. *Cf*. Hazen [1976], p. 41: "The actuality operator can (. . .) exempt clauses about the actual world from the influence of modal operators within whose scopes they lie".

6. Or rather, what is proved there is the completeness of these schemata together with $A(A\alpha \rightarrow \alpha)$; this last schema, however, is derivable from (A1)–(A4).

7. The only reason for avoiding uniform substitution is that it would not, if applied generally, preserve well-formedness (because of the unusual formation rules of L_C); there would be no objection to using the rule if it were appropriately formulated.

8. Since our underlying modal logic is S5 here, the present schema could be weakened by requiring that α be purely truth-functional, rather than merely S-free.

9. Castañeda's work in deontic logic has been unjustly ignored for this reason. For example, in one of the few professional discussions of this work (Powers [1967]), L. Powers more or less complains at the triviality of Castañeda's innovations at the propositional level, failing to observe the genuine pay-off in terms of expressive power these innovations bring in the full setting of a first-order language.

10. The point about outer occurrences idling applies also to the deontic logic sketched in the final paragraph of Section 2.

11. For brief discussion and references see Lewis [1973], pp. 63*f*. A more general discussion may be found in Segerberg [1973].

12. This is occasionally referred to as an 'actualist' interpretation of the quantifiers; we avoid the term here because it may suggest that the range of any quantifier is to be the actual world (w^*) of the model. (Ideally, one would like a word bearing the same relation to 'actual' as 'current' bears to 'present'.) This is the usual, and philosophically preferable, interpretation of the quantifiers.

13. This is the approach outlined by Quine in the first section of his [1956], and which may be found in various of Prior's writings, as indeed in countless other places. (Quine's criticisms of the approach and his worries about quantification into intensional contexts have nothing to do with the present criticism, and the point made here about 'wants that' will remain however such worries are to be settled.)

14. I am unable to track down the source of this objection, cast in terms of this example, but I am sure I have seen it somewhere in the literature. [[See 'Updates and Afterthoughts'.]]

15. For example, in the first section of Quine [1956], Quine observes that Spanish resolves, by means of the indicative/subjunctive contrast, the ambiguity of 'I am looking for a dog that talks'. (What he does not notice is that, as the F.B.I. example shows, not every similarly resolvable ambiguity can be treated – as he suggests the talking dog example may be – as a simple scope ambiguity.)

16. Those representations will not read very plausible on a symbol-by-symbol rendering into English, I'm afraid, suffering in this respect as does (4) of Section 2: the explicit conjunction here is better replaced by a relative clause construction.

17. This may be why desire contexts have a 'subjunctive feel' about them not possessed by belief contexts.

18. The same versatility in inhibitive effect is evidenced by the ambiguity of 'The builder said that the plumber believed that the pipe was longer than it actually was'. (I might mention here that further work on these matters would need to consider the simultaneous presence of alethic and deontic concepts, and of 'believes' and 'wants' in one language. It may, for all I know, turn

out that an L_A style treatment is better for some cases, and an L_C treatment for others, so that a language containing both inhibitors and activators was required.)

19. In the Appendix to this paper, as originally published, an adaptation of the familiar 'canonical model' style of argument was given to prove completeness for the system S5C as axiomatized in Section 2. Castañeda sketches a completeness argument for an axiom system of his own in Castañeda [1972]; the considerable differences in the formation rules for his language and mine (to say nothing of the fact that the $\models/\Vdash$ distinction is not employed by Castañeda) make a comparison between that proof of completeness and the present one difficult and unrewarding. In addition, there appear to be certain technical errors in Castañeda's presentation, as I indicate in my review (Humberstone [1975]) of Castañeda [1972] – or rather of the volume in which that paper appeared.

'Scope and Subjunctivity': Updates and Afterthoughts

As is clear from its content, the paper forming this chapter is very much a sequel to Crossley and Humberstone [1977], where what is here called the 'Actually' language is developed. Although the latter paper is co-authored, the main ideas—and certainly all of those used here—were worked out by me as a graduate student at Oxford, and presented in a paper I gave to a discussion group there in 1974. For confirmation of this, I refer the reader to footnote 5 on p. 24 of Forbes [1989]; Forbes was in the audience at the time. (What Professor Crossley did was to work out a completeness proof for the axioms whose completeness was in my earlier work only conjectured; the technique is fairly routine but at the time we began the joint paper – in 1975, on my arrival at Monash, where the earlier version was delivered at a staff seminar attended by Crossley – I was not familiar with the details of this technique.) The references to the logic of "actually" given in note 1 should be supplemented by one to a published descendant of Martin Davies' Oxford D. Phil. thesis (Davies [1976]), namely Davies [1981], as well as by our joint paper Davies and Humberstone [1980].

It has been a source of great *chagrin* to me that 'Scope and Subjunctivity', as printed in *Philosophia*, was so full of typographical errors that parts of it were virtually unintelligible; the original page proofs were in an even more seriously garbled state, and I was (of course) not given a chance to review the corrected (!) proofs. This may partially account for the fact that in the ensuing years several people have proclaimed as new observations points which can be found in the original piece, given enough patience to work past the typos. The most frequently rediscovered such point is the observation (in the second paragraph of Section 1) that the use of an actuality operator can be avoided, in the representation of 'it is possible for everything which is (actually) φ to be ψ', if instead we quantify over sets of individuals, since we may then write:

$$\exists s[\forall x(x \in s \leftrightarrow \varphi x) \wedge \Diamond \forall x(x \in s \rightarrow \psi x)]$$

Exactly the same point turns up, I was surprised to find, in Teichmann [1990] – surprised because he there cites Crossley and Humberstone [1977] (though he cites it as being by J. L. Crossley and I. M. Humberstone, thus managing to the initials wrong for both authors) and thinks that the availability of such a representation can in some way be used to cast aspersions

on the 'Actually' language. A somewhat different example is used to make the same point in Bostock [1988]; although Bostock discusses in some detail Davies and Humberstone [1980], he does not refer to 'Scope and Subjunctivity'. A variation on the alternative representation would use, not quantification over sets, but distinctively plural quantification over individuals, to make a similar point. This can be found on p. 93 of Forbes [1989], where again I was disappointed – though perhaps hardly surprised – to find no mention of 'Scope and Subjunctivity'. A similar proposal may be found in Bricker [1989]. (A comparative discussion of this proposal with the 'actually' treatment is provided by Chapter 6 of Cresswell [1990].)

Wehmeier [2000*a*], recently sent to me by its author, has, by contrast, grasped the main points of 'Scope and Subjunctivity' in spite of the typographical difficulties, and discusses an argument of Kripke's about the non-synonymy of proper names with definite descriptions with the aid of similar apparatus. He shows how, at least as standardly presented, the argument fudges indicative/subjunctive contrasts present in the definite descriptions concerned. Here we consider only his departure from the apparatus of 'Scope and Subjunctivity'. The relevant passage is quoted from p. 9 of Wehmeier's typescript (though in quoting I have altered the boldface **S** Wehmeier uses to a regular S for conformity with our earlier notation):

> The obvious way to go would be to use Humberstone's (1982) system with a sentential subjunctive operator "S" (…) I should like to propose a different solution here, essentially for the following reasons: Sentential operators can usually be iterated, but Humberstone's semantics cannot accommodate iterations of 'S' (and consequently, he does not count expressions with iterated subjunctivity operators as well-formed). From the point of view of natural language, this is of course no loss, since there is no such thing as iterated subjunctivity; still, one wonders whether, under such conditions, a sentential operator is the adequate formalisation. Similarly, his semantics cannot handle possibility operators within the scope of 'S'. Again, the exclusion of such expressions from the class of formulae is not objectionable as such, but the necessity of this measure suggests that subjunctivity does not work like a sentential operator.

Wehmeier goes on to sketch his preferred alternative, according to which formulas are built from atomic predicates of two types: indicative and subjunctive (written as F, say, for an indicative predicate letter, and F^* for the corresponding subjunctive predicate letter), and in addition an indicative and a subjunctive existential quantifier ("$\exists$" and "$\exists^*$" respectively). The idea of drawing the subjunctive/indicative distinction at the level of atomic predicates lies behind Castañeda's original language and one of the departures I was conscious of making in passing from it to our "Castañeda-inspired" language L_C was in drawing the distinction instead at the level of arbitrary formulas rather than of arbitrary formulas: indicative unless explicitly subjunctivized with the operator 'S', so as to present the neat correspondence with L_A which has the opposite presumption: subjunctive unless explicitly indicativized with the operator 'A'. But Wehmeier's thoughtful remarks, quoted above, suggest that the matter merits some reconsideration. (A later version of this material, Wehmeier [2000*b*], is somewhat differently organized, and the passage most closely corresponding to that quoted above appears in this version on p. 21*f*.)

Accordingly, let us review the facts that make possible such a dramatic restructuring of L_C. In fact, it is preferable to abstract from that particular case so that we can also address the situation of L_A. For this reason, rather than writing 'S' in the following equivalences, we write 'O' (for 'Operator'), with a view to considering their status not only for the case of 'S' but also for the case of 'A'. We use '≡' to indicate that the formula on the left and the formula on the right are fully interchangeable *salva provabilitate* or *salva validitate* w.r.t. some unspecified proof-system or semantical apparatus. (We are especially interested in the semantical apparatus associated in 'Scope and Subjunctivity' with the languages L_A and L_C, of course.) "φ", possibly subscripted, is a schematic letter for arbitrary formulas. Equivalences 1–3 pertain to the propositional languages on which we concentrated the formal developments in 'Scope and Subjunctivity' (though see Section 3 for some steps beyond this area), and 4 is added because of the interest of the associated quantificational languages, though we leave the right-hand side blank for the moment:

(1) $O\#(\varphi_1,\ldots,\varphi_n) \equiv \#(O\varphi_1,\ldots,O\varphi_n)$ *for each n-ary boolean connective* #.

(2) $OO\varphi \equiv O\varphi$

(3) $O\Box\varphi \equiv \Box\varphi$

(4) $O\exists x\varphi \equiv$ ___

The general question as to when the operator represented by O can be replaced without loss of expressive power by the device of special-purpose atomic formulas which are to be equivalent to those of the form Oφ for atomic φ receives an affirmative answer for the propositional languages if any formula containing as a subformula one of the forms on the left of (1)–(3) can be replaced by the corresponding formula on the left, since these equivalences allow us – by an induction on formula complexity – to transform any formula into an equivalent formula in all occurrences of 'O' are pre-atomic. In the case of (propositional) L_C and taking O as 'S', we have equivalence (1), this being another way of putting the point made in 'Scope and Subjunctivity' in Kripke-semantical terms by describing this as a monocosmic operator. So non-pre-atomic occurrences of 'S' can be driven inwards across boolean connectives. (2) and (3) allow us to drop an occurrence of 'O' with another such occurrence in its immediate scope or with an occurrence of '□' there, again helping to show that only pre-atomic occurrences need to be contemplated. In L_C, though, because of its unusual formation rules, the cases presented by the left-hand sides of (2) and (3) cannot arise in the first place. For (4), we need to take special measures to fill the blank, and, as already remarked, Wehmeier introduces for this very purposes a new primitive subjunctive quantifier (this being a feature absent from Castañeda's treatment), filling the blank with "$\exists^* x\varphi$". Thus with this single innovation, we can indeed say that every formula has an equivalent in which the only occurrences of 'S' are pre-atomic, and therefore replace, as Wehmeier does, such occurrences by trading in 'S($Ft_1\ldots t_n$)' with n-ary F and terms (individual constants or variables) $t_1,\ldots,t_n$, for the new atomic formulas $F^*t_1\ldots t_n$. It is worth noting that tense and mood behave differently in this respect, since there would, even with tensed quantifiers added, be a reduction in the expressive power of Prior-style tense logic if only tensed atomic predicates were allowed and not scope-bearing sentential tense operators. This is because of the failure, in, for the sake of simplicity, just the language with the past tense operator

P, equivalence (1) above: we will be able, with past-tense applied only to atomic predicates, be able to say that at some earlier time *a* was *F* and at some earlier time *b* was *F* (P*Fa* ∧ P*Fb*) but not that at some earlier time *a* and *b* were both *F* (which in Prior's language would be expressed by P(*Fa* ∧ *Fb*).

It is also worth noting – and this is why we formulated (1)–(4) in schematic terms with 'O' – that a similar move is possible with quantified L_A, provided an additional quantifier, say, "$\exists^{@}x$" (quantifying existentially over the 'actual' world in the model in question). (1) still holds, and this time the left-hand sides of (2) and (3) do arise as well-formed but since, these equivalences also hold, they can be replaced by their right-hand sides until all occurrences of 'A' are pre-atomic unless they precede a quantifier as in (4), in which case the subformula in question is replaced by "$\exists^{@}x\varphi$". Now, as in L_A itself, the presumption is of subjunctivity unless there is explicit 'indicativization', so a plain predicate letter '*F*' is understood in terms of the world relative to which the formula containing it is evaluated, and the supplanting predicates – suppose that we write the indicativized version of '*F*' as '$F^{@}$' – are introduced as before to replace 'S($Ft_1 \ldots t_n$)' with '$F^{@}t_1 \ldots t_n$'. The main point of 'Scope and Subjunctivity' remains intact – that we can achieve the same effect either by making the default interpretation indicative and subjunctivizing to override this, or by making the default interpretation subjunctive and indicativizing to override this. Indeed, the point is rather clearer in this setting, in view of the fact, noted by Wehmeier, that the syntax of the original languages L_A and L_C was not quite parallel because of the restrictive formation rules of the latter language.

(It may all the same be worth giving some attention to a potentially disquieting general feature of the above line of thought. As a general rule, it is not a good idea to argue against a more comprehensive language – such as one in which 'S' (or 'A') can operate on an arbitrary formula and not just (in effect) on an atomic formula – and against a less comprehensive one, on the basis of an observation that everything that can be said in the richer language has an equivalent in the poorer, since if we work only with the poorer language, we can no longer formulate the observation in question. The point is from Smiley [1996], and is further applied in §3 of Humberstone [2000].)

The discussion in Section 4 of 'Scope and Subjunctivity' oversimplifies when it claims that nothing like the scope phenomenon presented by such cases as the FBI example (17) of that section for desire arises for belief. (Incidentally, the source of this example, which I mentioned in note 14 having seen in the literature but been unable to trace, is: Prior [1971], pp. 137*f*.) Dummett [1973*a*], p. 272, mentions an example which he attributes (without citing a published source) to Geach ("following Russell"): "There are more people here than I expected." Of this example, Dummett says:

> This means 'For some *n* there are more than *n* people here and I expected that there would be at most *n* people here', and certainly not, 'I expected that, for some *n*, there would be more than *n* people here and there would be at most *n* people here'.

But it is far from clear that the sentence means what Dummett says it means (though it certainly doesn't mean what he says he certainly doesn't mean), especially as he goes on to clarify that the variable '*n*' ranges over numbers. (Dummett's discussion focuses on the worry that to

appear in both an opaque and a transparent context in the above regimentation, and thus from a Fregean point of view, to require to be taken on one occurrence as occupying the position of a term whose reference is a number and at the other as occupying the position of a term whose reference is the sense of such a term.) The following objection to Dummett's claim may well appear somewhere in the literature, but I have not been able to track it down, if it does. One could surely be surprised at the size of the crowd and therefore truly say "There are more people here than I expected" without there being any definite number of people concerning which one expected there to be at most that number of people, whilst there were in fact more. For one may only have had a vague specification ("a few") in mind as to the expected number, a vague specification which is nevertheless clearly exceeded by the number that turned up.

Because of its apparent involvement with issues of vagueness, it is difficult to know what to make of the above example. (In taking it as an example of the 'belief' rather than 'desire' type, we note expectation falls on the 'thetic' rather than the 'telic' side of the fundamental distinction between propositional attitudes discussed in Chapter 6 below.) However, there are cases in which this awkward feature of Dummett's example is absent, which seem to point to the presence for belief-attributions of the complexities emphasized in Section 4 for desire-attributions. Heny [1970] presents such an example, which he describes (p. 116) as a case of "the transparent interpretation of a non-specific phrase within a potentially opaque context". The case has some distracting complications, but here it is, as presented by Heny:

> If Alice in her dream believed in the existence of what we may call "playing-card-people", that is, entities somewhat like playing cards but able to act like men, and which the speaker knows in reality to be playing cards, then he is perfectly justified in reporting Alice's dream thus:
>
> (56) Alice believed that some playing cards were going to hurt her.
>
> If Alice has not idea which playing-card-men were likely to injure her, the indefinite NP must be interpreted non-specifically for the sentence to be true, and yet it must be interpreted transparently too. These sentences make it quite clear that specificity cannot be identified with either opacity or transparency. And transparency is not just wide scope.

For our purposes, it is the final remark which is relevant, though we need not exercise ourselves over exactly what Heny understands by 'transparency' here. What is confusing about the example is the suggestion that there are amongst the denizen's of Alice's dream things which the speaker (of (56), that is) "knows in reality to be playing cards": what *can* this mean? However, suppose that Alice was not dreaming and believed that at least one of a bunch of objects, which are in fact playing cards though believed by her not to be, was going to hurt her. Further, to avoid a complication about tense and quantification, let us think of a present tense way of describing her state of mind would have been as of the time in question. It cannot be (*) or (**):

(*) $\exists x$(x is a playing card $\wedge$ Alice believes(x is going to hurt Alice))

(**) Alice believes $\exists x$(x is a playing card $\wedge$ x is going to hurt Alice)

(*) is false since there is no particular individual, playing card or not, that Alice believes is going to hurt her, and (**) is false because Alice does not believe that playing cards are dangerous

and so does not believe that at least one playing card is going to hurt her. It would seem than that here we have a case for either the 'Actually' style of treatment – put the "*x* is a playing card" part of (**) into the scope of an occurrence of 'A' – or a subjunctive style of treatment – put the "*x* is going to hurt Alice" part of (**) into the scope of 'S', in each case reconstruing the unembedded parts of the representations along the lines of L_A and L_C as explained in Section 2 of 'Scope and Subjunctivity', respectively. The two complementary procedures we stressed in the paper for desire-ascriptions do after all seem to present themselves for service no less in the case of belief-ascriptions, contrary to the tenor of the remarks in Section 4. We return to this issue in the 'Updates and Afterthoughts' section of the following chapter.

Finally, it should have been mentioned that the nomenclature of 'S5A' *vs*. 'S5C' for the "actually" or "A"-equipped and the Castañeda-inspired (i.e., "S"-equipped) versions of S5 clashes with nomenclature used elsewhere. In particular Porte [1981] uses 'S5C' for a certain *Carnap*-inspired version of S5.

Some of the material in these Updates and Afterthoughts, especially regarding the work of Wehmeier, has been included (in a modified form) in Humberstone [2004].

From: *Canadian Journal of Philosophy* **17** (1987), 49–62. Notes begin on p. 38.

CHAPTER 2. WANTING AS BELIEVING

1. Preliminaries

An account of desire as a species of belief may owe its appeal to the details of its proposal as to precisely what sort of beliefs desires are to be identified with, and its downfall may be due to those details it does provide. For example, it may be proposed that the desire that α is in fact the belief that it ought to be that α, or is morally good or desirable that it should be the case that α.[1] Here the appeal might be that of forging a link between the holding of a moral belief and the acknowledgment that one has a reason for acting a certain way; and the shortcoming of the suggestion is its evident implausibility: even if the 'necessity' direction could be established, having a desire hardly seems sufficient for the holding of any such belief. Again: it might be proposed, perhaps simply to bring some order into the realm of propositional attitudes by reducing some to others, that the belief with which we should identify *a*'s desire that α is *a*'s belief that he will or would be happy if α. This proposed identification can be seen to be incorrect by consideration of examples such as the following, due to J. Gosling.[2] An aging and ailing parent might forego numerous pleasures in order that his children should reap the benefits of his saving and have a good start in life – perhaps by receiving an expensive education – after his impending death. Clearly he may want that they should so benefit even if he does not believe in an after-death existence in which he might come to know of, and so take pleasure in, his children's subsequent well-being. So his wanting that they should prosper cannot consist in his believing that he will be happy if/when they do, since he has no expectation of even being in existence in that eventuality. As Gosling puts it, there is a clear difference (illustrable with far less dramatic examples than this one) between thinking that something's coming about will bring one pleasure, on the one hand, and viewing the prospect of its coming about with pleasure, on the other.

In what follows, I too will be suggesting that desires are best construed as beliefs of a certain sort, while, in contrast, saying conspicuously little about what that sort is. Nor will this suggestion be urged as the expression of any special insight in philosophical psychology. It is, rather, from philosophical logic (broadly understood) that the inspiration for this account comes: though 'account' is too strong a word for it. For all that I shall try to show is that there is a case for regarding attributions of desire as ascriptions of belief, with '*a* wants that α' seen as having the form '*a* believes that $\_\_\alpha$', where the blank is filled by some intensional sentence operator, which I shall be writing as 'D'. I might mention that, just to have some way of reading this, I read 'D' as 'it is desirable that.' Such a reading must not be taken literally, of course, since we have already seen that the usual normative overtones of 'desirable' are quite out of place here: such a usage is merely suggestive, and there may be nothing left to the meaning of 'desirable' once such normative elements are subtracted. Indeed it would surprise me little if there turned out to be no word or phrase of English or any other natural language which can appropriately fill

the blank indicated, and even less if it could be shown that there was no way of explaining what my 'Dα' is supposed to amount to which did not already make use of the concept of desire. Thus there is a certain sense of 'reduction' in which it would be absurd to characterize the present enterprise as the reduction of desires to beliefs.

There is, however, one kind of context for 'D' in which a relatively natural paraphrase is available, if the arguments of the following section are successful in making the identification of a desire that α with a belief that Dα plausible. I am thinking of the case where 'D' occurs as the main (and sole) operator in a sentence, For here, given the identification, we may read 'Dα' as 'Would that it were that α'. The point is simply that for any β, the way for *a* to express the belief attributed to him by the sentence '*a* believes that β' is to say: β. So, *given* the identification, we take β as Dα. Such a reading is not generally available, though, since the 'would that' construction does not embed: but it does suggest that, if we wanted a word to describe 'D', we could call it an *optative* operator.[3] With this, we arrive at a point of close similarity between the present suggestion and one made by Kenny and developed by Hare.[4] Kenny's idea, roughly, was that to desire that a is to say in one's heart: *let it be that* α. Hare notes that if we take the italicised phrase imperatively, we get a proposal suited to the case of 'full-blooded' wants, as opposed to mere wishes, etc., while if we take it instead optatively we have something appropriate for wants in the catch-all general sense including both of these. It is in the latter, somewhat artificially inclusive sense that I use 'want' and 'desire' in this paper. If we take 'saying in one's heart' to mean simply *believing*, then we get the proposal I am making. It seems, though, that Kenny does not so take it, seeing each of belief and desire as composed out of saying in one's heart *plus* something else (an indicative or assertoric element in the former case). We shall see, in discussing the first argument in the following section, that there are reasons for preferring a less symmetric treatment: it is, specifically, the notion of desire that should be seen as composite, being composed out of belief and something else (namely, 'D').

2. Some Arguments for this Proposal

I shall argue for the composite treatment of ascriptions of desire described in I by presenting some examples of things we want to say but cannot represent in a language not making the compositional treatment explicit. The framework for this procedure is that of the 'modal' parsing of propositional attitude constructions: we write 'Wα' for '*a* (or some other specified individual) wants that α' and 'Bα' for '*a* believes that α, treating these as intensional operators forming sentences from sentences. Then the three representability arguments to be given are designed to show that certain sorts of propositional attitude ascriptions involving belief and desire cannot be expressed in this language unless it is supplemented by the operator introduced in I as 'D,' with 'BDα' and 'Wα' equivalent (thus rendering 'W' itself definable). Of course there are some well-known difficulties with this framework. In particular there are some who feel that the constructions involved are referentially opaque, and that this creates a problem about the intelligibility of quantifying into the scope of the operators concerned, a procedure I shall certainly be availing myself of below.[5] Further, the alleged hyperintensionality of these constructions is supposed to lead to difficulties in giving a systematic semantic account of the

framework, and that, more generally, the only way to press the analogy with modality is thought to embody unrealistically strong rationality assumptions (as in the usual systems of doxastic logic, etc.). I am going to ignore these difficulties because the problems we shall be looking at have nothing to do with opacity or hyperintensionality and will remain however they are to be overcome (if there is any overcoming be done).

It is interesting to note that according to one point of view, such problems – or at least those involving opacity and 'quantifying in' – are already met by the interpretation afforded by possible worlds semantics for the languages concerned, while according to another point of view the introduction of this apparatus serves instead to highlight the difficulties: the heuristic usefulness of the apparatus is thus hardly to be denied by either party. And although I shall not be bringing it into play here because a crisper statement of the position I am urging is possible without invoking possible worlds, I should like to say a few words on the subject before we get to the promised arguments. We are to understand the locutions involved in terms of universal quantification over a range of worlds: what is believed by *a* in *w* is what is the case in each world *w*' which is a doxastic alternative (for *a*) to *w*, while for what is wanted by *a* in *w* is what is the case in each (shall we say?) boulomaic alternative to *w*. The hypothesis I tentatively favour for the novel operator 'D' would be similar in requiring for the truth of Dα at *w* the truth of α at each of a range of worlds (those which are 'D-accessible', one might put it). In conjunction with the equivalence of Wα with BDα, this has the effect of requiring that the accessibility or alternativeness relation associated with W should be the relative product of those associated with 'B' and 'D.' Thus, unless restrictions are imposed on the models to the effect that each world has only one D-alternative, we have that in general D$\neg\alpha$ is a strictly stronger statement than $\neg$Dα (as well as that 'D' distributes over conjunction but not over disjunction). This means that there are more distinctions to be drawn than one might initially have thought, a point which we may illustrate by considering the concept of (boulomaic) indifference. In the standard framework with unanalyzed 'B' and 'W,' the subject's indifference as to whether or not α is represented thus: $\neg W\alpha \wedge \neg W\neg\alpha$. In the suggested replacement, with 'B' and 'D' (and 'W' as a defined symbol), this emerges as $\neg BD\alpha \wedge BD\neg\alpha$, and is to be distinguished from, in particular, the following: $B\neg D\alpha \wedge B\neg D\neg\alpha$ (equivalently: $B(\neg D\alpha \wedge \neg D\neg\alpha)$). The former may be the indifference of the person who has not made up his mind as to which possibility is the more attractive, while the latter is the indifference of the person who holds them to be equally attractive. Such notational multiplicity may be held to show the B-D proposal to be encumbered by an *embarras de richesses*. We are, it might be objected, accustomed to hearing such sentences as 'John doesn't want it to rain tomorrow' as two-ways ambiguous, but does the proposal not unfortunately predict a three-way ambiguity here? My reply would be, first, to repudiate the supposedly atheoretical conception of ambiguity with which the objector seems to be operating, and then, to suggest that the interesting question in this area is not: "What ambiguities of scope does a given English sentence present?", but rather: "What distinctions of scope are there to be drawn here?" We turn now to the promised arguments.

The first example concerns an unhappily married couple, John and Mary. So unhappy have they been lately, indeed, that Mary has actually left John and gone to stay with her sister, and, further, far from wanting her back, John's feelings for her remain steadfastly negative. Wary of his malicious streak, she has bought a large dog for protection, and as a result of hearing it

barking when he was in the vicinity of his sister-in-law's house (we do not ask why he was there), he has surmised as much. While John has in fact no intentions of trying anything, his ill feelings towards his estranged wife do combine with his having read in the paper of a possible outbreak of rabies in the area (and a little wishful thinking) to lead him to hold the attitude reported in (1):

(1) John thinks that Mary owns a mad dog, and he wants it to bite her.

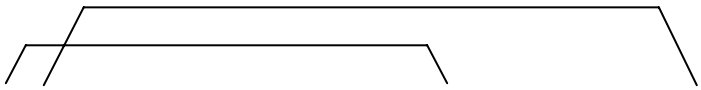
Perhaps slightly more natural under the circumstances would be '… and he hopes it will bite her', but let that pass.[6] Further, I take it—and you may modify the story slightly if you disagree—that we cannot represent (1) by:

(2) $\exists x$(B(x is a mad dog owned by Mary) $\wedge$ W(x bites Mary))

or by:

(3) $\exists x$(x is a mad dog $\wedge$ (B(Mary owns x) $\wedge$ W(x bites Mary)))

or indeed by anything else which has 'B' within the scope of the existential quantifier, since while John believes Mary to own a mad dog, there is not a particular dog he believes her to own. Accordingly, that quantifier must lie within the scope of 'B'. ('B' here for 'John believes that', of course.) But it must bind the 'x' in the scope of 'W,' and so must have the indicated occurrence of 'W' in its scope. Yet this is wrong: he doesn't just believe that he wants the dog to bite her, he actually wants it to. One's first reaction may be[7] to try and write down something like (4):



(4) B$\exists x$(x is a mad dog owned by Mary) $\wedge$ W(x bites Mary)

The idea is that the overhead ties extend from a scope-bearing element to the end of its scope. Now there is simply no way, within the received framework, of engineering such 'overlapping scope' representations. If we take 'W' out of the scope of 'B', we are left with the terminal occurrence of 'x' dangling idly free. So much the worse, one might conclude, for the received framework, which should somehow be liberalised so as to make something along the lines of (4) well-formed. Perhaps a detailed semantic account for a language containing such constructions could be given. But there is a good reason for not choosing this path: the resulting formalism would render not only (4), but also the result of switching 'B' and 'W' in (4) well-formed, in spite of the marked contrast in intelligibility between (1) and:

(5) John wants Mary to own a mad dog, and he thinks it will bite her.

The only way, as far as I can see, of making sense of (5) is to interpret what follows 'and' as 'believes that, if she does own a mad dog, it will bite her,' while no such conditional reinterpretation is required in the case of the original (1). So the real problem is not just to find a way of representing (1), but to find some way of doing so which does not simultaneously afford something analogously related to (5).

This problem is solved on the B-D proposal by a manoeuvre that proposal makes available and which works for a number of different sorts of case, which we might describe as: Divide and Insert. Here is the representation:

(6) $B\exists x$(x is a mad dog owned by Mary $\wedge$ D(x bites Mary))

The overlapping scope phenomenon has been dealt with by replacing the single operator 'W' by the pair 'B' and 'D' which acting in concert, do its work, while in being distinct allow for the interposition of an existential quantifier. Thus we are able to avoid attributing to John (as in (2) and (3)) attitudes of inappropriate specificity, without ending up ascribing to John a belief about a desire instead of a fully fledged desire. No analogous representation is thereby made available for (5), since it is desire and not belief that is treated as the composite operator. Incidentally, note that we would lose this advantage if we decided to treat *both* operators as composite: this appears to be a ground, as promised at the end of §1, for preferring the present asymmetrical account over the more even-handed suggestion we found in Kenny's work.[8] There is more to be said than I shall say here about this first argument for the proposal. I must just mention, though, one objection to this argument which will probably have occurred to some readers. Geach has raised a superficially similar problem about – as he calls it – intentional identity[9] with a case too well known to repeat here involving such identity across the propositional attitudes of different attitude holders, rather than, as with (1), across the attitudes of a single person. His case might be held to raise the same problem but, unfortunately, to resist solution on similar lines. I cannot discuss this objection here since I have resolved to keep the formal semantics in the background and it really is needed here to draw a distinction (with regard to what might be called the relative specificity of the attitudes) between the two kinds of cases. But I agree with the methodological stance of the imagined objector: if we were dealing with the same problem in the two cases (which I hold not to be the case) and my solution did not adapt to Geach's case (as it certainly does not), then my solution would be wrong for my case.

Our second argument does not involve quantification, as will no doubt come as a relief to those who suspect (and I cannot evaluate this suspicion) that the root of the difficulties over (1) lies in the general inappropriateness of the *Existential quantifier + bound variable* apparatus for the representation of natural language *indefinite noun phrase + anaphoric pronoun* constructions, a disquiet coming increasingly to be expressed.[10] We shall be interested instead in a particular kind of conditional desire ascription, whose peculiarity is that it could only misleadingly described either as the ascription of a conditional desire or as the conditional ascription of a desire. Since some members of audiences to which I have presented this material before are apt to produce blank stares when the crucial example is advanced, I shall have to pay careful attention to getting the stage-setting right to ensure that the example is intelligible.

Begin by casting your mind back to the final months of Richard Nixon's presidency. Rumours were in wide circulation, from different sources to the effect (i) that he was a witting party to certain criminal actions of a politically motivated nature, or at least to the subsequent attempt to thwart the course of justice by engineering a cover-up of those activities, and (ii) that he was suffering from a potentially serious illness. Now during this period, suppose, a certain elderly lady called Hermione comes to hear of these stories. She knows neither whether the suspicions of his misconduct nor the rumours about his health are correct, but she does think that if he really

is ill, then this can only be because he has in fact done what he is alleged to have done and his illness has been visited upon him by God as a punishment. Further, since such skullduggery really would have been, had it occurred, quite reprehensible in Hermione's opinion, she would be rather glad to see this particular stroke of divine intervention. In these circumstances, we could say something like this: she would like him to be ill, should he in fact be ill; or:

(7) Hermione wants, if Nixon is ill, that he should be ill.

How, writing '*p*' for 'Nixon is ill', and using 'W' for 'Hermione wants that', are we to represent (7)? There are only two options:

(8) $p \rightarrow Wp$

and

(9) $W(p \rightarrow p)$

(I am thinking of '$\rightarrow$' as material implication but you may read it as your favoured conditional without jeopardy to the argument.) Now (8) is obviously not what is called for. For suppose that, unbeknownst to Hermione, Nixon is in fact ill. It would then follow by *modus ponens* that she wants that he should be ill (i.e., she wants him to be ill: this is of course not a moral 'should'). However, in these circumstances she simply does not have any such categorical desire, even though she would have it if she discovered about his illness and underwent no change of heart. On the other hand (9) won't do either: I have not had to go into the details of Hermione's psychology to this extent in order to justify attributing to her any such tautologous desire.

It is quite difficult to characterize the respect in which she does not have the trivial desire that if *p* then *p*, because we have to contend with the fact that nothing Hermione could learn about Nixon's health could disappoint her. But it seems a mark of nontriviality that a representation of her state of mind must be logically efficacious in ensuring the consequences that she desires outright that *p* when coupled with the statement that she believes that *p* – this being another way of making the point just made in cross-temporal terms about what she would want if she were to discover that *p* without undergoing a change of heart - and no such desire as the desire that if *p* then *p* could be so efficacious. To sum up, then: if all we have to play with are 'W' and '$\rightarrow$', then we must choose between (9), which wrongly attributes a trivial desire to Hermione, and (8), which attributes to her, not the conditional desire in question, but the absolute desire that *p*, should the relevant condition be fulfilled.

It is time to announce the solution. I think the representation should look like this:

(10) $B(p \rightarrow Dp)$

where, needless to say, 'B' is for 'Hermione believes that'. The representation is efficacious in the way the last paragraph suggests is required, because (11) is a correct principle for rational belief:

(11) $B(\alpha \rightarrow \beta) \rightarrow (B\alpha \rightarrow B\beta)$

and we take '*p*' for α, 'Dp' for β. But, not suffering from the fault we found in (8), there is no inference from (10) and '*p*,' tout court, to 'BDp' (alias 'Wp'). Most of all, though, doesn't (10)

seen to get things just right? Isn't Hermione precisely, when she turns her attention to the problematic attitude, thinking to herself: If Nixon is ill, well that's fine by me? The problem has gone: so ends our second argument for the B-D treatment.

In connection with this second argument, I know of no interesting objections, beyond the reaction I've already mentioned to the effect that the key example (7) doesn't make sense, a reaction I've been doing my best to avert. There is, however, an additional point of interest which I think can at least be mentioned without going into the semantic details I am suppressing throughout this paper. Before giving (7), I prefaced it with a gloss using the words 'in fact': Hermione wants Nixon to be ill, should he in fact be ill. This seems a natural enough way of putting the matter: but what rôle are these words 'in fact' playing here? This phrase, like the word 'actually,' often functions, when in the scope of an intensional operator O to indicate that what follows is not itself to be taken as lying in the scope of 'O' from a semantic point of view.[11] When an occurrence of such a scope-immunising device lies within the scope of one intensional operator, 'O_1' itself within the scope of another, 'O_2', we get an ambiguity according as we take the material it immunises as outside the scope merely of the inner 'O_1' or as right outside even the outer '0_2' (as in – to adapt an example of Russell's – 'Betty claimed that Harry thought that your yacht was longer than it actually is'). Now, if we write 'A' for 'actually' (or 'in fact'), what the gloss given before (7) suggests is: W(A$p \rightarrow p$). And since the only operator appearing with 'A' in its scope is 'W', the only exscopating manoeuvre available yields something equivalent to the unwanted (8).[12] What has gone wrong here? I suggest that the 'A' is trying to get half way out of the scope of 'W', rather than completely out. This makes sense, of course, only if 'W' is seen as disguisedly a structured operator, as indeed it is on the B-D account. We are in the situation just described schematically with 'O_1' and 'O_2', these being, in the present case, 'D' and 'B', respectively. The intended re-scoping of W(A$p \rightarrow p$), or as we may more perspicuously write it, 'BD(A$p \rightarrow p$)' is not '$p \rightarrow$BDp' but rather – just having the 'Ap' hop over the inner 'D' – 'B($p \rightarrow$ Dp),' i.e., (10). Someone who maintained the traditional view that 'W' was a single unstructured operator would have to find it puzzling that the words 'in fact' occur in contexts like this at all.

The third and last of our representability arguments for the proposed dissection of 'W' into 'B' and 'D' involves the reasons people want things. I give up all pretence of conducting the discussion against the background of a formal language for propositional attitude ascription of the customary sort, which I hold to be incapable of expressing certain ascriptions, since we need to use the word 'because' essentially in the examples, and I am not familiar with a formal theory of that connective. Still, for abbreviative purposes and in the interests of making scope distinctions easily visible, we can agree to write 'α because β' as treating "$\because$" as a binary sentence connective. We are familiar with two ways of taking 'because' clauses appended to want-ascriptions: on the one hand to supply an explanation for why the attitude-holder has the desire, and on the other, to give his reasons (or grounds) for the desire. The distinction we normally mark with some such words as these, however, is not one easily recorded in the language containing the unitary 'W' operator alongside 'B'. For let it be desired by *a* that *p*, *a*'s grounds being that *q*. Then we could consider the following family of candidates:

(12) W$p \because q$

(13) W(p ∴ q)

(14) Wp ∴ Bq

(15) W(p ∴ Bq)

In (12) and (14), we have explanatory reports on desires, rather than reports on desires together with their reasons, considered as internal to the individual's attitudes. This is clear with (12), but is also the case with (14): perhaps having a certain belief will explain someone's having a desire, without its content being any part of what he would cite in answer to the question 'Why do you want that?' (i.e., as one normally puts it, this will be part of the reason he has the desire but no part of his reason for having it). As a report on the situation schematically described, (15), too, is considerably off target, unless something very strange is going on. Suppose a wants his children to wear seat-belts in the car because driving without them is dangerous. Then he may indeed have the additional desire that they should wear seat-belts because he believes not to do so is dangerous, where this is understood to have him desiring that that conduct is not only that required by safety, but also that it is caused by his having this belief. But this is an extra desire and not the original desire together with its grounds. In fact the situation is similar with (13) since here too we have '∴' in the scope of 'W' so that what is said to be wanted is that a certain causal connection should obtain. To illustrate the difference between this and the usual grounds-giving 'because' construction I am claiming to escape representation in this restricted language, suppose that a has a low opinion of the human race, and wants it to die out because it is inherently wicked. In reporting thus, we do not commit ourselves to the truth of 'The human race is inherently wicked.' Rather, we present this as *a*'s reason for wanting what he wants. It is quite possible that *a*, for all that has been said so far, would feel his desire would be satisfied if humanity became extinct as a result of a second ice-age, the earth being struck fortuitously by a shower of meteors, or whatever. What he wants is that the species should disappear, and when you ask why, he makes these remarks about inherent wickedness. Contrast *a*'s situation with that of *b*. What *b* wants is that the human race should become extinct, and his reason has again to do with its inherent wickedness, but, unlike *a*, *b* has a more specific desire - what he wants is only fully described in some such words as these: the human race dies out because of its wickedness. No cosmic accidents will satisfy this desire. A nuclear war would be good. Perhaps a shower of meteors, provided it was a piece of divine retribution. Evidently, it is a desire to this sort that is reported by (13).

So what becomes of the more straightforward case of a desire had for a reason, as with *a*? It will come as no surprise to hear that my suggested representation exploits once again the familiar Divide-and-insert strategy:

(16) B(Dp ∴ q)

Because we have the '∴' in the scope of 'B' we are dealing, as required, with reasons internal to the individual's attitudes, while because the first constituent of the '∴'-construction is 'Dp' and not 'p', we are not inadvertently turning a desire about what should be the case into one about why it should be. This concludes our presentation of the case for the B-D proposal by arguments about the inexpressibility of thought in the 'unitary W' language.

3. Closing Comments

I end by trying to settle some lingering qualms the proposal I am urging may have aroused, and making various other defensive remarks. The first concerns the very idea that desires might be beliefs. Surely, it will be protested, a reason for action must consist of both beliefs and desires: my proposal must be wrong because everybody knows that, by themselves, beliefs cannot motivate action. I want to address a word of reassurance to anyone worried on this score. The distinction between beliefs and desires survives somewhat transformed as the distinction between ordinary beliefs - as we might say - that is, those beliefs ascribable without 'D' in the scope of 'B,' and desiderative beliefs those not so ascribable. The doctrine that only desires can motivate then becomes the doctrine that only desiderative beliefs can motivate. I say the distinction survives *somewhat transformed*, because the desiderative beliefs will include not only those ascribed by reports of the form BDβ, which are the desires proper, but also those reported by statements of the form Bβ where β contains 'D' but is not equivalent to anything with 'D' as the main operator. This is the type of case illustrated by the three arguments of §2, and I take it as clear that such reports do describe considerations that would motivate the propositional attitude holder in question to action. One might say: yes, belief and desire are both involved in action, but so inextricably so that they cannot even be guaranteed to be separable constituents of agents' reasons for action.

Another sort of ground might provoke the same kind of fundamental resistance to the B-D proposal. If *a* wants it to rain (on a certain day) and *b* wants it to remain fine, so the objection runs, *a* and *b* have conflicting desires in the sense that not both sets of desires can be satisfied, but they surely need not have conflicting beliefs – where this means beliefs not all of which can be true together. Hence, it is concluded, desires cannot be beliefs. I am not sure that there really need be no conflict of belief in such a case, but a reply is available without pressing the point. Nothing I have said rules out taking 'D', just like, 'W' and 'B', as person-relative.

Finally, I want to mention an objection to which I am not quite sure what to say in response. The worry is that the thesis that desires are beliefs seems implausible when applied to non-linguistic animals.[13] Perhaps the idea is that in such cases the ascription of a belief is a rather hypothetical affair, a matter of what the creature would say if it could talk, while the ascription of a desire is a much more concrete business, a matter of how the creatures behaves in certain circumstances. I am not sure I agree with this picture. Certainly it may be that in the case of non-linguistic animals, we should be hard-pressed to find grounds for attributing the kind of propositional attitudes (with separated 'B' and 'D') which were used in §2 to argue for the thesis. If so, we may not be under the same pressure to hold that animals' desires were beliefs, though it would be very odd to hold that they definitely weren't, while human beings' desires definitely were. When it comes to the behaviour of non-human creatures, actual or possible, I suppose that the B-D account at least strongly suggests that even if there could be creatures capable of belief but incapable of desire, there could not be creatures capable of desire but incapable of (ordinary) belief. I think this suggestion is correct.[14]

NOTES.

1. A discussion of several philosophers who have taken such a line may be found in Stocker [1979].

2. Gosling [1969]; the example appears in Chapter 1.

3. Another handy way of reading 'Dα' is 'it would be nice if α', but this should not be taken as a conditional prediction of happiness or any other psychological state on the part of the person concerned, for the reasons already given in connection with Gosling's example. [[Note: I have moved the position in the text to which this footnote is appended. In the published version, the footnote flag appears at the end of the following sentence, for which note 4 is more appropriate. Note 4 appears amongst the footnotes in the published version, but there seems to be no flag to it in the main text.]]

4. See Kenny [1963], esp. Chapters 10 and 11, and Kenny [1966]. Hare's discussion is to be found in Hare [1971].

5. The *locus classicus* for qualms on this score is Quine [1956].

6. Recall that I use 'want' as the generic verb for desiderative attitudes. I might mention here that this example is only a slight variation on some provided by Pieter Seuren, who notes the expressibility problems involved, in Seuren [1977]. It was cases like this one, mentioned by Seuren in seminars at the University of Oxford in 1973, that first suggested the B-D analysis to me. The points made about this example, I might add, can also be made about the example get on replacing 'wants' (or "hopes") that' by 'fears that,' showing that a thesis like that maintained for desire in this paper can be argued on behalf of the affective (or 'conative') propositional attitudes more generally.

7. *Cf*. homework problem #19 in Kaplan [1973].

8. It might be said: surely the asymmetry in question amounts to no more than this – than one can have desires about whom one merely believes to exist but not beliefs about what one merely desires should exist. But such a remark still leaves open the question of how to represent the attitude-ascriptions we are interested in.

9. Geach [1967]. The most interesting discussion of Geach's problem to be found in the literature is that of Saarinen [1978], though I do not find myself in agreement with his conclusions. An example to the same effect as our (1) was given in an earlier discussion of Geach's paper by D.C. Dennett (Dennett [1968]), namely 'Tom thinks there is a bogey man in his bedroom and wishes he would go away.' Dennett's proposed way of representing this sentence has 'thinks' appearing (incorrectly) in the scope of 'wishes.'

10. See, for example, esp. sections 4 and 6 of Evans [1977]; as well as Kamp [1981]; and note also Hintikka's remarks at p. 201 of his contribution to the volume cited in which Kaplan [1973] (cited in note 7) appears, about indefinite descriptions and Hilbert's ε-terms.

11. For more details on this, see Humberstone [1982] [[= Chapter 1 of this thesis]]; want-ascriptions are discussed in Section 4. For present purposes, 'Actually' may be taken as a simple backward-looking operator in the sense of Saarinen [1978].

12. The notion of equivalence involved here is that called real-world equivalence in Crossley and Humberstone [1977]. The discussion in the present paper is oversimplified in order to get a point across briefly: it should not be concluded from what is said here that in general any formula has an 'A'-free equivalent (this fails, in particular, for quantified formulae). I plan to discuss in more detail the interplay between the B-D proposal and the material of the papers cited in this and the preceding note, in a separate piece on the fine structure of propositional attitude ascriptions. [[This records an intention never acted on, to try to reconstruct the original paper delivered to the Victorian Branch of the AAP, mentioned in the Foreword. Some comparative remarks on the treatments desire]]

13. A worry on this score was expressed to me by U.T. Place, on whose comments the view expressed in the following sentence is based.

14. This would be one reason for being sceptical about an analysis, asymmetrical like mine, but in the opposite direction, analysing beliefs as a species of desire. Such a proposal has actually been made, at p. 451 of Grandy [1973]. Grandy analyses beliefs as higher order desires (desires about desires), surely a kind of desire not possessable by all creatures sophisticated enough to hold beliefs.

'Wanting as Believing': Updates and Afterthoughts

Probably the most potentially serious threat to the main thesis of 'Wanting as Believing' is a style of argument, of which the prototype may be found in Lewis [1988], which purports to show that desires cannot be beliefs because the ways beliefs are rationally revised under the impact of new evidence leads, given orthodox decision theory and the hypothesis that desires are beliefs, to unacceptable results about the ways desires are rationally revised. As the reference to decision theory indicates, Lewis is working in a framework according to which beliefs and desires come in degrees—as subjective probability and expected value, respectively—rather than in the framework in which our work has been cast, where belief and desire (on the part of a given agent, with respect to a given proposition) are on/off matters. But Collins [1988] gives a non-quantitative analogue of the Lewis argument, in which belief and desire are on/off states, and shows that plausible assumptions about their revision likewise lead to trouble for the thesis that desires are beliefs. (One might think that Hume had shown, not simply that *not all* desires were beliefs, but that *no* desires were beliefs. I tried to disarm this kind of thought—which has been given a forceful formulation in Smith [1987]—with the opening paragraph of Section 3 of the paper; I will return to the point in Chapter 6 – see especially note 5 thereof and the text to which it is appended.) Although he works with non-quantitative notions of belief and desire, Collins uses the framework of (Gärdenfors-inspired) belief revision theory rather than that of intensional logic. A belief state is represented as a set K of possible worlds, and the result of revising that belief state so as to accept the proposition A (again thought of as a set of worlds) is denoted by K_A. This revision operation is subject to various conditions we will not go into here. Then there is also to be a set V of worlds containing each world such that the agent would be

satisfied if it were the actual world, and no other worlds. The relevant part of the Desire-as-Belief thesis (appearing in Collins' paper on p. 339 as a weakening of a stronger version of the thesis which it would take us too far afield to discuss) is then the hypothesis that for every proposition A there is a proposition Å such that for every belief state K:

(*) $K_A \subseteq V$ if and only if $K \subseteq$ Å.

The circle notation in 'Å' (read 'A halo') is Lewis's, by the way; it would clearly be a very cumbersome notation to employ when the expressions it is to sit on top of have any internal complexity, though these will not be needed here. For this to bear on the B-D proposal of 'Wanting as Believing', we are presumably to make the following connection: if A is the proposition expressed by a sentence α, then Å is the proposition expressed by Dα. Then writing $R_D(w)$ for the set of worlds to which a world w bears the accessibility relation associated with D (*cf.* the second paragraph of Section 2 of the paper), this proposition is $\{w|\ R_D(w) \subseteq A\}$; similarly, the proposition expressed by Wα is $\{w|\ R_W(w) \subseteq A\}$. The B-D proposal is then, in terms of this apparatus, that the latter proposition is the proposition that the former proposition is believed (by the subject in question), that is:

(**) $\{w|\ R_W(w) \subseteq A\} = \{u|\ R_B(u) \subseteq \{w|\ R_D(w) \subseteq A\}\}$

or, to put it into greater alignment with (*) above:

(***) $R_W(w) \subseteq A$ if and only if $R_B(w) \subseteq \{u|\ R_D(u) \subseteq A\}$

(We have changed variables to avoid double use of '*w*' as free and bound.) There is no free world variable—analogous to the '*w*' here—in (*), but we should think of fixing on some value of this variable, which can be thought of as the representing the world presumed actual, in '$R_B(w)$' and '$R_W(w)$' do get what Collins represents as K and V, respectively; as already noted our '$\{u|\ R_D(u) \subseteq A\}$' corresponds to his 'Å'. Translating (***) back, then, what we get is

(****) $V \subseteq A$ if and only if $K \subseteq$ Å

whose right had side is the same as (*)'s, but whose left-hand side is very different, most conspicuously in respect of what side of the inclusion sign the 'V' appears. Thus Collins [1988], in which trouble is made for (*), does not—despite his footnote citing of 'Wanting as Believing'—bear adversely (or otherwise) on the B-D proposal, whose closest approximation in his terms would be (****). Despite of my repeated pleas to Collins (on his several visits to

Monash) to clarify the situation, and show definitively either that because of such differences, the proposal is not vulnerable to any plausible variation on the attack of his [1988], or that in spite of them, it is thus vulnerable, no such clarification has been forthcoming. I turn therefore to another line of criticism.

Max Cresswell, in a paper discussing problems about anaphoric pronouns in and out of propositional attitude contexts (Cresswell [1990*a*]), devotes a section (§II) to the proposal of 'Wanting as Believing' insofar as it bears on these problems. After summarising the proposal as it applies to a variant on that given in Section 2 of the paper (formulation (6) in our numbering), Cresswell writes (p. 3):

> If Humberstone's analysis is to be plausible something must be said about the meaning of B and D. Suppose we take it that B is the belief operator. What about D? One of Humberstone's glosses of D is 'let it be that', but while this might be a helpful thing to say about what goes on when Susan sincerely *says* it, it doesn't really give truth conditions. In particular nothing that Humberstone says suggests that D has truth conditions any different from the 'wants' operator he is trying to analyse.
>
> I suggest that we look at D a little differently. Suppose that Susan wants an ice cream. I take her to a shop and point to some. She says
>
> (8) I want one of those.
>
> Unknown to Susan they are all plastic imitations. Does she really want one of them? There is surely a sense in which she does and a sense in which she does not. It is the sense in which she does not that D represents. For a world in which her desires are realized is *not* a world in which she holds in her hand a plastic imitation ice cream. She utters (8) only because she believes that those are real ice creams. Insofar as I can report her as wanting one of those, I must say something like
>
> (9) BD(Susan has one of those)
>
> What my remarks amount to is that although I am denying that wanting should be *analysed* using B and D, and therefore I am not claiming that desire is a species of belief, I think it quite plausible that many reports about what others want *are* like (9). Sentences about wants exhibit a structural ambiguity according as we are reporting real wants or believed wants.

First, let me raise some minor quibbles. I don't like the labelling, in the opening paragraph here, of the B-D proposal as an 'analysis', or the suggestion of a contrast with my position embodied in Cresswell's emphatic denial that wanting can be *analysed* using B and D. As I said in the first section of the paper, it would not surprise me at all "if it could be shown that there was no way of explaining what my 'Dα' is supposed to amount to which did not already make use of the concept of desire. Thus there is a certain sense of 'reduction' in which it would be absurd to characterize the present enterprise as the reduction of desires to beliefs." Reductive analysis is not to the point, then. (For more on what is at issue here, see the distinction between accounts of the application conditions of a concept, and analyses of that concept, in Chapter 7 below.) A

second quibble arises over Cresswell's claim that "nothing that Humberstone says suggests that D has truth conditions any different from the 'wants' operator he is trying to analyse." Yet in Section 2 of the paper, in the discussion following representation (3), I pointed out that while, to avoid overspecificity, the existential quantifier involved "must lie within the scope of 'B'. ('B' here for 'John believes that', of course.) But it must bind the '*x*' in the scope of 'W,' and so must have the indicated occurrence of 'W' in its scope. *Yet this is wrong: he doesn't just believe that he wants the dog to bite her, he actually wants it to.*" I have italicized the last sentence in this quoting of the earlier material in order to emphasize that I was assuming a distinction between wanting that α and believing that one wants that α; since the proposal is that wanting that α is believing that Dα, I can hardly be accused of saying nothing to suggest that D "has truth conditions any different from" (Cresswell must mean "makes a contribution to truth conditions any different from", since sentence operators do not themselves have truth conditions) "the 'wants' operator he is trying to analyse".

Turning now to Cresswell's positive suggestion, that there is a sense of 'wants' according to which what Dα means, when the D is understood as the desirability (or 'would-be-nice') operator relative to some given subject, is that the subject in question wants that α, I find myself in agreement. Suppose you are told that you can have either the contents of the box on the left or those of the box on the right, but not both, and that one contains $1 and the other $100. You are then asked which box you want to pick, in response to which you may well say "I don't know which one I want"; and this certainly suggests that there is one you want and one you don't want, though you do not know which is which. Cresswell would put this by saying that although you *really* want one of them—the one on the left, let us say—you don't have a belief that you really want that one, and so don't in the other sense he thinks 'wants' has, want that one. It is clear from his discussion that he takes the latter sense to be derivative. My own view is that if either sense is to be taken as primary, it is the latter sense, and that because there is no natural way of expressing 'D' we misappropriate the talk of wanting to do the job. In other words, the reply, natural though it may be, "I don't know which one I want", is not literally correct in the circumstances. Something similar can happen with belief itself, though in this case there is no missing 'D' to supply something of a justification for departing from the literal. Suppose you hear two conflicting accounts of what transpired on some occasion and are asked which account you believe. You may well reply "I don't know which I believe". Again this suggests that there is one that you do believe ("*really* believe"?) though you don't know which it is: but surely here the suggestion is wrong. We may sometimes need to think a bit, or perhaps to undertake some kind of clinical treatment (or whatever), to know what we believe on a certain matter, but no such course will help here. What you are ignorant of is not what you believe about the two accounts but rather which of the accounts gives a true description of what happened on the occasion in question. The appropriate reply, if one is being a little more careful, is not "I don't know which one I believe", but: "I don't know which one to believe". Similarly, though it has a less idiomatic ring to it, I would replace "I don't know which one I want", in the earlier example, with "I don't know which one to want". When you don't know which to believe, out of an account according to which α and an account according to which *not*-α, this is because you don't know whether α or *not*-α; when you don't know which to want, out of an outcome in which in α (= say, "you get the box on the left") and an outcome in which *not*-α, this is because you don't know whether Dα or D(*not*-α).

This last formulation of course brings in a controversial commitment to the B-D proposal defended in the paper, but the point about which usage of 'wants' to treat as fundamental and which as derivative can be put independently of any such commitments. I take that wanting, like believing, is a propositional attitude in the narrow sense distinguished in Chapter 0 above. We do not need to accept an incorrigibility thesis concerning such attitudes, to the effect that one can never be less than fully informed about which attitudes one holds to which propositions, in order to make sense of the *possibility* of perfect accurate introspection on that score. A perfect introspector knows what his or her propositional attitudes are, when it is propositional attitudes in the narrow sense that are at issue. But no amount of introspective omniscience will reveal to Susan, faced with her plastic ice creams, or to the subject we just imagined faced with the two boxes to choose from, what is, in Cresswell's sense, *really wanted*. So 'real wants' are anything but a species of wants, assuming that wanting is a propositional attitude of the kind delimited; and in writing 'Wanting as Believing' I had in mind, as the referent for each term, a propositional attitude of just this kind.

For the sake of engaging Cresswell on more or less his own terms, I have gone along with his interpretation of 'D', forsaking the neutrality of 'Wanting as Believing' on this score. As to quite how to characterize that interpretation, I am unsure. One way is in terms of what will make the subject happy, or satisfied: this, after all, is what the plastic ice cream won't do (for Susan), and it may be on such grounds that Cresswell says that she doesn't 'really want' it. However, for reasons given in the opening paragraph of 'Wanting as Believing'—the aging, ailing parent example—this is not what anyone is likely to mean by 'really want'. That is not to say we do not sometimes measure the strength of desires by consider the strength of anticipated subjective satisfaction, for the case of desires which are not of this 'posthumous satisfaction' kind: this, at any rate, will be argued in Chapter 3 below (where it will be contrasted with another measure of strength of desire – motivational efficacy). It is simply to say that if what we are after is a general reading of 'D' which will always render 'BD' equivalent to '(the subject) wants', taking the 'wants' here subjectively rather than as 'really wants', then the reference to happiness or felt satisfaction will not do the job. One alternative would be interpret 'D' as applying according to what the subject would want *if fully informed*. Such full information would drive out Susan's mistaken belief about what she takes to be an edible ice-cream, and it would fill in the epistemic lacuna of my subject above, presented with a choice between the two boxes. Certain problems about the conditional formulation arise which needn't detain us here; there is a discussion of similar formulations and the needed qualifications (though not in defence of the B-D proposal) in Smith [1994].

Another writer who should be mentioned *à propos* of (especially Section 2 of) 'Wanting as Believing' is Alan Reeves. Even though his work was published when that paper was written, I was not aware of it. Reeves [1975] is actually discussing some claims by Barbara Hall Partee (Partee [1972]) about the following sentence, here given with Partee's (and Reeves') numbering):

(10) John wants to murder the man who lives in Apartment 3.

What is referred to as claim (a) in the following passage is the claim that the ambiguity of (10) and similar sentences is one in which different senses correspond to giving the existential quantifier different scopes. We quote from p. 224 of Reeves [1975]:

> First, it is true that there is one reading of (10) which does carry the suggested presupposition or implication of existence. This is the reading which is recognised by Partee. In accordance with claim (a), this reading might be written as (10′). However, there is another possible interpretation of (10), one that Partee fails to recognise. Suppose that the speaker knows that John holds both the false belief that Apt. 3 is inhabited and that the sole inhabitant is a male. He also knows that John is greatly disturbed by noise that he wrongly believes to come from the vacant Apt. 3. John says that he wants to kill the man in Apt. 3 because of the noise. Under those circumstances, the speaker can say (10) and say something that is true even though (10′) is false. This other reading of (10) might be written in accordance with claim (a) as:
>
> (10″) John wants that there should be a man who lives in Apt. (3) and is murdered by John.
>
> This way of rewriting (10) does not read well because it obliterates the distinction between what we might call John's primary want and what John presupposed in forming that primary want. A more satisfactory version might be:
>
> (10‴) John believes that there is a man in Apt. 3 and wants that he is murdered by John.

What Reeves has not noticed is that the interplay between the propositional attitude verbs and the quantifier in his (10‴) is problematic in exactly the way we saw that it was in the case of our own example (1) in 'Wanting as Believing' ((1) = "John thinks that Mary owns a mad dog, and he wants it to bite her").

From: *Philosophical Papers* **19** (1990), 99–118. Notes begin on p. 58.

CHAPTER 3. WANTING, GETTING, HAVING

1. Propositional Attitudes and Affective States

I will be interested in the relation between two aspects of the complex conceptual apparatus we employ in psychological description. On the one hand, we use a variety of verbs followed by 'that'-clauses to attribute what philosophers have come to call *propositional attitudes*: 'Sandie believes that it is raining, knows that it will soon stop, hopes that tomorrow will be fine, but fears that it will only rain again.' While such constructions have traditionally been found problematic for systematic semantic theory, they are understood well enough by all of us (as speakers) for present purposes, and I shall have nothing to say about the problems. Nor will more be said than already has been, with the citing of these examples, and the allusion to received usage, to demarcate precisely and in general terms what the propositional attitudes are. On the other hand, we also have the apparently simpler practice of giving non-propositional descriptions of our states of mind. Here I am thinking of the application of those concepts such as that of being happy, or being afraid, or being disappointed, or being angry, no full account of which—whatever else it might involve—could omit mention of the distinctive way being in such a state *feels* to one who is in it. I will call such states *affective states*, in order to avoid having to attend to any distinctions there may be between feelings, emotions, and moods; these states are what are often called 'occurrent emotions', the intention in calling them affective states being to allude to the essential phenomenological dimension just mentioned.[1]

A reference to affective states is involved in taking certain propositional attitude ascriptions at face value. For example, we say that Sandie is happy that it has stopped raining, or disappointed that it rained for so long, or angry that she was not elected. We call the attitudes so ascribed *affective propositional attitudes*.[2] The literature on emotion makes frequent mention of directed emotions with specific objects. Arguably, so many different relations have been considered as 'the' emotion/object relation that this terminology does more harm than good.[3] However, in the cases in which emotions are said by those given to this way of speaking to be directed toward propositional objects, it is affective propositional attitudes that are at issue (as in 'angry that—*or*: at the fact that—she was not elected' rather than 'angry with those who did not vote for her'). There is of course no suggestion that to have an affective propositional attitude at a time, one must actually be in the associated affective state at that time (in our example: be feeling angry).

Philosophers have paid considerable attention to the relationship between propositional attitudes and affective states, often noting that certain states are not available to an individual lacking an attitude falling within some range. For example, Philippa Foot [1958] cited the

following examples involving ('directed') affective states with correlatively required beliefs: (i) *a*'s being afraid of *b*, *a*'s believing that in some way *b* is a potential danger to *a* (ii) *a*'s feeling proud of *c*, *a*'s believing that *a* is in some way favourably associated with (in the most straightforward case this being by being causally responsible for the merits of) *c*. Desires, too, have been amongst the attitudes it has seemed not unreasonable to require for certain affective states. We may use again the case of (occurrent) fear, with the associated desire—not, of course, one necessarily overriding all others—being to get away from that which is feared or else to in some other way remove the threat it is perceived as presenting.[4] In what follows, we shall pursue connections somewhat less direct that appear to obtain between certain attitudes and states. Specifically, we will be concerned with the propositional attitudes of *wanting* and *being happy* (*that*), the latter being one of the above-mentioned affective attitudes, and with the affective state of happiness (in the sense of *feeling happy*). In §2, we note that there is actually something problematic about the classification of desire (wanting) as a propositional attitude; nevertheless it is a view we continue to work with, after finding a way of construing it so as to make it plausible. §3 suggests that the connections between this attitude and *motivation* have been overplayed at the expense of other connections it has with certain affective states (principally happiness). The latter connections arise from more direct connections with being happy that such-and-such is the case, which will occupy us in §4.

2. Wanting and Having

First, to clear up the situation with wanting. Let us see how much of a problem for the view that ascriptions of desire can always be presented in the verb + *that*-clause format befitting a propositional attitude is posed by examples not at first sight conforming to this pattern. Consider:

(i) *a* wants to φ

(ii) *a* wants *b* to φ

(iii) *a* wants ___

where in (i), (ii), 'φ' is schematic for a verb phrase and the blank in (iii) for a noun phrase. The first two seem unproblematically renderable in the preferred idiom, with sentential complements 'that a'φ and 'that *b* φ', as long as no fuss is made over the *de se* nature of (i), which it had better not be on pain of not counting '*a* believes that he/she will φ' as a propositional attitude ascription. (No commitment is intentionally incurred here on the question of whether, given suitable senses of 'object' and 'proposition', propositions are in fact the objects of the propositional attitudes—on which see Lewis [1979], or Castañeda [1966], for the seminal observations. There is also an interesting question, raised in Lakoff [1970], skirted here, about a *de se* version of (ii)—giving rise to the construction '*a* wants himself/herself to φ'—and how this is related to (i).) The third construction raises some difficult issues. Consider some instances:

(A) Tom wants a house in the country.

(B) Tom wants a glass of milk.

(C) Tom wants a holiday in Fiji.

One natural line to take would be to say that the desires attributed to Tom in these cases are, respectively, a desire to own a house in the country, a desire to drink a glass of milk, and a desire to go on a holiday in Fiji. This would give a reduction to the form of construction (i), and thence by the earlier recipe, to the *that*-clause format. But wait. Perhaps Tom wants only to rent a house in the country. It is still right to describe the real estate agent Tom consults in his search as knowing, when he goes through his files discarding various urban and suburban properties, that his client wants a house in the country. Yet (A) is not to be regarded as *ambiguous*, presumably. Think about (B). Perhaps Tom is a painter, who wants a glass of milk so he can put it on the table between the flowers and the bowl of fruit, for a still-life arrangement. We can hardly regard (B) as ambiguous between 'Tom wants to drink a glass of milk' and 'Tom wants to paint a glass of milk': pretty clearly, there will be no end to the number of such would-be disambiguations: a particularly virulent case of what transformational grammarians used to call the 'recoverability of deletions' problem.

From a linguistic point of view, we may get around the difficulty by supplying, in, e.g., the case of (A), 'Tom has a house in the country' as the sentential complement.[5] The strategy seems general enough – whether he buys or rents it, if Tom's desire is satisfied, it will be true that he has a house in the country; and what the painter wants, no less than one who is thirsty, is to have a glass of milk; note also that (C) could reasonably be expanded to 'Tom wants to have a holiday in Fiji'. (The renting/buying, drinking/painting indeterminacies do not quite appear to afflict (C): there is very little you can do with a holiday in Fiji other than take—*alias* have—one.) But philosophically, such a resolution of the difficulty may well seem less than satisfactory, as is noted by Anscombe in a suggestive passage from her [1957] which I shall quote at the end of this section. What exactly is the proposition that Tom has a house, anyway? Once we notice how much further afield our use of 'have' stretches than our use of 'own', its use appears so all-encompassing that it seems no more than a dummy verb. For reasons already rehearsed in connection with 'Tom wants a house in the country', we don't want to say that 'Tom has a house in the country' is *ambiguous*. The usual alternative is to say that we have simply a case of non-specificity. To say that Tom travelled to Europe is not to say anything ambiguous, though it is to leave it open whether he went by plane, by boat, or by some combination of these and/or other means of transport. But this doesn't quite seem to fit the present case. It's not that there are various ways of having a house, one of which is renting and another of which is owning: it's more that we don't have any generic conception of having a house to start with, of which these might be seen as further specifications. For what it's worth, the 'dummy verb' account of 'have' seems to me to be made even more plausible if we switch from the proposed filling out of Tom's desire in (B), 'Tom has a glass of milk', to this variant: 'Tom has a glass of milk beside him'. What does this mean beyond simply: 'There is a glass of milk beside Tom'? And doesn't 'This hill has no sheep grazing on it' mean just 'No sheep are grazing on this hill'? This suggests that the word 'have' comes in in the process of reallocating previous non-subject lexical material into grammatical subject role.

But could any such sketch be extended to cover all uses of 'have' (with the exception of perfect 'have' which I believe to be at most etymologically related to the homophonous main verb)? Notice how different 'Tom has a glass of milk' and 'Tom has a glass of milk beside him'

are in at least the following respect: only the former has a genuine present progressive, because having a glass of milk (i.e., drinking one) is something you can do, while having a glass of milk beside you is not something you can do. (This makes having a glass of milk unlike having a house in the country.) Nevertheless one has the impression that this 'have' is semantically neutral, merely indicating some sort of vague association between Tom and a glass of milk, which, people and glasses of milk being what they are, *defaults* to the more specific relation of drinking. When our expectations provide us with no such default setting, we are at a loss as to how to construe a 'have' sentence directly, and grope about amongst such analogies as occur to us. Thus, 'These plants have green leaves' poses no problem, while 'Those nurses have green leaves' does. You might say that the latter sentence, unlikely indeed to be heard outside of a science fiction context in which the possibility is being taken seriously of people spontaneously sprouting leaves, is straightforwardly false, and so the only problem it poses is the problem of how anyone might try to use it, such contexts aside, to make a reasonable assertion. And this problem won't arise if enough of a background is in place. Perhaps the nurses are given leaves of various colours to wear on their shoulders as indications of seniority, and you explain to someone who cannot see that group's leaves: those nurses have green leaves. Do not be misled by the proximity of the leaves to the nurses in this example, thinking of this as a relevant similarity with the case of the trees. It might instead be that each nurse is assigned a rank which is indicated in the register of hospital staff by the placing of an appropriately coloured leaf by that nurse's name: we can again say of some of the nurses that they have green leaves, of others that they have brown leaves.

It may here be worth entering a reminder that the move away from proximity can also go the other way. A hill, to recall an earlier example, can not only have sheep on it but also have bumps on it: and these bumps are not just close to the hill, they are part of it. We say: things *have* parts. They also *have* shapes. Philosophers have often had occasion to remark on the diversity of what counts as having, for example, in criticizing an excessively perceptual model of sensation by accusing its proponent of assimilating having a toothache to having a toothbrush; or again in diagnosing one source of dualism as overlooking the possibility that 'People have bodies' may be more like 'Objects have shapes' than like 'Kings have palaces'. The last point I shall make to defend the suggestion—admittedly deserving further elaboration—that 'have' is a semantically neutral dummy verb concerns relational expressions. Consider, 'Tom has a sister'. It will not be hard to get agreement that this simply says someone stands in the *sister-of* relation to Tom. Taking 'has' as a two place predicate in its own right would lead one into such as invalidities as:

Tom has a sister.

Every sister is a daughter.

∴ Tom has a daughter

(This recalls an argument about small elephants which was popularly urged against classifying attributive adjectives as one-place predicates.) An even more obvious consideration in the present case would be the oddity of saying, that if Tom has a sister and that sister is Sarah, then Tom has Sarah. Of course one can use the form of words, 'Tom has Sarah as (or 'for') a sister', but here we really have the semantically unstructured binary predicate '*x* has *y* for a sister', the converse of the relation '*x* is a sister of *y*'; if we take the above suggestion about the role of 'have' in indicating a re-allocating to cases role, one might almost say here, the *passive*

of '*x* is a sister of *y*'. You may say that this is all very well for having a sister, having a daughter, and so on, but perhaps we could make out a case for hiving off *this* use of 'have' from the others, for which a more traditional account, as indicating a very non-specific binary relation, might be salvaged. I defend the 'dummy verb' account by replying that this would be to make an ambiguity-claim which is very implausible in the light of such remarks as 'Tom is so fortunate: he has a magnificent house in the country, an interesting job, a beautiful wife and two lovely children', which induces no zeugma-reaction.

Where does all of this leave our working hypothesis that all wanting is wanting that? The situation is somewhat delicate. The attempt at reducing wanting a such-and-such to wanting that one should have a such-and-such was an attempt at procuring a uniform paraphrase in the canonical idiom for propositional attitude ascriptions, but we found that the apparent uniformity was spurious. The linguistically uniform results involved havings as relations of all sorts, and it seemed doubtful that we could postulate one sort of which they were all subsorts. A fallback position would be as follows. Drop the attempt at uniform paraphrase, and say instead that in any given case of an individual's wanting a such-&-such, his desire is a desire that *p* for some *p*. (Again, we suppress explicit *de se* complications here.[6]) To make clear what the desire is, you can specify the *p*, thereby producing an attitude ascription of the favoured form. But we pull back from the claim that the ascription so offered is a *paraphrase* of the bare '*a* wants a such-&-such' form.

It remains to explain to those familiar with Anscombe [1957] how the point fussed over here differs from one she made much of there. Apparently less persuaded than I am on the secondary status of talk of wanting a such-&-such and the primacy of wanting that, she conducts her discussion in the former idiom, and holds that in all but special cases, a person cannot intelligibly 'just want' something, and must be able to supply an answer, which she calls a desirability-characterization of the object wanted, to the question 'What do you want it for?'. The famous case of the man who wants a saucer of mud, so amusingly described by Anscombe, is worth quoting in full:

> But is not anything wantable, or at least any perhaps attainable thing? It will be instructive to anyone who thinks this to approach someone and say: 'I want a saucer of mud' or 'I want a twig of mountain ash'. He is likely to be asked what for; to which let him reply that he does not want it for anything, he just wants it. It is likely that the other will then perceive that a philosophical example is all that is in question, and will pursue the matter no further; but supposing that he did not realise this, and yet did not dismiss our man as a dull babbling loon, would he not try to find out in what aspect the object desired is desirable? Does it serve as a symbol? Is there something delightful about it? Does the man want something to call his own, and no more? Now if the reply was: 'Philosophers have taught that anything can be an object of desire; so there can be no need for me to characterize these objects as somehow desirable; it merely happens that I want them', then this is fair nonsense. ([1957], p. 70)

A person can't just want a saucer of mud, if Anscombe is to be believed. This is reminiscent of our conclusion, above, that the whole truth has not been told when we say that an individual wants a glass of milk: before we can know what is wanted we need to know if what the person wants is to drink the milk, look at it, paint it, or some other such thing. But this is *not* at all the same as what Anscombe is getting at, for while the completion I request is supplied when I am told that the man who said he wanted a saucer of mud tells me that he wants to sit beside it – for he has told me what proposition has to be true for his desire to be satisfied – Anscombe's request to be told in virtue of which of its features he manages to find that prospect desirable remains to be met. In fact Anscombe briefly alludes to the distinction between these two matters in the discussion of a man who wants a pin. Again, I must quote; the closing sentence can be regarded as having triggered the extended discussion above on the subject of the dummy verb 'have':

> Now saying 'I want' is often a way to be given something; so when out of the blue someone says 'I want a pin' and denies wanting it for anything, let us suppose we give it him and see what he does with it. He takes it, let us say, he smiles and says 'Thank you. My want is gratified.'– but what does he do with the pin? If he puts it down and forgets about it, in what sense was it true to say that he wanted a pin? He used these words, the effect of which was that he was given one; but what reason have we to say that he wanted a pin rather than: to see if we would take the trouble to give him one? (…) It is not at all clear what it meant to say: this man simply wanted a pin. Of course if he is careful always to carry the pin in his hand thereafter, or at least for a time, we may perhaps say: it seems he really wanted that pin. Then perhaps, the answer to 'What do you want it for?' may be 'to carry it about with me', as a man may want a stick. But here again there is further characterisation: 'I don't feel comfortable without it; it is pleasant to have one' and so on. To say 'I *merely want* this', without any characterisation is to deprive the word of sense; if he insists on 'having' the thing, we want to know what 'having' amounts to. (Anscombe [1957], pp. 70f.)

There emerges from this discussion, then, a distinction between two kinds of answer to the 'what do you want it for?' question which Anscombe has been pressing. One kind of answer does no more than spell out what the desire is, by supplying a propositional object: 'What do you want a pin for?' – 'I want to hold a pin with me as I walk around'. A second type of answer to the question, which may be re-asked at this stage, might be 'Because it makes me feel comfortable', or 'To ward off rheumatism'. This is where the business about desirability-characterizations comes in. Supplying them makes the desires intelligible only in the sense of enabling one to understand the person's having such a desire. But an answer to the question at the first stage is needed before one can even say what exactly the desire *is*. One might venture further here, and say that such an answer is required for one to even understand the desire-attribution, but I think this goes too far. If supplying the *that*-clause (or its infinitival reduction) were disambiguating a multiply ambiguous sentence, then this would be right, since, pending such a disambiguation, one would have no idea which sense was intended. But we found, in

discussion of (A), (B), (C) above, that the multiplicity of would-be senses was too great to make the ambiguity charge stick with any plausibility.

3. Wanting and Getting.

The passages quoted in the last section from Anscombe appear in her [1957] close on the heels of one of that book's most often cited *dicta*: the primitive sign of wanting is *trying to get*. Let us attend to the role of trying to get, and of what happens when one does or does not get, what one wants. The first thing to note is that talk of getting is paired with talk of wanting *a thing*, rather than the idiom I prefer to take seriously, of *wanting that* (subsuming *wanting to*). Transposed into this preferred idiom, the dictum becomes: the primitive sign of having a desire is trying to satisfy it. There is, this point aside, another disadvantage to Anscombe's original formulation, nicely observed by Alan White[7]:

> This common philosophical linking of wanting with getting is paralleled by the equally common linking of wanting with attaining some end. Philosophers have over-concentrated on examples like 'I want my car today, so I'll go to the car-park' to the exclusion of examples like 'I want my car today, so I cannot lend it to you'.

Here, the desire it would be plausible to ascribe to the subject in each example, incidentally, would be the desire that he have the use of his car on the day in question. The second case is a bit complicated; White may have in mind the fact that no special effort is required to have the use of the car – one simply has to refrain from lending it, and so on. Here one should note that Anscombe spoke of the 'primitive sign' of wanting, rather than of a necessary condition. The 'sign' talk suggests that the trying in question be regarded as the display of a disposition, as one might say that the canonical display of solubility in water is: dissolving in water. So preparedness to try to realize the desire, should the need for effort arise, is what we expect from the wanter, when it is this aspect of the concept of desire that we have uppermost in our minds. (Some other aspects will emerge presently.) As White notes, an excessive zeal to connect wanting with trying to get may also betray a mistaken belief that one can only want what one does not have.[8] The formulation in terms of propositional wanting covers, I take it, not only trying to get, but trying to keep: here there is a desire that something should remain the case rather than that something should become the case. Both sorts of desire share this feature: they concern the future. What about a person's wanting, e.g., to be in the bath right now, as opposed to wanting to get into or to stay in the bath? There may be pressure here to change verbs. It might be said: if the person's not in the bath, then the most he can do is wish he were or want to get in as soon as possible. This goes along with the line that desires whose unsatisfiability (given the circumstances) is evident to the desirer are not wants but ('mere') wishes. Next suppose that the person who wants to be in the bath right now is in the bath. Then it might be held that it would be misleading to say that he wants to be in the bath, and better to say that he is glad to be in the

bath. I sympathize with the reaction to the first case, though we need not go into the subtle interplay between desire (used as a generic term here), perceived impossibility, and preparedness to act, which gives rise to our threefold distinction between wanting, wishing, and hoping. In the case of the person who is already, and knows he is, in the bath, I do not agree that there is anything wrong with saying he both wants to be, and is, in the bath. But it is, I suppose, correct to observe that there are dangers of misleading an audience by saying too little, even when it is true. If I say 'Sandie wants to be a university lecturer' and no more, I tempt you to wonder 'And what is she going to do about it?', and it may be non-plussing for me to reply: nothing at all, she already is one. The oddity can be cancelled, it would appear, by heavy stress on 'wants', in any case. What does seem useful about this response is the appropriateness of talk of the individual's being glad to be in the bath when his current desire to be is currently satisfied. Here we come face to face for the first time since we set them aside in §1, with what were there labelled affective propositional attitudes. I take it of course that being glad or being happy to be φ is being glad or happy that one is φ.[9]

Preparedness to act on a desire ('trying to get') is, then, one aspect of having that desire. Now that the affective propositional attitudes are back in play, we can discern another. Being happy at the satisfaction of a desire seems to me to be not just a contingent feature of desire satisfaction. The 'trying to get' theme, however it is eventually to be worked out, focusses on the motivational aspects of desire but ignores the emotional side of the picture. This feature helps to create the oddity in Anscombe's pin-wanting example. We imagine the man taking the pin and saying in a very impassive sort of way 'Thank you. My want is gratified.' He puts the pin down and walks away from it, just as miserable as he was before we went to all those efforts on his behalf to procure that pin. In a television documentary about a beauty contest, the winner, newly crowned, wipes away the tears and says: I never realised how much I wanted to win. Presumably she was reflecting not on how much effort she had put in – she knew that already – but on how good it made her feel to have won. This affective dimension competes with the 'trying-to-get' dimension as a measure of the strength of desires. Anticipation of great pleasure on the satisfaction of a desire may not go hand-in-hand with preparedness to expend great efforts to procure that satisfaction. You might say: then the subject does not really much want whatever it is. But this is only one way to measure the strength of desire – the beauty contestant used another. We have both dimensions, neither of them the right way to tell how strong a desire really is, each of them providing a legitimate notion of strength of desire.[10]

We need to attend to both of these respects in which a desire may be strong, for example, to get a reasonable understanding of what laziness is. It is hard to understand the way we feel about laziness if the strength of a desire is just given by its motivational efficacy. If I just happen to prefer lying here on this couch to working on the wall-papering, then, other-regarding repercussions aside, what's wrong with my acting on the stronger desire, the desire which shows its greater strength by the fact that I'm still here on the couch? Is it that my longer term interests are thereby jeopardised? Well, yes: but what does this mean? In this example, there is no pleasure to be got out of wall-papering, though there is plenty to be got out of having finally finished the job. Change the example. Instead of lying on the couch, I could be playing badminton, and moreover, know that this very activity, rather than some end to which it is a means, would give me more pleasure than remaining here. But I just can't be bothered.

Measured by emotional payoff, I prefer the prospect of the game to the prospect of extended languishing. Measured by motivating force the direction of preference is reversed. This seems to be what is involved in our conception of laziness as something like a fault in (practical) rationality. There is a violation of a normative principle we might call

'GO FOR IT': *Make your efforts at satisfying desires proportionate to the extent to which you'd be happy to have them satisfied.*

Stated baldly thus, this principle may not seem very appealing. It ignores, in particular, both the cost of efforts at desire satisfaction and the probability of their not being successful. But it does not seem impossible that some reasonable variation on the 'Go For It' theme will emerge, a principle which enjoins a matching up of the two measures of strength of desire.[11] Indeed, the injunction to maximize expected utility may already be such a principle, given an appropriate understanding of 'utility'. The inappropriate understanding is that on which this goes by preference satisfaction, with strength of preference revealed by choice of action. When subjective probabilities are similarly calculated from observed behaviour, everyone turns out to be maximizing expected utility all the time: if you think not for some case, that just shows you miscalculated the strength of desires and beliefs for that case. So for the principle to have normative bite, the notion of utility should be that of the extent to which one would be happy to see various outcomes, rather than the extent to which one is motivated to secure those outcomes. You might say this is just replacing preference egoism by hedonistic egoism, by analogy with the parallel distinction amongst utilitarianisms, but there are several things wrong with saying this. First, the suggestion of egoism is wrong – you may anticipate being happy that others are doing well as part of your altruistic desire that they should do well. Second, I am trying to claim that the term 'preference', meaning wanting one thing more than another, because it involves a comparison between strengths of desires, can cover both the action-producing measure of strength and the more affectivity-oriented notion. There is not a change from preference to something else. Third – the appropriateness of terms like 'hedonistic' in this connection is as yet unsettled, because we may wish to interpret 'Go For It' as alluding to the affective propositional attitude of being happy that something is the case, rather than in terms of the affective state of being happy. Nothing has yet been said on the relation between these two things. It will occupy us a little in the following section.

In the meantime, I conclude this discussion of the affective dimension of desire by mentioning a third measure of the strength of desire. This is the extent to which one would be, as far as one can now judge, unhappy to see the desire unsatisfied. (Like the second measure, then, this consideration connects with affectivity rather than with motivation.) Again, let us remain neutral as to what exactly this means – whether it means how unhappy, *tout court*, one would be should that happen, or on the other hand how unhappy one would be that it happened. I am also fudging here, as in the discussion above of happiness at having one's desires satisfied, the question of which we count, *a*'s estimation of how unhappy a would be, *vs*. the extent to which a in fact would be. At the moment all I want to say is that something like one of these – the relation between them being touched on in the following section – gets into the act when we

assess strength of desire, and that it need not march in step with either the 'trying to get' measure, or the 'how-happy-you'd-be-if-the-desire-is-satisfied' measure. To show the contrast with the former, consider the case of a man devastated by the departure of a wife he was not prepared to do anything to keep from leaving, or—a mismatch in the opposite direction—your own case when you learn how little you wanted the job you'd been trying your hardest for by noting how little disappointment you feel at not getting it. As to the contrast between the positive and the negative affective measures of strength of desire, we may do worse than to consider, as an attempt to nullify the contrast, this principle that might be put forward as a norm of rational attitude-management:

PRINCIPLE OF PARITY: *Make the extent to which you would be happy that something is the case be equal to the extent that you would be unhappy that it wasn't the case.*

Again, considerations of probability make this principle implausible just as it stands. For example, while you might be ecstatic, or at least highly delighted to win the state lottery on the next draw, you will not be correspondingly upset not to win. Such an imbalance does not seem to indicate that your attitudes are in any kind of tension. What about when you rate the chances of the thing you want at about 50–50? Say you have applied for a job for which you and one other candidate, reckoned by you to be of roughly equal merit in the eyes of the appointment committee, have been shortlisted. There does seem to be something odd in being of such a mind that the following two conditionals are both true: (i) if you were to fail to get the job, you would be heartbroken; (ii) if you were to succeed in getting it, you would be either indifferent or at best, mildly pleased. It would be interesting to spell out what is involved here, making a case for some modified form of the principle of parity. Some of the relevant considerations will be aired in the following section. In the meantime, let us simply note that if the other job candidate is violating the principle in the opposite direction (he too rates his chances at about 50–50, would be over the moon if he got the job but not at all mortified to miss out) then there is no simple answer to the question which of the two of you wants the job more. As I say, there are simply several different – I've mentioned three – independent dimensions along which strength of desire can be measured. 'Independent' here means that the measures involved can (and often do) come apart, not that this may be so even for someone whose desires are perfectly in order—a point which would of course be contested by a defender of 'Go For It' or of the Principle of Parity.

4. Affective Propositional Attitudes

We will soon be back in the area of normative considerations about appropriateness and reasonableness of affective states, but first a few ground-clearing remarks and questions about the affective propositional attitudes are called for. We will concentrate on the example of a person's being happy that something is the case. Compare (1) with (2):

(1) John is happy that *p*.

(2) John is happy because *p*.

Well, a common observation to make about sentences like (2) is that they are ambiguous. Perhaps what is intended is a causal explanation, or at any rate some kind of explanation whose correctness is not jeopardised by John's failure to believe that *p*, entertain the thought that *p*, or even be so conceptually equipped as to be capable of entertaining that thought. On the other hand, we may be trying to give John's reasons for being happy. But taken either way, (2) entails that John is happy. It is his being in a certain affective state that we are trying to explain, whether externally or 'from the inside'. (1) on the other hand, does not seem to entail that John is happy, *tout court*.[12] John might be happy that *p* and unhappy that *q* at the same time, where we can take 'unhappy' to imply at least 'not happy'.[13]

So much for ground-clearing. Now a question. Would something like the following do as an account of how the attitudes are related to the states in such cases? Being happy that *p* is being such that were one to fill one's mind with reflection on one's belief that *p*, to the exclusion of distractions, one would be happy. (Actually it would need to be added that the belief in question was true—the construction in question being 'factive'—and perhaps even a piece of knowledge.) Thus although there is a distinctive way it feels to be happy, and no distinctive way it feels to be happy that *p*, being happy that *p* is to be understood in terms of one's feeling the former distinctive way under certain conditions.[14] There may, however, be insurmountable circularity objections to any such account offered as an *analysis*. (What are the distractions to be excluded, if not the things one is not happy about?) Whatever its merits, I think it may be useful to consider some everyday aspects of the kinematics of the emotions with an eye to the role that focussing of attention has on converting affective propositional attitudes into affective states.[15]

Consider, for example, the phenomenon of consolation, self- or other- administered. Your house has just been burgled and you've lost your television and video recorder. Still, says consoler, at least they left the record-player and all the furniture, and did no structural damage. Or you've broken your right leg while skiing – but, says consoler, the man in the next bed has broken both legs and his collar-bone. Consolation works, when it does work—and perhaps as often as not one just feels like saying 'Get lost. This morning everything was fine and just look at me now.'—by releasing the latent affectivity in your being pleased not everything was taken, or glad that it was only a leg you broke. However miserable you may be, the consoler—who may, as I said, be you yourself—wants to turn the fact that you are happy that things aren't worse into happiness, or at least a moderation in your unhappiness, because things aren't worse. And the process works by focussing your attention on the proposition your 'happiness that' was directed on, and distracting it from those other propositions you are inclined to dwell on.

It's interesting that not just any old truths can be expected to be paraded by the consoler before you with any chance of success, however glad you may be that they *are* truths. Suppose that after the burglary, someone says: 'Never mind, at least you don't have a dentist's appointment tomorrow', or after the accident, someone says 'Well, look on the bright side – at least you had the good fortune twenty years ago to be educated at Oxford University'. The trouble here is that to work well the bright side has to be the bright side of the same general

picture as you were previously focussing on the darker side of. (Under some circumstances, such as when trying to get an overall perspective on your life while you wait to die, this may well work: here the picture is bigger.) The need for some kind of relevance in the propositions the consoler makes salient is especially clear when part of the consolation would have you feel *relieved* that things weren't worse, rather than just happy they weren't. If everyone else who had been burgled had lost more than you, or the man in the next bed had his accident as a result of the same avalanche as you, relief has a chance of setting in. You say '*Whew*—that nearly happened to *me*. Thank heavens I got away so lightly.' There is, by contrast, no chance if, when you are distressed by the inconvenience of the intermittent nosebleeds you've been suffering for the last month, someone says '*You* should worry – there was once a man in Tibet who bled to death after his hand was ripped off by a tiger.' There seems no way in which you can respond by seeing your admittedly, by comparison, minor problem, as a lucky break to have escaped *his* fate.

Though consolation works especially well to bring about a favourable mood change when the element of relief is present, relief has no special connection with consolation. After all, you may relieved not only that things aren't worse, but that they aren't bad at all: for example relieved that you were found not guilty. Here, as always, a bad prospect is lurking, and its salience, as in the burglary and injury examples, is due to its non-negligible probability of eventuating.

Another way for the bad prospect to be made salient is for it, or more accurately something very much like it, to have been recently experienced. You are relieved when a pain stops even when you were absolutely convinced that it was going to stop then. The contrast is provided not by the likelihood of the pain being present but by its all too recent actuality. I think that something close enough to relief for it to be petty to quibble about the name can arise when neither your recent nor your feared-present life harbours what you're relieved not to be going through. Seeing how bad it is for ill-treated prisoners somewhere as you watch a television documentary (or even a fictional film) may leave you not only sorry for them but relieved not to be undergoing what they are. The bad prospect is made salient by empathy with the victims rather than by a past or risked involvement on your own part.

As a past misfortune recedes relief loses its grip but we are still glad it's over. The consolation move would have us dwell on this where that may help to dispel, for example, brooding resentment over the incident's having happened at all. Is some such move always legitimate? Here you are, happy that something is no longer the case but unhappy that it ever was in the first place. Which of these, let's suppose perfectly appropriate affective attitudes, is the one that deserves to determine occurrent affective tone? We should resist the reply that whenever your are happy that p and unhappy that q, what you should be is happy, *tout court*, rather than unhappy, because it's a better state of the world for more rather than fewer people to be happy. This, or some egoistic analogue of it, may be a good answer to some question, but not the one we were considering. Something more like our usual conception of justification for beliefs, as distinct from a parading of practical reasons for holding them, surely arises for emotions too.[16] The way it would be consequentially advantageous to feel just might not be the way it's appropriate to feel under these circumstances. And a negative verdict on this score need not come in only for those affective states which are sustained by false beliefs. But I have no theory of these matters. (Have you ever had anyone try and cheer you up by saying, "There's

no point being sad about it now – that won't do any good"? As if you were being sad for a *point*.)

People notoriously differ in temperament as to which way, if either, to resolve the affective tone question. Two who have been through the same bad experience may focus, one on its badness, the other on how good it is that it's over. Of another pair who grew up together under idyllic circumstances but have now fallen on harder times, one will be happy because, and not just happy that, things were so good, using – for example – reminiscing as a 'focussing' device, while the other will be unhappy because, and not just unhappy that, it's all over. Actually such temperamental differences may even extend further, so that, in the case of the latter individual he is not only not made happy by reflection on past happiness but isn't even pleased *that* things went well in the past. After all, not every marriage between a theory of value and a philosophy of time will yield amongst its offspring the claim that if it was good that something happened at a past time, then it is now good that it did happen then. The person we were just imagining denies this claim. There seems no inconsistency. But it's a disconcerting style. Getting off the plane and making for home, you say: Wasn't that the most *marvellous* holiday!— Everything went so well. They say: Yes, but that was *then*. It doesn't help us *now*.

Let me close with an example not involving present-to-past contrasts. Suppose I have been looking forward to attending a party at which there will be a chance to meet up with some good friends I haven't seen for a long while, but unfortunately, I have just come to be afflicted by a distressing but not incapacitating sore throat. It is nasty, not only 'in itself', but because it would certainly take the edge of enjoying the party. On the other hand, the party would at least offer some distraction from the sore throat. Given that the party is on, I'd rather not have the sore throat than have it (well that would be the preference anyway, but especially given the party); but given the sore throat, I'd rather the party were on than not. So what's the problem? – you ask, tabling my preferences:

1. Party, no sore throat
(2. No party, no sore throat)
3. Party, sore throat
4. No party, sore throat

The second-ranked item is in parentheses as its position is not forced by my description of the situation – just a plausible enough filling out of the case. I reply that I know all about the preferences, and all about my propositional attitudes, affective or otherwise, and that I am now on my way to that party. But I don't know how to feel. Because of the sore throat, I'm glad the party is on; because of the party I'm (extra) annoyed about the sore throat. These affective attitudes entail no particular affective states, though. It may seem far-fetched to entertain the possibility of normative principles for bridging this logical gap; I hope only to have raised interest in its existence.[17]

NOTES

1. See Stocker [1983] for the view of occurrent emotion here assumed.

2. A closely related concept is isolated under the description 'emotional thoughts' in Stocker [1987].

3. Such a case is made in Nissenbaum [1985].

4. See also the more general discussion in Thalberg [1964]. [[This reference was unfortunately missing in bibliography of the present paper as originally published, but may be found in our list of references at the end of this thesis.]]

5. See McCawley [1972].

6. As well as suppressing a further complication—also, as it happens, first uncovered in the work of Castañeda—raised by examples such as '*a* wants to rescue the victims of the next Mexican earthquake' in which it is no part of what is wanted that there should be an earthquake; see for example Castañeda [1967*a*] , and, for further references and a discussion focussing on desire-attributions, §4 of Humberstone [1982] [[= Chapter 1 above]].

7. In White [1975], p. 109.

8. The same point was made in Matthews and Cohen [1967].

9. As these constructions figure in the above examples, that is. Of course, there is also the construction with 'happy', as in '*a* is happy to go' in which 'happy' means something like '(very) willing'. (This does not exist with 'glad': we have only '*a* would be glad to go'.)

10. There would be a circularity in using real or anticipated happiness as a test for desire if happiness were itself construed simply as the state of having one's desires satisfied; but of course the view of the text is that happiness is an affective state, more specifically a matter of feeling good. Even many who take the alternative satisfaction-of-desires line on happiness acknowledge the existence of a genus containing pleasure, contentment, etc., which is what they are keen to distinguish happiness – as they see it – from. There is, however, a certain hard core opposition, represented perhaps by Ryle [1955], for example, which would deny that any affective state or (to use Stocker's term) 'psychic feeling' can be intrinsically pleasurable. The argument for this view runs: let X be any type of state a subject S can be in; surely it must remain an open possibility that S should be in state X and not enjoy being in state X. Thus no type of state, whether physical or psychological, can be of its very nature such as to please any subject in that state. To react to the argument we should disambiguate the part about the possibility of being in state X and not enjoying being in state X. The sense in which this may have to be a possibility is when this means: and not be happy that one is in state X. But what the argument needs is: and not be happy, even though one is in state X. And this is question-begging: for where has not been shown that we may not take state X to be happiness itself, to get something which is certainly not an open possibility?

11. Talk of a matching up suggests that our formulation of 'Go For It' leaves open the question of what is to be hot to match what. Compare the request to 'make the lengths of stick A and

stick B equal', which could be done by changing the length of stick A to match that of stick B, or *vice versa*, or by changing both. For the application to laziness below, one would have in mind the more specific injunction to match one's efforts to the (actual) extent to which one would be happy to see one's desires satisfied. Turning from a non-specificity to an ambiguity in 'Go For It', we note that the 'you'd be happy to see them satisfied' can be understood (to use the terms of Hare [1981]) along now-for-then lines or along then-for-then lines; this is not an ambiguity we need to resolve for present purposes.

12. An alternative response would be to hold that 'John is happy' does indeed follow from 'John is happy that *p*', appearances to the contrary being explained in the manner of Grice as due to our reluctance to assert that John is happy, *tout court*, lest this give a misleadingly rosy impression. This would be somewhat analogous to a view according to which a wall which is partly white sand partly red counts as white (and also as red). We need not debate its merits since even if this line is taken the irreducibility of 'happy that *p*' to 'happy because *p*' is not threatened.

13. In her discussion (Greenspan [1980], p. 229f.) of conflicting emotions, Greenspan writes that 'philosophers may try to dismiss their logical conflict by fiddling with the object of contrary emotions, building into it the reasons for a positive or negative reaction, so that contrary emotions may be directed toward different objects'. There is no 'fiddling' going on here: one just *can* be happy that *p* but unhappy that *q*. The element of conflict that Greenspan worries is being buried here emerges when the question of which affective attitude is to be chosen as determining affective tone, a question raised in these terms below. It remains an interesting further question, not considered here, why only *some* instances of (e.g.) being glad that *p* and being sorry that *q* issue in the experience we describe as having mixed feelings.

14. If some such suggestion is correct, this partly legitimates the 'fudging' in Section 3, between affective states and affective propositional attitudes over the interpretation of such principles as 'Go For It'. One imagines the situation of learning that *p*, when one has been trying to realize the desire that *p*: this is a situation in which one's thought is appropriately focussed on the fact that *p*, so that being happy and being happy that *p* coincide, according to the suggestion.

15. Here we have attempted to throw light on being happy that *p* by saying something about being happy *tout court*; for an attempt to reverse the direction of explanation, giving an account of happiness in terms of *being happy that*, see Davis [1981]. There need be no tension between the two proposals, though the latter account – which, as Davis spells it out, measures happiness by the extent to which one is happy that *p*, over the various propositions *p* that one occurrently believes – risks losing sight of the affectivity involved in being happy. The remarks in Section VI of Davis' paper about 'looking on the bright side' are very much in the same spirit as the discussion of relief and consolation below.

16. *Cf.* the talk of appropriateness of emotion on the editor's introduction to Rorty [1980].

17. I am very grateful to discussions with Michael Smith on the topics of this paper; notes 10 and 11 were directly prompted by his comments.

'Wanting, Getting, Having': Updates and Afterthoughts

The discussion of 'have' as a dummy verb in §2 is reminiscent of something to have been published since the original paper appeared (and entirely independently of it). On p. 116f. of Dixon [1991] the following we read:

> *Have* refers not to an activity but to a general relationship between two roles; it has a very wide semantic range, e.g. *I have a stereo*/*a daughter*/*a headache*/*a wonderful idea*/*a good dentist*. (…) The general relationship indicated by have can alternatively be expressed by the clitic *'s* (which goes onto the last word of an NP), by the preposition *of*, or by the verb *belong to*.

I agree with everything in this passage except for the last remark: John's headache— the headache John now has—is hardly something which belongs to him. The ellipsis marked by the '(…)' contains an observation of some incidental interest. Dixon writes that "*Have* can substitute for most (perhaps all) instances of *possess*, but only for *own* in general statements), not in deictically referring expressions (*John owns that car*, scarcely **John has that car*)."

From *Dialectica* **42** (1988), 183–200. (At the time of this writing only articles published in Dialectica since 1992 are in the on-line archive at http://www.blackwell-synergy.com/loi/DLTC.)
Notes begin on p. 74.

CHAPTER 4. SOME EPISTEMIC CAPACITIES

1. Introduction

The following is a convenient terminology with which to mark certain familiar distinctions.[1] Say that an individual has a *positive recognitional capacity* with respect to a predicate Φ if he can recognize any Φ object[2] with which he is presented as a Φ object, a *negative recognitional capacity with* respect to Φ if he can recognize any non-Φ object he is presented with as not being Φ, and a *general recognitional capacity* with respect to Φ if he can tell of any object with which he is presented whether or not it is Φ. The terms defined here will be abbreviated to 'p.r.c.', 'n.r.c.' and 'g.r.c.'. Note that the definitions imply that one has a g.r.c. with respect to Φ iff one has both a p.r.c. and a n.r.c. with respect to Φ. Two further clarificatory remarks are in order. First, the terms 'recognize' and 'tell' in the definitions are to be understood as requiring that the subject comes by *knowledge* (rather than, e.g., merely true belief) that the object presented is, or that it is not, Φ, through exercise of the capacity in question, or at least – for he may already have the appropriate knowledge concerning that thing – that exercising the capacity *would* suffice for the acquisition of knowledge. Secondly, note that the defined expressions are intensional in predicate position, attribution, for example, of a p.r.c. with respect to Φ not sufficing for attribution of a p.r.c. with respect to Φ not sufficing for attribution of a p.r.c. with respect to some co-extensive predicate Ψ.

I want to consider a particular argument concerning the relations between these various capacities, which from the assumption that an individual possesses, and knows that he possesses, a p.r.c. with respect to some predicate Φ, moves to the conclusion that the individual concerned has, at least if he is able to attend accurately to his own state of mind and prepared to draw the logical consequences of his beliefs, a g.r.c. with respect to Φ. Taking this 'at least if' qualification as read, and noting that one cannot know one has something unless one has it, the argument seeks to establish the anyone who knows he has a p.r.c. with respect to Φ has a g.r.c. with respect to Φ (by conditional proof, discharging the above assumption). The argument would thus establish, for example, that if you can tell of any divorced woman you meet (say, by a distinctive 'look' about her) that she is divorced, and you know you can do this, then you can tell of any arbitrary woman you meet whether or not she is divorced (or more accurately: of any individual whether or not that individual is a divorced woman). The argument is very simple, and is perhaps best introduced still in terms of this fanciful example. You meet a woman. If she is divorced then, since you have the p.r.c. in question, you can tell that she is. If she isn't, then you

can't tell that she is; but you know, if she were, you *would* be able to; so you are in a position to conclude that she isn't. Further, since you really do have the required p.r.c., and also the knowledge that you do, it seems that your conclusion counts also as knowledge. In schematic terms the proponent of the thesis that a known-p.r.c. implies a g.r.c. argues thus: it suffices to show that known possession of a p.r.c. with respect to Φ implies possession of the corresponding n.r.c.; but the subject cam always interpret his failure to recognize as Φ any non-Φ object with which he is presented as evidence, which, since he knows he has the p.r.c., he can justifiably take as conclusive evidence, that the object is not Φ.

Regard this initial presentation of what we may call the *Basic Argument* as serving, for the moment, mainly as a familiarization exercise in the new terminology. The conclusion of the argument, to the effect that, for any choice of Φ, any (sufficiently rational and self-aware) individual who knows he has a p.r.c. with respect to Φ has a g.r.c. with respect to Φ, we may call the 'known p.r.c.' thesis. In §2 we will consider how well the Basic Argument in support of this thesis fares against objections. I will mainly be concerned to emphasize the availability of replies to such objections, since I am sympathetic to the spirit of this argument, though not committed to the letter of the above formulation of it. In §3, we consider two objections in the form of (putative) direct counterexample to the known-p.r.c. thesis itself. Of course, if successful, these would show that there was something wrong with the Basic Argument, since they would show it to issue in a false conclusion; all the same it will be convenient to consider them together in a separate section. First, some preliminary remarks.

(1.) It is possible to become confused over the p.r.c./g.r.c. distinction by a double usage of 'if' in English (with a corresponding ambiguity in some other languages). On the one hand there is the genuinely conditional usage; on the other, the use of 'if' to mean 'whether'. Intonation and tempo will usually give it away in speech which use is intended, with the presence or absence of a comma doing the job in writing. Thus we have, for example, "He can always tell if she's lying" (attributing a g.r.c.), *vs.* "He can always tell, if she's lying" (attributing a p.r.c.); the if-clause can move to the front of the sentence only in the latter sort of construction.

(2.) Since there is an evident similarity between the p.r.c./g.r.c. distinction and the distinction in the theory of computability between the concepts (respectively) *recursively enumerable* ('semi-decidable') and *recursive (*'decidable'), I will restrict myself here to the disanalogies. (*a*) The intensionality of the present concepts has already been remarked on. (*b*) What keeps recursively enumerable sets from all being recursive is the irreducible variability in lengths of computation, so that the absence of a 'yes' answer by a certain stage cannot be interpreted as a 'no', since a 'yes' may yet be forthcoming at a still later stage. This feature is entirely absent from the situation we are considering, and for examples of the kind we are interested in there is no harm in making it explicit as an assumption for any individual with a p.r.c. with respect to Φ that there is some fixed length of time such that for any Φ object he is presented with, he can tell that it is Φ within that length of time.[3] In the face of this disanalogy with the computability theorist's distinction, our distinction may seemed threatened with collapse, but note that (*c*) Even the weaker notion of semidecidability for a set S is characterized using a biconditional: we require the existence of an algorithm which for any input will yield a 'yes' answer after finitely many steps if *and only if* the input denotes an element of S; whereas our potential Φ-recognizer

was just required to say 'yes' if presented with a Φ object. (Problems of overclassification will occupy us in §2.)

(3.) The Basic Argument may seem suspiciously to conjure knowledge out of ignorance, since it passes from a failure to recognize something as Φ to successful recognition of it as non-Φ. When people – rightly or wrongly – criticized the principle of indifference as involving some such conjuring trick, they had in mind the illicitness of relabelling ignorance (of which of several alternatives was more probable than the others) as knowledge (that they were equally probable). The sense in which the argument with which we are concerned gets 'knowledge out of ignorance' is quite different: it is not a question of relabelling anything but of inferring something from the fact that one is ignorant of something else. There is certainly nothing illicit about this. Many-person analogues of the kind of case we are interested in are quite common, where someone infers that something must be the case with regard to a certain subject-matter because of someone else's ignorance of a fact relating to that subject-matter. Think of those puzzles which have people standing around looking at each other trying to work out what colour hat they are wearing on the basis of (i) knowledge of constraints they are assured tile colour distribution will satisfy, (ii) what colour they can see other people's hats are and – the important point here – (iii) such indirect information as they can glean about their hat's colour from the fact that others have not been able to work out theirs[4]. Or, consider this hackneyed little dialogue:

A: Are you in love with her?

B: I just don't know.

A: Then you can't be, because when anyone's in love, they know they are.

Of course, this is strikingly like the argument with which we are concerned, except that here (A) is helping (B) out with some of the reasoning. (I doubt if anyone in B's position has ever been persuaded by this sort of argument, there being so little reason to believe in the allegedly universal p.r.c.).

2. To and Fro over the Basic Argument

There is an obvious weakness in the Basic Argument from known-p.r.c. to g.r.c. The argument seeks to get you from your failure to recognize an object as Φ to recognizing it as non-Φ), with the aid of the auxiliary premise that you have the p.r.c. and know that you have, since you can use this knowledge to infer that if the object had been Φ you would have recognized it as such. Now, to be able to use this knowledge you must be *aware* of not recognizing the presented object as Φ. However, all that you could be in a position to be aware of is that you have not formed the *belief* that it is Φ; you can pass from this to being aware that you have not recognized the object as Φ, since belief is required for knowledge. But it may happen that you do not recognize the object as Φ not because you lack this belief but that for one reason or another, the belief does not constitute knowledge. Especially important for present purposes is

that the belief is false: the object is not Φ. In any such case, your failure to recognize the object as Φ will not be available to you as a datum from which to proceed as the Basic Argument supposes. In terms of epistemic logic, what we have here is a tacit appeal to the notoriously invalid **S5** principle for knowledge: if you don't know that *p*, then you know that you don't know that *p*.[5] (This principle is ridiculous because it means that when you think you know something you in fact don't know, you also know – and therefore think – that you don't know it.)

Let us put the difficulty in terms of a concrete case. A final year medical student may have learnt enough about measles to be able to diagnose successfully any case of measles she is presented with. Thus she may have, and indeed know that she has, a p.r.c. with respect to cases of measles (alternatively: with respect to measles-sufferers). We count her verdict, in passing such a diagnosis in such a case, as expressing knowledge because it is appropriately based on evidence. Now suppose that there is another disease – jungle fever, say, which presents symptoms so like those of measles that the student, who has in any case never heard of jungle fever, diagnoses those few patients she sees suffering from this disease as suffering instead from measles. So from what was said earlier, we have someone who knows she has a p.r.c. with respect to measles, but from what has just been said, who lacks the g.r.c. because there are some non-measles cases (namely, the jungle fever cases) which she cannot recognize as non-measles cases. How should the proponent of the thesis respond to this kind of example?

There is an obvious line of reply to try, arising out of attention to the discriminatory aspect of the concept of knowledge.[6] Since – the reply might run – the student cannot tell the difference between measles symptoms and jungle fever symptoms, she does not after know of every measles case she is presented with that the case is one of measles, rather than jungle fever: the most she could be said to know – and this in turn might have to be weakened if further indistinguishable ailments are added to the list – is that she is faced with a measles-or-jungle-fever sufferer.

Or so, at any rate, the defender of the known-p.r.c. thesis might argue, here appearing to avoid the charge of committing the '**S5** fallacy' in its defence. More generally, the vindication of the thesis is seen to depend on this principle: where there is a proneness to errors of overclassification, even correct classification does not constitute knowledge. That is to say, there cannot be possession of the universal ability to recognize Φ things as such. There is no intention to deny that some recognition of Φ as Φ is secure in spite of a tendency to overclassify, as long as the cases in question are distinguishable by the subject from the non-Φ things which, when she errs, she classifies as Φ. Goldman (in the work cited in note 6) gives the example of a man who cannot tell alsatians (German shepherds) from wolves as not knowing of an alsatian in front of him that it is a dog, whatever the man's belief may happen to be, but who is not thereby deprived of the ability to recognize a dachshund as a dog.

From such considerations there may yet be extracted an objection to tile known-p.r.c. thesis. To elaborate: Goldman not only drew attention to the discriminatory aspect of knowledge exploited by the proponent of the known-p.r.c. thesis in rebutting the suggestion that the medical student example offered a counterexample to that thesis, he also made the further observation that we only require the knower to be able to tell the kind of case before him *from relevant*

possible cases that would mislead him. Quite what constitutes relevance here is unclear, but it has been suggested that we rule out as irrelevant cases which the knower would be extremely unlikely to encounter, as the world is at the time at which he is being said to know whatever it is. Goldman's famous example illustrating this sort of ruling out involves the person who has just driven into a region of the United States where, amongst the ordinary farmers' barns, and from the road indistinguishable from them, are located *papier-mâché* barn-façades. In this case, although he may happen to be looking at a genuine barn and have visual evidence which in normal circumstances would suffice for the belief he expresses by "That's a barn" to constitute knowledge, we are reluctant to say that he knows it's a barn. Goldman suggests – and isn't he right about this? – that we show much less reluctance on this score if we make the assumption that the only facsimile barns in existence are thousands of miles from our American driver, over in Sweden.

Accordingly, runs the objection, we should now consider a variation on the example of the medical student unable to tell measles from fever in which it is specified that jungle fever is found only in (say) New Guinea, just as Goldman considered removing all the facsimile barns to Sweden. The example now appears quite different in force from the case in which perhaps every hundredth apparent measles sufferer in the Aberdeen area (where she lives) is in fact a victim of jungle fever. We may agree that in the latter case, we do not get a counterexample to the thesis, since the subject does not succeed in having the p.r.c. required, being unable to tell cases of measles from relevant misleading non-measles cases, but what about the former case? If, as it has been set up, you are still inclined to take the New Guinean jungle fever sufferers as relevantly in need of being distinguished by the student from measles sufferers, then suppose that not only has jungle fever never been known outside New Guinea, but in addition that it has been extinct there for a hundred years; or … I shall simply write on the assumption that the original New Guinea case has jungle-fever sufferers relegated to the status of examples of non-measles-sufferers sufficiently remote to leave intact the student's ability to recognize measles when she sees it. (I don't mean 'spatially remote', but that the possibility of her encountering them counts as sufficiently remote). Do we not have here, the objection proceeds, a counterexample to the known-p.r.c. thesis? Having made jungle fever too *recherché* a possibility for confusion with measles to undermine the student's p.r.c. with respect to measles, or indeed her knowledge that she has this capacity, does it not remain the case that she lacks the corresponding n.r.c.? For are there not some non-measles-sufferers who, were she to be presented with them, would elicit an incorrect measles diagnosis? Intriguingly, there may be, as we shall see below after a discussion of some counterfactual constructions, a way of taking the attribution of recognitional capacities that does not force us under those conditions to deny that the student has the n.r.c. in question.

The question we must now address is that of how attribution of a p.r.c. is to be construed in order for a reply along the lines envisaged above, in terms of excluding remote possibilities as irrelevant, to be available. After all, taken literally, it would seem that to have such a capacity with respect to Φ is to be able, for *any* object which is Φ, to tell on being presented with the object for inspection, that that object is Φ. We cannot exclude the irrelevant candidates if we say 'any'; yet it would seem *ad hoc* to explicitly disallow the counterexamples. By looking at some aspects of the role of counterfactuals in a possible formulation of p.r.c. attributions, it will

become evident that there is a natural formulation available which has the effect of *implicitly* excluding the unwanted cases. While the issue raised here affects negative and general recognitional capacities as well, we focus on the positive case in our discussion.

We use the symbol '$\Box\!\!\rightarrow$' for the subjunctive conditional, reading it as David Lewis does,[7] and use 'Pres(x)' to mean that the subject is presented with the object x, and 'K' for the epistemic operator 'the subject knows that'. As candidates for an appropriate precise formulation of the claim that the subject has a p.r.c. with respect to Φ, we might consider choosing from amongst:

(1) $\forall x(\Phi x \mathrel{\Box\!\!\rightarrow} (\mathrm{Pres}(x) \mathrel{\Box\!\!\rightarrow} \mathrm{K}\Phi x))$

(2) $\forall x(\Phi x \mathrel{\Box\!\!\rightarrow} (\mathrm{Pres}(x) \rightarrow \mathrm{K}\Phi x))$

(3) $\forall x(\Phi x \rightarrow (\mathrm{Pres}(x) \mathrel{\Box\!\!\rightarrow} \mathrm{K}\Phi x))$

(4) $\forall x((\Phi x \wedge \mathrm{Pres}(x)) \mathrel{\Box\!\!\rightarrow} \mathrm{K}\Phi x))$

(5) $\forall x(\mathrm{Pres}(x) \mathrel{\Box\!\!\rightarrow} (\Phi x \mathrel{\Box\!\!\rightarrow} \mathrm{K}\Phi x))$

(6) $\forall x(\mathrm{Pres}(x) \mathrel{\Box\!\!\rightarrow} (\Phi x \rightarrow \mathrm{K}\Phi x))$

(7) $\forall x(\mathrm{Pres}(x) \rightarrow (\Phi x \mathrel{\Box\!\!\rightarrow} \mathrm{K}\Phi x))$

All these formulations would come out equivalent if we were to read '$\Box\!\!\rightarrow$' as (like '$\rightarrow$') expressing material implication, in view of familiar portation principles. But such are the complexities of the logic of intensional conditionals that no two of them are equivalent in any of the systems of counterfactual logic presented in Lewis (*op. cit.*) though there is some collapse in the presence of additional assumptions it may, in the present context, be plausible to make, such as:

(Inv.) $\forall x(\Phi x \rightarrow (\mathrm{Pres}(x) \mathrel{\Box\!\!\rightarrow} \Phi x))$

or the strengthening of this assumption we get by converting the first '$\rightarrow$' into a '$\Box\!\!\rightarrow$'. The assumption is that the property Φ is 'presentation-invariant': things with the property would not lose it if presented for inspection. (On one understanding of 'grue', this predicate will not replace 'Φ' in (Inv.) to yield a truth; no doubt other more natural examples can be found, depending on what exactly one thinks of 'Pres(x)' as coming down to. But it is reasonable here to set such cases aside as peripheral.)

The point to note about (1)–(7) above is the effect of the initial universal quantifier in each of them, which may be held to require too much on the part of the recognizer. Weakening this requirement makes room for a certain reply by a proponent of the known-p.r.c. thesis to our case of the measles-recognizing medical student. This reply would concede, unlike the defence considered earlier, that the student does have a p.r.c. with respect to measles but *deny* that she lacks the corresponding n.r.c., in the case where jungle fever is confined to New Guinea. I am taking it that the question of how to formulate an n.r.c.-attribution is determined by a decision on

the question for p.r.c.'s., since an n.r.c. is just p.r.c. with respect to the complementary property. The line of reply now envisaged will maintain that there is some sense in which it is true that the student *would* recognize as a non-measles case any non-measles case she came across, even though jungle fever would confuse her, on the grounds that if she were to come across ('be presented with') a non-measles case, it would not be one of the jungle fever cases. It is far from obvious how the precise conditional construction involved here is to be represented in the language in which (1)–(7) above are cast; the difficulties will emerge if we consider the following example, well removed in subject-matter from the epistemic capacities with which we have been occupied.

The American Embassy has been seized by terrorists in some middle-Eastern country in an effort to secure compliance by the U.S. Government over some of their political demands. To emphasize their seriousness they have grouped the embassy staff into pairs, and ranked the pairs in order: if the demands are not granted within a week, the first pair will be shot; then, if no action is forthcoming in the second week, the second pair will be shot, and so on. As it happens, the first pair (and no other pair) on the list is a married couple. Consider the conditional:

(C) If an American male hostage were to be shot, so would his wife be.

In terms of the similarity semantics for counterfactuals, and with what Lewis (*op. cit.*) calls the Limit Assumption in force, we can say that (C) requires for its truth that in the most similar worlds to the actual world in which an American male hostage is shot (over the period of the troubles) his wife is shot also. Suppose that in the actual world a rescue party manages to free the hostages before the deadline for the first pair of executions expires, contingency plans for a second rescue mission before the end of the second week having been drawn up in case the first plan should fail. In these circumstances, it is reasonable to hold (C), in the way I am understanding it, to be true. Let us try to represent it in the quantified $\Box\!\!\rightarrow$-language. To avoid exposing irrelevant structure, we may take the quantifiers as tacitly restricted to ranging over American male hostages, with 'Φx' for 'x is shot' and 'Ψx' for 'x's wife is shot'. Notoriously, if we try something directly mirroring the structure of (C) we end up with the unwantedly open formula (C1):

(C1) $\exists x(\Phi x) \Box\!\!\rightarrow \Psi x$

But the usual strategy at this sort of point, rebracketing and changing quantifiers, to yield:

(C2) $\forall x(\Phi x \Box\!\!\rightarrow \Psi x)$

gives us something too strong, and indeed false in the case as envisaged, since, not every hostage being spouse-paired, it is not true that for any American male hostage, if he were to be shot, so would his wife be. Reverting, in the face of this unwanted universality, to an existential quantifier, but giving it still broad scope, gives us:

(C3) $\exists x(\Phi x \Box\!\!\rightarrow \Psi x)$

which is now too weak, since it would have been true if not the first man, but the second man on the list, had been paired with his wife, for now someone is indeed such that in the worlds most closely resembling the actual world in which he is killed, his wife is too. If we are prepared to deviate more considerably from the structure of (C), a rendering may be possible with some repetition of constituents, along the lines of:

(C4) $\forall x[\exists y \Phi y \mathrel{\Box\!\!\rightarrow} \Psi x) \rightarrow (\Phi x \mathrel{\Box\!\!\rightarrow} \Psi x)$

or perhaps that a variant with the initial '$\Box\!\!\rightarrow$' replaced by '$\Diamond\!\!\rightarrow$' which does at least appear to have the right truth-conditions.

We now return to the Aberdeen medical student. The purpose of isolating the special sense of (C)-like constructions is to offer such a sense to the attribution of a n.r.c. with respect to measles in the hope of having something which is true given the facts as imagined in connexion with that example. That attribution runs: if she were presented with a non-measles sufferer, she would recognise that individual as a non-measles sufferer. Now it looked initially as though this had to be false, because if the student were presented with a New Guinean jungle fever sufferer, she would not recognize the individual as not suffering from measles. But this is like taking (C) as amounting to (C2), which, it was urged, is too strong for at least one possible construal of (C). Rather, we are interested in worlds closest – or at least reasonably close – to the actual in which a non-measles sufferer is presented to the student, and we want to say that in such worlds, the presentee will be correctly recognized by the student as not suffering from measles. It is plausible to hold that this is the case; for these non-measles sufferers will comprise such people as Aberdeen school children with heavy chest colds, housewives with rheumatism, and perfectly healthy bus conductors. You can of course add to this list those suffering from ailments in danger of being confused with measles – such as, for example, German measles. These our student had better be able to fell from measles itself in order to be granted the p.r.c.; but you are not at liberty to add jungle fever to the list because, as we are imagining things, if a non-measles sufferer were presented to the student it would not be a jungle fever victim. Here we have an attempt to marry the sense in which jungle fever encounters are too remote a possibility to count as relevant (to use Goldman's term) for undermining knowledge claims, with the notion of remoteness made available by the similarity-metric in play in the Lewis-Stalnaker possible worlds semantics for conditionals. Obviously much more would need to be said to establish that such a marriage can be made to work, but it looks provisionally like a promising line a proponent of the known-p.r.c. thesis might develop to defend that thesis against apparent counterexamples like that provided by the aspiring measles-diagnoser.

A second worry about the known-p.r.c. thesis arises out of Goldman's observation that we do not deny that someone unable to tell alsatians from wolves may know that the dachshund in front of him is a dog. Now suppose that there are no alsatians, but our subject is still apt to judge wolves he encounters to be dogs. Suppose in fact, that there are no dogs at all that he confuses with wolves. All dogs are, in this respect, in the same category as dachshunds, so, consistently with Goldman's observation, we are presumably forced to say of this individual in the alsatian-free world, that he has a p.r.c. with respect to dogs (or with respect to the predicate 'is a dog', as the official formulation runs). Since he mistakenly takes some non-dogs, namely wolves, to be dogs, he lacks the n.r.c. with respect to dogs. Do we not here have a simple counterexample to the thesis? Well, not quite yet, for we have yet to make out a case for his not only having, but also knowing that he has the p.r.c. in question. And perhaps this is a simple matter. Could he not become a certificated dog-recognizer by attending the local dog shows and identifying samples of every breed correctly? To get his certificate, he has to give his reasons for thinking that this or that animal is a dog, and he cites just the sort of features as anyone else would for saying that this poodle or that corgi is a dog, so convincing the examiners that he has

the p.r.c.; the stringency of the examination and the fact that he got his certificate appear to give him good reasons for believing, and believing truly, that he has the p.r.c. So perhaps he knows he has. But let us take a closer look, supposing, for simplicity, that the only kinds of dogs there are are poodles, corgis, and spaniels. Using obvious abbreviations, (*a*) is true:

(*a*) $\forall x[\mathrm{Dog}(x) \rightarrow (\mathrm{P}x \vee \mathrm{C}x \vee \mathrm{S}x)]$

and the circumstances also make true not only (*b*) but also (*c*):

(*b*) $\mathrm{K}\forall x[(\mathrm{P}x \vee \mathrm{C}x \vee \mathrm{S}x) \rightarrow \mathrm{K}(\mathrm{P}x \vee \mathrm{C}x \vee \mathrm{S}x)]$

(*c*) $\mathrm{K}\forall x[(\mathrm{P}x \vee \mathrm{C}x \vee \mathrm{S}x) \rightarrow \mathrm{K}(\mathrm{Dog}(x))]$

where for simplicity of exposition I have opted for a form of representation which eschews '$\Box\!\!\rightarrow$' and 'Pres', as not being relevant here. (*c*) follows from (b) in the circumstances envisaged, since we may accept that the subject knows that all poodles, corgis, and spaniels are dogs. Now from (*b*) minus the outermost 'K', and therefore from (*b*) itself, there follows, given (*a*):

(*d*) $\forall x[\mathrm{Dog}(x) \rightarrow \mathrm{K}(\mathrm{P}x \vee \mathrm{C}x \vee \mathrm{S}x)]$

from which in turn, for the reason just given for passing from (*b*) to (*c*), it follows that:

(*e*) $\forall x[\mathrm{Dog}(x) \rightarrow \mathrm{K}(\mathrm{Dog}(x)]$

Here we have arrived at the concession that the subject does indeed have a p.r.c. with respect to dogs. But how, from the materials assembled here as does the contested conclusion formulable as (*e*) prefixed with a 'K' follow? We cannot of course simply replace '$\mathrm{P}x \vee \mathrm{C}x \vee \mathrm{S}x$' in (*c*) by the co-extensive (according to (*a*) '$\mathrm{Dog}(x)$', since we are within the scope of a 'K', and while we used, in moving from (*b*) to (*c*) the assumption that

(*f*) $\mathrm{K}\forall x[(\mathrm{P}x \vee \mathrm{C}x \vee \mathrm{S}x) \rightarrow \mathrm{Dog}(x)]$

what we should need here to justify a similar replacement of '$(\mathrm{P}x \vee \mathrm{C}x \vee \mathrm{S}x)$' by '$\mathrm{Dog}(x)$' would be, not (*f*), but (*f*) with its '$\rightarrow$' reversed. Yet this – the assumption that the subject knows that the only dogs there are are poodles, corgis, and spaniels – is conspicuously something the example forces us to consider as false. Thus this latest objection to the known-p.r.c. thesis cannot be taken as conclusive.

3. Two Final Objections

The two objections reserved for discussion in this section have the following in common: the Φ they focus on is naturally thought of as a property of the recognizing subject rather than of some external object under examination. A consideration of the first objection, suggested by Bernard Williams' discussion of anoxia and related conditions, will draw out an assumption not so far commented on, but needed for anything resembling the Basic Argument to succeed. This will have the effect of forcing a qualification of the known-p.r.c. thesis. The second difficulty to be raised for the thesis arises from taking as properties whose recognition is at issue properties already characterized in terms of the presence or absence of knowledge on the subject's part.

After introducing (something tantamount to) the distinction between positive and negative recognitional capacities in his discussion of Descartes' views on dreaming,[8] Williams considers a range of cases in which a subject with a p.r.c. lacks the corresponding n.r.c., and muses in a memorable passage:

> I recall a lecture on the symptoms of anoxia (lack of oxygen), against which high-altitude pilots have to be on their guard. One symptom was blue finger-nails; another was overconfidence, which led one to neglect such things as blue finger-nails. On a rather idealized version of this phenomenon, it might well be that A could not tell that he was anoxic when he was; but it would surely be paradoxical to suggest that therefore A could not tell that he was not anoxic when he was not (for instance, A is you, now).

If we agree with what Williams here writes, then the known-p.r.c. thesis may appear to be in jeopardy, since one could not only have the p.r.c. with respect to non-anoxic states of oneself, but know that one had this p.r.c., while apparently lacking the n.r.c. with respect to non-anoxic states (which is the p.r.c. with respect to anoxia, cancelling the negations).

Actually the passage quoted is rather slippery. We are to find it paradoxical that A "could not tell he was not anoxic when he was not", and told to suppose, to reinforce this, that A is *you,* sitting comfortably in a chair reading Williams' book. This is an invitation to the reader to think: yes, surely it is clear that I know that I am not now suffering from anoxia. But for this to show the words just re-quoted to embody a paradoxical (or counterintuitive) suggestion, they must be interpreted as meaning that A could on no occasion tell that he was not anoxic when he was not – only then would the invitation to the reader work to exhibit a counterexample. On the other hand, for the quoted words to be relevant to showing that the lack of a p.r.c. with respect to anoxia accompanies the possession of an n.r.c. with respect to anoxia, they would need to mean that A *could not on every* occasion tell that he was not anoxic when he was not. For having the n.r.c. presumably amounts to being able to recognize *every* instance of non-anoxia in oneself, not just being able to recognize *some* instance. Imagine the reader of Williams' words to be, though now seated in a comfortable armchair with a copy of *Descartes* in his lap, a professional high-altitude pilot by day, incurring (so the example may be somewhat dated) the anoxia risks of which Williams speaks. Is it so clear that this individual, while flying in a normal state, and free as far as he can tell from any symptoms of anoxia, knows that he does not have anoxia? Wouldn't he be aware that if he were suffering from anoxia, he would make similarly negative judgments as to the presence of symptoms? And aren't these anoxic possibilities relevant alternatives?

These last considerations may make us suspicious about the idea that the p.r.c. with respect to non-anoxia is quite so easy to possess. One would have to deal with possibility of states confusable with anoxia and we should be back in poodles-corgis-&-spaniels territory, with the chance of a concession that the p.r.c. might be *unknowingly* possessed – again, no threat to the known-p.r.c. thesis. Perhaps the questions at the end of the last paragraph could only be answered with more information about the nature of idealized anoxia. This neglect of the blue fingernails (etc.) – is it that they are not seen, or that they are seen but not 'noticed', or that they are noticed but that their state is not then put to use as evidence that one is suffering from anoxia, or … ? The possibilities are considerably narrowed down by a later parenthetical remark Wil-

liams makes while cataloguing various different types of epistemically incapacitating states, to the effect that idealized anoxia illustrates the type of state in which one cannot "rationally *tell that* anything, come to any rational conclusion about anything". Now this characterization (whether or not it is true to the idealization encapsulated in the anoxia passage quoted above) has a very direct bearing on the ability of the Basic Argument of §1 to establish the known-p.r.c. thesis in the unrestricted form in which we have been formulating it. That bearing is negative. When we see the needed adjustment, we shall also see the immunity of the thesis to examples drawn from other parts of Williams' catalogue of recognitional pathologies.

Go back to the Φ-recognizer from §1. He not only has, but knows he has a p.r.c. with respect to Φ. The Basic Argument was that this individual, if rational and self-aware, could tell a non-Φ object when he encountered one, by bringing to bear, in the absence of a positive verdict for the object concerned, his knowledge of the relevant p.r.c., to infer that a negative verdict was called for. But now – who says that he can bring this knowledge to bear? The assumption we make, concerning someone with a known-p.r.c., is that he is presented with a non-Φ; we have not explicitly assumed that *were he to be presented with such a thing* he would still possess both the p.r.c. in question, and also, most crucially, the knowledge that the possessed this p.r.c. But precisely this is what must be assumed in order for the reasoning reconstructed in the Basic Argument to be available. In §2 we had occasion briefly to mention an invariance condition to the effect that an object did not change in respect of the property under investigation upon being presented to the subject. Now we encounter another kind of invariance condition, noting that we have tacitly assumed that the known-p.r.c. thesis has whatever plausibility it has, only when restricted to properties Φ confrontation with non-instances of which leaves unaltered the subject's possession of the p.r.c. with respect to Φ as well as his knowledge of that possession. So we should make this condition on Φ (and the cognizing subject) explicit in the formulation of the known-p.r.c. thesis.

Taking the qualification just indicated as being in force to cover the present cases in which what one is presented with is states of oneself rather than various independently existing objects, we can see how other cases Williams considers might be dealt with. Take the case of being conscious. As Williams points out – concerned, again, to illustrate the possibility of a p.r.c. without the corresponding n.r.c., rather than to undermine the *known-p.r.c.* thesis, one can tell that one is conscious when one is but one cannot tell that one is not when one is not. Since one can, however, know that one has the former capacity, the unqualified p.r.c.-thesis would be falsified by such cases. But the qualified thesis is safe, since being unconscious is something which deprives one of (occurrent) knowledge of anything, one's p.r.c.'s included.[9]

The second of the two problems to be raised here for the known-p.r.c. thesis arise over a very special case of that thesis; it is not clear to me how best to deal with the objection this special case poses, and I shall indicate two different strategies, uncertain as to which (if indeed either) is the appropriate response.

The difficulty is made visible by a crude formulation of the known-p.r.c. thesis in which the only non-truth-functional connective to appear is the operator 'K' ('the subject knows that') of epistemic logic; this formulation is the claim that every instance of the schema (*) is true:

(*) $\quad \mathrm{K}\forall x(\Phi x \rightarrow \mathrm{K}\Phi x) \rightarrow \forall x(\neg\Phi x \rightarrow \mathrm{K}\,\neg\Phi x)$

Here, I have not only taken all conditions materially, but also suppressed the reference to the condition 'Pres(x)' from §2 to the effect that the subject be presented with the object x. Now suppose that we add (*) as a new axiom to a system of quantified epistemic logic having the strength of **S4**. That is, we add all particular instances of the schema (*) as axioms. These include of course all instances of the following more specific form:

(**) $\quad K\forall x(K\Psi x \rightarrow KK\Psi x) \rightarrow \forall x(\neg K\Psi x \rightarrow K\neg K\Psi x)$

But since all instances of the antecedent of (**) are already provable, we may detach their consequents: precisely the **S5** principles whose unacceptability was remarked upon at the beginning of §2. This now turns into a argument against the known-p.r.c. thesis, represented as an unrestricted commitment to the schema (*). For though there is room for argument over the acceptability of the **S4** principle for knowledge, even its detractors would agree that the **S5** principle is much more objectionable. So (*) is to rejected because it leads from an (at worst) mildly implausible assumption to a grossly implausible conclusion.

This argument relies on taking (*) as appropriate in the formalization of the known-p.r.c. thesis. One might question this. One might ask to see explicit mention of the qualification concerning Φ that our discussion of Williams forced us to note, and one might want to see 'Pres' re-instated, along with '$\Box\rightarrow$' for '$\rightarrow$'. But these points do not seem very promising. One strategy along similar lines which does show more promise would be to restore the temporal aspects of the recognition situation which were mentioned in I and promptly shelved. First the subject is presented with the object whose classification is at issue, and then he comes out with a verdict. In the most direct cases, there will be a time lag between presentation and recognition. What we have called the Basic Argument treats an indirect method which relies on observing the direct method to have failed, and therefore involves an even greater lag. We set aside these considerations (note 3), but what might now be noted is that the verdict to the effect that the object is Φ is really a verdict to the effect that the object was Φ at the time it was presented, rather than that it is now (at the time of the verdict) Φ. For all the examples we have mentioned, there has been no point in dwelling on this distinction, since the property in question can be assumed to persist over the whole of the period concerned. Now even if we can make this assumption about a property Ψ, it does not follow that we can make it about the property of being know by the subject to be Ψ. So let us distinguish x's having the property of being known at the time of presentation – call it t – to be Ψ, from whatever is known at the time of the recognition verdict – call it u. A p.r.c. with respect to 'is known by me to be Ψ' then implies that:

(1) $\forall x(K_t\Psi x \rightarrow K_uK_t\Psi x)$

and from the knowledge (over the period from t to u inclusive) that one has this capacity would according to the known-p.r.c. thesis and subject to the qualifications we have already made, give one the n.r.c. recorded in (2):

(2) $\forall x(\neg K_t\Psi x \rightarrow K_u\neg K_t\Psi x)$

Some such temporal shift may be naturally read into natural language conditional constructions, counterfactual or otherwise, as in "If ice is placed by a fire, it melts", which predicts a melting *after* rather than *as* the ice is placed by the fire. But now that we have made this temporal

element explicit, we can see that the references to **S4** and **S5** may not be so pertinent after all. These principles speak of knowledge of knowledge or knowledge of ignorance, with the times of the second order knowledge and of the first order knowledge (or ignorance) are the *same*. It would be require further work to bring the comparative merits of the **S4** and **S5** axioms to bear adversely on the known-p.r.c. thesis *via* a consideration of the transition form (1) to (2).

The second strategy which it might be worth investigating in order to defend the known-p.r.c. thesis against this latest attack would be, unlike that just considered, to concede the substantial correctness of (*) and its instance (**), as commitments of thesis, and to conclude that the apparent greater acceptability of the **S4** principle for knowledge over the **S5** principle is *only* apparent. In the interests of still further simplification, we may strip (*) of its quantificational structure, arriving at (3):

(3) $K(p \rightarrow Kp) \rightarrow (\neg p \rightarrow K\neg p)$

Put '$\Box$' (for 'it is necessary that') where 'K' appears, to obtain an alethic modal principle, which after a contraposition in the consequent, looks like this, abbreviating '$\neg\Box\neg$' to '$\Diamond$':

(Recession Axiom) $\Box(p \rightarrow \Box p) \rightarrow (\Diamond p \rightarrow p)$

This principle has a secure place in modal logic's Hall of Fame on two counts. First, it has figured prominently in the axiomatization of normal modal logics determined by no class of Kripke frames: the so-called 'incomplete' modal logics. (I've named it here after one of its semantic properties to have emerged in the relevant literature.[10]) Its second claim to fame, which goes back many more years,[11] is as the best known example of how a choice between modal logics bears on a philosophical issue other than one in modal metaphysics or philosophical logic. The principle, provable in **S5** but not in **S4**, drives the Modal Ontological Argument (take p = "God exists"). Further, **S4** extended by this principle as a new axiom is *precisely* **S5**. So not only does a choice between these logics affect one's view on the validity of that argument (its soundness being another matter of course), but that choice can be regarded as being a choice over the acceptance of the Recession Axiom, no less that over the more popularly debated '$\neg\Box p \rightarrow \Box\neg\Box p$' and its variants. Given **S4**, rejecting such variants is tantamount to rejecting the Recession Axiom. More to the point here: objections to these variants, coupled with a prior commitment to that axiom, constitute objections to the **S4** axiom. And—the suggested second strategy runs—as in alethic modal logic, so in epistemic logic.

Inconclusive as the discussion has been, I hope at least to have assembled some of the ingredients likely to bear on a considered assessment of the known-p.r.c. thesis, to have shown that thesis not to be as immediately vulnerable to decisive refutation as it might first have appeared, and to have indicated some of its connexions with other central issues in epistemology and epistemic logic.[12]

NOTES

1. Familiar, for example, from Appendix 3 of Williams [1978]. Williams' own views on the subject will be discussed in §3 below.

2. By a 'Φ object' here is meant an object to which the predicate Φ applies; this abuse of language makes possible a convenient formulation. For stylistic variation below, we occasionally speak of recognitional capacities with respect to properties which should be understood as carrying the rider 'under that description', so that having a positive recognitional capacity with respect to the property of being metallic is a matter of ascertaining the presence of that property under the description 'the property of being metallic', and not necessarily also under non-synonymous descriptions picking out the same property. Further, depending on context, 'Φ' is to be taken as schematic either for a predicate or for the name of a property.

3. The above formulation of the Basic Argument ignored this feature, treating the matter of recognition as more or less instantaneous, but strictly the individual with the p.r.c. needs to know not only that he has such a capacity but also what the associated recognition period is, since it is only when he knows that it has elapsed without issuing in a recognition that he can conclude that the object under examination lacks the property he can recognize objects as having.

4. An informative review of such cases and the perplexities they give rise to may be found in Gardner [1977].

5. This matter is briefly alluded to on p. 106 of Hintikka [1962], and fully explained at p. 79 of W. Lenzen [1978]. Recent entrants into epistemic logic from the computer science field remain blissfully ignorant of the point; two (from amongst many) examples that may be cited are the papers Halpern and Moses [1985], and Parikh and Ramanujan [1985]. [[In the published version of the present paper, the authors for the former reference were incorrectly given as D. Harel and A. Pnueli. See *Afterthoughts* to this chapter.]] In the present paper talk of self-awareness and the like as an idealization of the epistemic subject means awareness of the presence and absence of belief, not awareness of the presence or absence of such knowledge.

6. Goldman [1976].

7. Lewis [1973].

8. Appendix 3, 'Dreaming', in Williams, *op. cit.*

9. In the interests of brevity I have made various simplifications in this paragraph. For example, Williams does not consider unconsciousness as such and I here amalgamate several of his cases (death, dreamless sleep, etc.) on the assumption that the reader will be able to find some reading of 'conscious' on which my comments are acceptable. Nor is the parenthetical 'occurrent' quite what is wanted since it is not the occurrent/dispositional distinction that is relevant, at least when that distinction is so drawn that someone who is asleep, as well as someone who is wide awake and thinking out a chess problem, can be said then to know 'dispositionally' what her telephone number is. In the latter case, but not the former, the knowledge in question is currently "on call" and this is what "occurrent" has to mean in the text (however exactly this is to be analysed). There is also the question of what "being presented" with a state of oneself involves: being in the state, or being in a position to be introspectively aware of that state, or actually introspecting, or…

10. See van Benthem [1978] or van Benthem [1984].

11. See for example p. 201 of Prior [1961]. Prior actually develops the point using a close relative of the Recession Axiom as formulated here: his has '$\Box p$' for the final 'p'; against the background of **S4** we need not distinguish these principles (as potential axioms, that is).

12. The material in §§1 and 2 of this paper was written before my attention was drawn – for which I am grateful to David Wiggins – to the relevance of Bernard Williams' discussion. I would also like to thank David Lewis for supplying (C4) in §2 in response to my speculating in an earlier draft that no representation of the intended sense of (C) was available in the language there under discussion. Comments from David Bostock and from John Collins have also led to improvements.

'Some Epistemic Capacities': Updates and Afterthoughts

As mentioned in note 5 of this reproduction of the article, the originally published version contained an erroneous reference at the corresponding point, citing a 1985 paper, entitled 'Towards a Theory of Knowledge and Ignorance', allegedly by D. Harel and A. Pnueli. The correct reference was given to – somewhat embarrassingly – by one of the two authors who really wrote the piece, Joe Halpern. I will quote from the letter Professor Halpern sent me (dated July 12, 1990). The 'comment' referred to is the content of my (original) note 5.

> I found this comment surprising on a number of grounds. For one thing, the paper 'Towards a Theory of Knowledge and Ignorance' was written by me and Yoram Moses, not by Harel and Pnueli. For another, I don't know on what basis you make the claim for "blissful ignorance". While I cannot speak for Rohit Parikh (…), I was quite well aware of both Lenzen's work and Hintikka's work at the time that paper was written (as well as numerous other works on that theme).
>
> It seems to me that the philosophy community (particularly in the work on epistemic logic in the 50s and 60s) has had the view that there is one "right, true" notion of knowledge, and it is their job to explicate this notion. Given that point of view, it makes perfect sense to say that certain principles of knowledge are ridiculous. The view that my colleagues and I have taken is quite different. There are many, closely related, ways that word 'knowledge' is used in English. While these usages share some properties, there is no unique set of properties that characterizes them. Depending on the context or the intended applications, different notions of knowledge are appropriate. (…)
>
> Given this viewpoint, it makes sense to try to construct models for some of the more important variants of these notions of knowledge, rather than arguing about which one is the right, true notion. It turns out that in distributed systems applications, the notion of knowledge that is characterized by the axioms of S5 is quite useful.

The erroneous attribution certainly needs acknowledging, and I do so – somewhat belatedly – in the final section of Humberstone [2002], where I also argue that there is no real need to exploit the extra strength of **S5** over **S4** in the discussions Halpern has in mind.

Another point which could have done with more acknowledgment is that the status of **S4** as an acceptable epistemic logic (by contrast with the unacceptable **S5**) is in fact highly contested in the literature. The characteristic **S4** axiom, $\mathrm{K}p \rightarrow \mathrm{KK}p$, has been subjected to numerous criticisms. Rather than repeat them here, I mention only two pertinent papers: Lemmon [1967] and Williamson [1992]. The former takes explicit issue with an argument from Hintikka [1962] defending the principle.

From *Philosophical Studies* **48** (1985), 401–423. In this reproduction, the Appendix has been omitted. Notes begin on p. 87.

5. THE FORMALITIES OF COLLECTIVE OMNISCIENCE

Consider the following variation on the concept of omniscience, where in describing a being as omniscient I mean, as usual,[1] that whatever is the case, the being knows to be the case: a pair of beings, *a* and *b* are collectively omniscient when whatever is the case is known to be the case by one or other of *a*, *b*. Then if each of *a*, *b* is assumed to know all the logical consequences of what he knows, the further assumption that *a* and *b* are collectively omniscient leads, somewhat surprisingly perhaps, to the conclusion that at least one of them is omniscient *tout court*. For suppose that *a* is not omniscient, so that, to use some obvious notation, for some statement *p*

(i) $p \wedge \neg K_a p$

and that b is not omniscient either, so that for some *q*

(ii) $q \wedge \neg K_b q$

Then, conjoining the first conjuncts of (i), (ii), we have that the antecedent of (iii) is true:

(iii) $(p \wedge q) \rightarrow (K_a(p \wedge q) \vee K_b(p \wedge q))$

a special case of the assumption of collective omniscience for *a* and *b*. But the consequent is false, since its first disjunct conflicts with the second conjunct of (i) and its second disjunct with the second conjunct of (ii). It is in drawing out these conflicts that we make use of the assumption that each of our knowers knows the logical consequences of whatever he knows. This assumption is what epistemologists sometimes call the closure assumption for knowledge, though here it is helpful to make a distinction as modal logicians have done, between a stronger and a weaker version of the idea that knowledge is closed under logical consequences. In general, one describes an operator O (which forms a sentence when attached to a sentence) as *monotonic* (or 'monotone') when OB is a consequence of OA whenever B is a consequence of A, for any A and B, and as regular when OB is a consequence of $OA_1,\ldots,OA_n$ ($n > 1$) whenever B is a consequence of $A_1,..,A_n$, for any $A_1, \ldots,A_n$, B; evidently, though the converse does not hold, any regular operator is monotonic (take $n = 1$), so that regularity for the operators K_a, K_b is the stronger closure assumption, while all that is required for the above argument is monotonicity.[2] It is to be observed, further, that though even this weaker assumption is sometimes held to be an absurdly strong demand on the deductive powers of the normal subjects of knowledge-attributions, the only extent to which it is exploited in the above argument

is inferring knowledge (on the part of *a* and *b*) of a conjunct from knowledge of a conjunction containing that conjunct, and the degree to which this appears an idealization is surely negligible.

The argument called attention to here can be recast in the bimodal logic of the two monotonic operators K_a, K_b, as a deduction of '$(p \rightarrow K_a p) \vee (q \rightarrow K_b q)$' from '$p \rightarrow (K_a p \vee K_b p)$', thus:

(1)	$p \rightarrow (K_a p \vee K_b p)$	
(2)	$(p \wedge q) \rightarrow K_a(p \wedge q) \vee K_b(p \wedge q))$	From 1, substituting $p \wedge q$ for p
(3)	$K_a(p \wedge q) \rightarrow K_a p$	By Monotonicity applied to $(p \wedge q) \rightarrow p$
(4)	$K_b(p \wedge q) \rightarrow K_b q$	By Monotonicity applied to $(p \wedge q) \rightarrow q$
(5)	$(p \wedge q) \rightarrow (K_a p \vee K_b q)$	Truth-funct'l conseq. of 2, 3, 4
(6)	$(p \rightarrow K_a p) \vee (q \rightarrow K_b q)$	Truth-funct'ly equivalent to 5

It is to be noted that neither of the disjuncts of (6) can be deduced from (1), only their disjunction; and in my opening remarks I said if *a* and *b* were collectively omniscient it followed that one or other of them was omniscient, not that one or other of them was such that his omniscience followed from the collective omniscience of the pair. Thus if we think of (1) as an axiom extending the smallest bimodal monotonic logic to give a system we might consider in its own right, we see that this system is, as they say, Halldén-incomplete (or Halldén-unreasonable). it contains disjunctive theorems without containing either disjunct as a theorem even when the disjuncts in question have no propositional variable in common. I have been careful here to describe the system as the extension of the minimal bimodal monotonic logic by the axiom (1), i.e., the smallest bimodal logic containing (1) rather than as the smallest bimodal monotonic logic containing (1) because a glance at the above deduction will reveal that we apply the monotonicity rule to principles (in fact, tautologies) available prior to the addition of (1), not to (1) itself or anything deduced from it.[3] The interest of this is that such an application would not be justified by the idea of monotonicity as formalizing a closure condition, since our assumption is simply that if something is the case it is known to be the case by *a* or by *b*, not that if something is the case then it logically follows that it is known to be the case by *a* or by *b*. This same point may help allay the qualms of those who feel that such systems as the one we have just isolated would be better described as *theories* of the intensional concepts involved than as *logics* in any philosophically committal sense.[4]

Those who don't simply object to the idea of 'epistemic logic' – for whom the last remarks were intended as reassurance – but wholly distrust the intensional idiom in which our discussion has been cast, will have to reconstruct the reasoning in first-order terms, quantifying over statements (or sentences), to which a truth-predicate ('T', let us write) applies, with 'K' reconstrued as a two-place predicate, whose first argument we now elevate from subscript position. The task is then to derive, in the presence of a certain ancillary assumption,

(***) $\forall x\,(Tx \rightarrow Kax) \vee \forall x\,(Tx \rightarrow K\Box)$

from

(**) $\forall x(\mathrm{T}x \rightarrow (\mathrm{K}ax \vee \mathrm{K}\Box))$

The ancillary assumption is designed to cope with the work done by the formation of conjunctions in the modal deduction (1)–(6), saying that for statements x and y there is a statement z which is as good as their conjunction for the purposes of our argument.

(*) $\forall x \forall y \exists z\, [((\mathrm{T}x \wedge \mathrm{T}y) \rightarrow \mathrm{T}z) \wedge ((\mathrm{K}az \rightarrow \mathrm{K}ax) \wedge (\mathrm{K}bz \rightarrow \mathrm{K}by))]$

Notice that we no not need to include in the matrix here the converse of the first conjunct, the fact that conjunctions imply their conjuncts having only been used in the modal deduction to yield a premiss for the application of the monotonicity rule, the conclusion of whose application appears as the second conjunct of (*). The derivation of (***) from (*) and (**) is straightforward and does not need to be reproduced here.

A point that is unlikely to have escaped the reader's attention is the irrelevance of the focus on collectively omniscient pairs in our original argument, which is easily adapted to show for any n, that if there are n individuals who are collectively omniscient (where this phrase has the obvious sense) then at least one of them is omniscient *tout court*. We picked on the $n = 2$ case for ease of exposition (brevity of proof, more precisely). To mirror the argument thus adapted in the first-order setting of the preceding paragraph, the assumption (*) should be modified, for the n knowers case becoming:

$(*)_n$ $\forall x_1,\ldots,\forall \mathrm{x}_n \exists z[((\mathrm{T}x_1 \wedge \ldots \wedge \mathrm{T}x_n) \rightarrow \mathrm{T}z) \wedge ((\mathrm{K}a_1z \rightarrow \mathrm{K}a_1x) \wedge$

$\ldots \wedge (\mathrm{K}a_nz \rightarrow \mathrm{K}a_nx))]$

For no $n > 2$ does this ancillary assumption (or its universal quantification in the a_i places) seem less reasonable than it does for $n = 2$. Since we are using x, y, z, … as sortally restricted variables ranging over statements, we install 'u' as a variable over knowers (or cognizing agents, if you prefer a more neutral term), in order to state the following rather striking – because of its general fallaciousness – quantifier shift principle:

(Q) $\forall x \exists u(\mathrm{K}ux) \rightarrow \exists u\, \forall x(\mathrm{K}ux)$

which is a consequence of the $(*)_n$ assumptions together with the hypothesis that there are only finitely many knowers. For from the hypothesis that there are at most n knowers (in first-order logic with identity: $\exists u_1 \ldots \exists u_n(u = u_1 \vee \ldots \vee u = u_n)$ we can of course deduce anything of the form:

$\forall x \exists u \mathrm{A}(u, x) \rightarrow \exists u_1 \ldots \exists u_n \forall x(\mathrm{A}(u_1, x) \vee \ldots \vee \mathrm{A}(u_n, x))$

and so taking $\mathrm{A}(u, x)$ as '$\mathrm{K}ux$' the antecedent of this schema matches the antecedent of (Q), while from the consequent of this instance of the schema we derive the consequent of (Q) by the argument we have been discussing, since it says that there are n individuals who are collectively omniscient (one of which must therefore, given $(*)_n$ be omniscient *tout court*). So for example, if the total human population spread across time is finite and every truth is sooner or later known by someone then at some time or other there is someone who knows all truths. (Note that we do not require that the totality of truths be finite.[5]) One might expect an argument to such a conclusion to involve some rather strong assumptions about the cumulative nature of epistemic

progress; but no, we draw our conclusion from considerations about the familiar properties of truth-functional conjunction.

We revert, for the remainder of the discussion to the intensional idiom. The reason for stressing the logical properties of the operators K_a, K_b, figuring in our original presentation of the argument was to show how widely the point, made as a point about collective omniscience, applies where nothing resembling (except in respect of monotonicity) knowledge is involved. In particular, notice that we never used the principle that knowledge implies truth, so that a parallel argument would go through, for example, in the case of belief. (We deal with a somewhat different argument – a dual form of that presented above, in fact – for belief, in the following paragraph.) More interestingly, we did not assume *regularity* for K_a or K_b, so that the argument will work for notions like possibility, permissibility and so on, for which such an assumption would be out of the question. (For knowledge and belief, regularity is a natural assumption to experiment with, even if one has doubts as to its plausibility; one could in any case read, e.g., 'K_ap' as saying that from what *a* knows, it follows that *p*.[6]) As an example, take permissibility, assuming, with most work in deontic logic, that this notion is reasonably treated as monotonic. Think of *a* and *b* not as knowers but as moral codes, with 'P_aA' and 'P_bA' as saying that it is permissible according to moral code *a* that it be the case that A, and that it is permissible according to moral code *b* that it be the case that A, respectively. Line (1) above then rewrites as:

$$p \rightarrow (P_ap \vee P_bp)$$

saying that whatever is the case is either permissible according to the one code or else according to the other; the earlier deduction then entitles us to conclude with (6), appropriately rewritten:

$$(p \rightarrow P_ap) \vee (q \rightarrow P_bq)$$

which tells us that at least one of the two codes is such that everything that is the case is permissible according to it. We can mix this case with the original case by combining permissibility according to an unspecified (single) moral code, represented here with an unsubscripted 'P' operator, and knowledge on the part of individual *a* ('K_a', as before), rewriting (1) as:

$$p \rightarrow (Pp \vee K_ap)$$

which says that a is aware of all the transgressions there are, from which we conclude the analogue of (6) to hold:

$$(p \rightarrow Pp) \vee (q \rightarrow K_aq)$$

which states that either there are no transgressions at all, or else a is omniscient, a conclusion one might not have expected to follow without prior acquaintance with the general form of argument in (1)-(6). As a final variation on the theme, let our two monotonic knowers a and b be joined by a 'regular guy', *c* (i.e., we assume regularity for the operator 'K_c'). Then make the assumption that *a* and *b* are 'collectively as knowledgeable as *c*':

$$K_cp \rightarrow (K_ap \vee K_bp)$$

and we can deduce, by a mild variation on (1)-(6) whose reconstruction can safely be left to the reader, either *a* by himself or *b* by himself is at least as knowledgeable as *c*:

$(K_c p \rightarrow K_a p) \vee (K_c q \rightarrow K_b q)$

It has become common, since at least William James's 'The Will to Believe',[7] to discern in the teleology of belief two separate aspects: the goal of having true beliefs, and the goal of avoiding false beliefs. If the notion of knowledge in our argument about collective omniscience is replaced by belief, then (1) says that *a* and *b* are collectively successful in attaining the former goal completely (while (6) says that in that case at least one of them taken in isolation has achieved such success). Now consider the analogue of (1) in connexion with the latter goal: make the hypothesis that whatever both believe is true. This apparently allows that they should each have beliefs not shared by the other which are false. But not so – at least if monotonicity is assumed for the belief attributing operators in question ('B_a' and 'B_b', as we shall write them):

(1′)	$(B_a p \wedge B_b p) \rightarrow p$	
(2′)	$(B_a(p \vee q) \wedge B_b(p \vee q)) \rightarrow (p \vee q)$	From 1′, subst. $p \vee q$ for p
(3′)	$B_a p \rightarrow B_a(p \vee q)$	By Mono. from $p \rightarrow (p \vee q)$
(4′)	$B_b q \rightarrow B_b(p \vee q)$	By Mono. from $q \rightarrow (p \vee q)$
(5′)	$(B_a p \wedge B_b q) \rightarrow (p \vee q)$	Truth-functional conseq. of 2′, 3′ 4′
(6′)	$(B_a p \rightarrow p) \vee (B_b q \rightarrow q)$	Truth-functionally equivalent to 5′

Thus if each of *a* and *b*'s shared beliefs is true, at least one of them has only true beliefs. The magic work performed in the earlier argument by conjunction is performed here by disjunction.[8] An informal presentation of the point, in the style of (i)–(iii) above for this case would run as follows. Suppose each of *a*, *b* is mistaken about something, say, *p*, in *a*'s case, and *q* in *b*'s. Since each of *p*, *q* is false, so is their disjunction: but this disjunction is one of their shared beliefs. So not all their shared beliefs can be true.

The applications of the monotonicity rule here may seem less plausibly an appeal to an incontrovertible deductive closure condition on rational belief than were the analogous steps in our argument (1)–(6). For while few have had occasion to entertain a conception of entailment according to which conjunctions fail to entail their conjuncts, that disjuncts should less obviously entail disjunctions in which they are constituents is evidenced by Parry's system of analytic implication.[9] As I understand it, it is with just such applications in mind as doxastic closure principles that Parry developed the idea. If I do not have the concepts involved in the formulation of a statement, is it so obvious that on pain of irrationality, I must believe the disjunction of that statement with, say, 'Some bulls have horns'? (This problem does not arise with conjunction-to-conjunct inferences; even without assumptions of rationality, a belief to the effect that *p and q* seems *inter alia* to be a belief to the effect that p. 10) Thus the envisaged objection to the deduction (1′)–(6′) as establishing what the surrounding informal gloss suggests would be that analytic implication (or some close relative thereof) rather than material implication is what should be meant by '→' in the formulation (Note 2) of the monotonicity rule, under which construal steps (3′) and (4′) fail. I mention this objection because it occurs to me as an

interesting one, rather than because anything interesting occurs to me to say in response to it. Connectedly, it is worth remarking that the original deduction does not really present a different argument from (1′)–(6′), when the '→' is read, as we have been presuming, as the material conditional. Rather, the latter argument is simply a dualized version of the former, which happens to connect more directly in its opening and closing lines, with easily intelligible theses about the intensional concepts involved. For example, it is the form we should have considered has we restated the argument about permissibility (using P_a, P_b) above in terms of the corresponding notions of obligation ($\neg P_a \neg$, $\neg P_b \neg$)In the present case, the (1)–(6) form of the argument would be what we got if we had formulated the hypothesis (1′) in terms of the dual modalities $\neg B_a \neg$, $\neg B_b \neg$. However, it is to be noted that the contraposition steps presumed available in describing the two forms of argument as equivalent notoriously fail for analytic implication (as contraposing is apt to violate the inclusion of concepts requirement).

The conclusion that if each of two (or of course more) people's shared beliefs are true, then one of them has no false beliefs, seems on the face of it counterintuitive. I think the reason this is so is because we tend to have in mind particular sorts of shared belief, for which is it not the case that truth of all shared beliefs of the given sort implies truth for each of some one individual's beliefs of that sort. For example, consider a group of political commentators with some competence at predicting the outcomes of elections in their country. It might well be that whenever they all agree as to which party will win the election, that party does indeed win. Does it follow that one of their number is such that whatever opinion he forms as to the which side will win is a correct opinion? Not at all: they could all be wrong. This is because not every belief about the outcome of the election counts as a belief 'as to who will win' the election. It is a question of specificity. If three parties, *X*, *Y*, and *Z* are in contest, then a belief to the effect that *X* or *Y* will win is not a belief which is in the intended sense a belief as to which party will win. However confident you are of such a belief, you will not think of yourself as knowing who will win (unless you narrow down your beliefs further and come to think either that party *X* will win or that party *Y* will win). Thus it does not follow from the hypothesis that the group are always right when they agree as to which party will win that they are always right when they agree about the outcome of the election, in this looser sense of 'agree about the outcome', and so the (1′)–(6′) argument does not in fact exclude the possibility of such a group of commentators. (The move to disjunctive belief-contents is the place at which the required specificity is lost.)

The (1′)–(6′) deduction can be run through with desire in place of belief, to identical effect. We represent *a*'s (*b*'s) wanting that A by $W_a A$ ($W_b A$), and take the argument as leading from:

$(W_a p \wedge W_b p) \rightarrow p$

down to:

$(W_a p \rightarrow p) \vee (W_b q \rightarrow q)$,

assuming of course that the 'W' operators, so read, sustain the crucial monotonicity inferences. The plausibility of this assumption may be questioned. For example, Frank Jackson[11] would deny that from *a*'s wanting it to be the case that *p*, it follows that *a* wants it to be the case that *p* or *q* (even for completely rational *a*) since *a* might very much want it to be the case that *p* and think that were it to be the case that *p* or *q*, that would most likely be because it was the case

that q (a prospect to which he is perhaps averse) and not p. For those with such reservations (which I shall not discuss), I suggest reading '$W_a p$' as '*a* has desires whose satisfaction entails that p', so that the monotonicity is guaranteed';[12] I shall continue with the '*a* wants it to be the case that p' rendering, however, in the interests of brevity. Then (1′)–(6′), with 'W' for 'B', opens with the assumption that whatever *a* and *b* both want to be the case is the case, and concludes that either all of *a*'s desires are satisfied or else all of *b*'s desires are. In a highly attenuated sense of 'omnipotent' (better expressed by saying, 'omnicontented') we can say that the deduction shows that if a group of individuals is collectively omnipotent then at least one of the group is omnipotent *tout court*. What makes this an attenuated sense of 'omnipotent' (and 'collectively omnipotent' when this phrase gets the obvious derived sense) is that, even if we were to accept that ability-attributions can be analyzed as conditionals—with a reference to wanting, trying, or intending, in the antecedent—the appropriate conditional is hardly to be taken as the material conditional. 'For all p, $W_a p \rightarrow p$' simply says that a has no unsatisfied desires, a condition which, as is often remarked in philosophical discussions of freedom, could be met by a powerless individual whose lack of abilities is matched by a corresponding lack of desires. So what we would prefer for omnipotence is something along the lines of 'for all p, $W_a p \mathbin{\Box\!\!\rightarrow} p$' where '@' is the counterfactual or subjunctive conditional, here appearing in the notation devised by David Lewis.[13] And, as the reader familiar with Lewis's work may verify, for no plausible account of the logic of this style of conditional will the argument (1′)–(6′) go through when '$\rightarrow$'s are replaced by '$\Box\!\!\rightarrow$'s.

Continuing to ignore this point about '$\Box\!\!\rightarrow$', I cannot resist remarking on the appearance the argument (with 'W') we have been reviewing presents as establishing a poor man's version of an Arrow-style impossibility result, the starting point of which (when numerically generalized) says that whatever everyone wants, goes – a sort of Pareto condition – and whose terminus says that one person is such that whatever that person wants, goes – a violation of the condition of non-dictatorship. This is not really analogous on closer inspection, however, to the situation discussed in the post-Arrow social choice literature, since in the first place it is preferences rather than outright desires that are at issue, and in the second because the relata (of the individual preference relations) are taken to be mutually exclusive outcomes (so that moves like the inference from '$W_a p$' to '$W_a(p \vee q)$' have no analogues. But most importantly, the point about '$\rightarrow$' would need to be brought in here to adequately capture in the present setting the intuitive idea of dictatorship, since this is the social power version of omnipotence. A dictator, in the current technical sense, is not someone such that whatever he personally desires to be the case emerges as being the case (or as being the socially preferred outcome) since this—to repeat the point just made—would be true of the powerless but lucky citizen, but rather someone who is such that whatever he were to want to be the case, would be the case. This distinction, captured here by distinguishing between different types of conditional, is captured in the social choice literature by quantifying over combinations of mathematically possible individual preference orderings, but remains, embarrassingly enough, a distinction which is crucially fudged in many expository discussions of the Arrow material, as was recently pointed out by Richard Routley.[14] Thus our 'poor man's version' turns out not to be as much of a caricature as one might have hoped.

The remainder of my remarks will be organized around three points of connexion between the concepts of omnipotence and omniscience, or, more generally, of ability and knowledge. The third will return us to the argument about collective omniscience (derivation (1)–(6)) with which we began. The first will be brief, since I intend only to mention and not to take up, a thought suggested by the discussion above of the $\rightarrow$ element in attributions of omnipotence (in the non-attenuated sense). The thought is that what I have taken to be omniscience is lower on the scale of excellences in respect of epistemic competence than the label might suggest. We could use the term 'strong' to qualify that omniscience that passes not just my test: that for all p, $p \rightarrow K_a p$, but the strengthened condition we get when this '$\rightarrow$' is replaced by a '$\Box\!\!\rightarrow$'. This brings in a measure of truth-tracking (to use a term from Nozick, *op. cit.*) over and above any such element that may be deemed to lurk within the here unanalyzed '$K_a p$' itself. As in the case lately reviewed, the argument with which we began does not survive this strengthening. What can be said of this notion of strong omniscience? First, that as currently formulated, it is perhaps too strong to be of much interest, though a reasonable weakening survives if the antecedent 'p' is replaced by '$p \wedge a$ exists', since this modification allows the concept to apply to beings existing only contingently. Second, that strong omniscience is of course not to be confused (even if as the result of some argument one may come to identify it) with the notion of necessary omniscience; a similar *prima facie* distinction is to be observed between the modified notion of strong omniscience just introduced and the familiar notion of essential omniscience.[15] But, as I say, I have nothing further to add on this topic, which I leave to your further investigation.

The second point about omniscience and omnipotence is a point of disanalogy, assuming I am right about the need for something like the '$\Box\!\!\rightarrow$' element in the latter concept. There is *prima facie* a contrast in the conceivability of these concepts being multiply instantiated. Two gods might unproblematically be assumed each to be omniscient, while there is something of a problem about an analogous plurality of omnipotent beings. True enough, it might be that whatever either decided to make the case ended up being the case, and even that its being the case was due to that god's so deciding, but such a course of events relies on the happy accident that the two gods never make decisions incapable of joint realization. So we cannot say that whatever either decided in favour of, even if the other decided against it, would end up being the case. I say that this consideration renders problematic, rather than clearly rules out, the possibility of multiple omnipotence, because it may be possible to concede that while it would have to be a contingent fact that the gods' desires never conflicted in this way, it may yet not be accidental, since it might be true that if either were to have a given desire, the other would not have a conflicting desire. I do not intend to discuss further the question of whether such a middle way does make sense of the possibility of multiple omnipotence.

The third point arising out of a comparison between knowledge and ability[16] concerns collaboration. One reasonable sense to give to the attribution of a collective ability to a group of agents would be that they can pool their efforts and resources so as to bring about a certain state of affairs. Clearly this may be the case even though no individual in the group could by his own efforts alone bring about the outcome. A similar interpretation can be given to attributions of collective knowledge: while perhaps no one member of the group knows that p, if they were to get together and pool their epistemic resources, piecing together the information and evidence they severally possessed, they could infer (subject to whatever stringency conditions you care to

impose for such inference to yield knowledge in the one person case) that p. Actually, this counterfactual paraphrase does not quite convey the collaborative interpretation I have in mind for collective knowledge, which is rather one of the notions of impersonal knowledge explored by Risto Hilpinen and is better explicated in the following way, here spelt out only for the case of a two-person group: a and b collectively know that r when there exist $p_1,\ldots,p_m,q_1,\ldots,q_n$ together entailing r and such that a knows that p_i (for each $1 \leq i \leq m$) and b knows that q_j ($1 \leq j \leq n$). The earlier counterfactual gloss is not quite accurate for various obvious reasons, such as that it may beyond their pooled logical abilities to see that the entailment mentioned holds, and that the envisaged act of collaboration would render false one or more of the pooled premisses. If we assume that knowledge for a, b is a regular modal notion then we can simply replace the '$p_1, \ldots,p_m$' by 'p', with a similar simplification in the case of the q_i. (Indeed to get a regular notion of knowledge for a single individual, amounting to the knowledge implicitly possessed by that individual, we can simply use the above route to the concept of what that individual 'collectively' knows (or what that individual and he himself collectively know, to stick pedantically to the form of the definition – which, however, was offered as illustrative of the general case, for fewer than two just as well as for more than two knowers).) If we go further and assume normality, so that the Kripke relational model-theory becomes available, then this concept of collective knowledge collaboratively construed has, as Hilpinen has observed, a very interesting semantics in terms of intersections of accessibility relations.[17] For present purposes, it will be worth devoting a separate paragraph to this general topic, with illustrations drawn from areas other than knowledge, before returning to that particular application of the idea.

Suppose that we have a language, like that for truth-functional logic but equipped with three primitive modal operators $\Box_1$, $\Box_2$, and $\Box_3$, interpreted by models $\langle W, R_1, R_2, R_3, V\rangle$ where the R_i are binary relations on the (non-empty) set W, each element of which, paired with an atomic formula, is an argument for which V is a function delivering as value a truth-value. V is extended to a truth-relation $\models$ for arbitrary formulae by the usual clauses for the truth-functions as well as:

$$\textit{for } w \in W, \models_w \Box_i A \Leftrightarrow \forall y(wR_iy \Rightarrow \models_y A)$$

for $i = 1, 2, 3$. Now we are to consider the set of formulae true at every element of each such model. This is of course precisely the smallest modal logic which is 'normal in' $\Box_1$, $\Box_2$, $\Box_3$, in the sense of being closed under truth-functional consequence as well as each of the three rules (for $i = 1, 2, 3$):

$$\frac{(A_1 \wedge \ldots \wedge A_n) \rightarrow B}{(\Box_i A_1 \wedge \ldots \wedge \Box_i A_n) \rightarrow \Box_i B}$$

If we are interested in the extension of this minimal tri-modal system which will capture precisely the formulae unfalsifiable at any point in each model $\langle W,R_1,R_2,R_3,V\rangle$ in which $R_3 = R_1 \cup R_2$, then one easily sees that the only additional axioms needed are those instantiating the schema: $\Box_3 A \leftrightarrow (\Box_1 A \wedge \Box_2 A)$, a point we could equally well put by saying that the operator whose

accessibility relation is the union of those for operators $\Box_1$, and $\Box_2$ of a bi-modal logic could be introduced into the language of that logic by taking the above biconditional as a definition.[18] Intersections of accessibility relations are much more interesting in that no such definition (it can be shown) is possible, and I shall not here enter into a discussion of the proof-theoretic codification of the set of formulae unfalsifiable in models $\langle W,R_1,R_2,R_3,V\rangle$ in which $R_3 = R_1 \cap R_2$. Rather, I want to consider some cases in which one might want such an intersective operator, which I will now write as '$\Box_1 \cap \Box_2$' instead of the unmemorable '$\Box_3$', an operator, that is, for which an appropriate semantic clause would be:

$$\vDash_w (\Box_1 \cap \Box_2)\mathrm{A} \Leftrightarrow \forall y((wR_1y \;\&\; wR_2y) \Rightarrow \vDash_y \mathrm{A})$$

The corresponding dual operator defined by $(\Diamond_1 \cap \Diamond_2)\mathrm{A} =_{df} \neg(\Box_1 \cap \Box_2)\neg\mathrm{A}$ works out semantically like this:

$$\vDash_w(\Diamond_1 \cap \Diamond_2)\mathrm{A} \Leftrightarrow \exists y((wR_1y \;\&\; wR_2y) \;\&\; \vDash_y \mathrm{A})$$

a point I mention because the first example I want to give of the extent to which these intersective operators are to be found in ordinary language is a deontic one more naturally expressed in terms of permissibility than of obligation. Think of $\Box_1\mathrm{A}$ as saying that according to a certain moral code it ought to be that A, and of $\Box_2\mathrm{A}$ as saying by contrast that a certain legal code requires that A. Now, if the moral code is your own, and the legal code is that under the threat of whose sanctions you live, you may take a special interest in those things which are both legally and morally permissible, and it is my impression that this must be taken in a sense in which to say:

(α) It is morally permissible that A and it is legally permissible that A

is to say something much weaker than to say:

(β) It is both morally and legally permissible that A.

(I am prepared to admit an ambiguity in (β), one sense being that of (α), the other being the stronger sense here at issue.) What I just termed your special interest is in facts of the type (β) rather than of the type (α). For an example, which will admittedly be somewhat farfetched, suppose that the only person it would be morally permissible for you to marry belongs to a category of persons it is legally forbidden that you marry. Then although it is legally permissible for you to get married, and also morally permissible for you to get married, it is – as I am inclined to put it – not legally and morally permissible that you should marry, since there is no way you can get married while fulfilling both your moral and your legal obligations. Putting it in terms of possible worlds: although there are morally permissible worlds in which you marry, and there are legally permissible worlds in which you marry (taking permissibility relative to this world in both cases), there are no worlds which are legally as well as morally permissible in which you marry.[19] For a second example: it might be consistent with what Harry says that *p*,

and consistent with what Mark says that p, but not consistent with what Harry and Mark say (when their testimonies are taken together) that p: and all this even when Harry's testimony and Mark's testimony are consistent with each other. These cases are both of the '$\Diamond_1 \cap \Diamond_2$' type. A case where one might want the dual operator arising from the study of tense logic rather than formalization of ordinary discourse itself is the following. Think of the carrier set W of the models as containing not worlds but moments of time, and of the relations R_1 and R_2 as converses; then the formula $(\Box_1 \cap \Box_2) \bot$, where $\bot$ is any contradiction, is true throughout a model iff R_1 (or equivalently R_2) is asymmetric. (No formula of the usual tense-logical language – in which $\Box_1$ and $\Box_2$ would customarily be written as 'G' and 'H' – has this property.)

Returning finally to the question of collective knowledge on the part of our knowers a and b, their knowledge assumed in each case to satisfy the normality condition and so to be construable in terms respectively of truth at all R_a-related and at all R_b-related worlds, we can think of attribution to them of collective knowledge *à la* Hilpinen that p as of the form: $(K_a \cap K_b)p$, where this is the intersective operator understood as above. Note that, although implied by, this does not imply $K_a p \vee K_b p$. Plugging this notion of collective knowledge into a definition of collective omniscience to give the collaborative conception thereof gives us the following assumption about a and b: for all p, $p \rightarrow (K_a \cap K_b)p$. Unlike our original 'disjunctive' conception, the assumption here does not yield as a conclusion that either a or b is omniscient *simpliciter* (as one may verify model-theoretically: if the intersection of two relations is included in the identity relation, it does not follow that one or of the relations must be included, even if the relations are assumed reflexive).[20] My conclusion then is that there is some notion of joint omniscience, though it is not the one with which we began in this paper, according to which a group thus blessed collectively need not contain a member omniscient individually, and that perhaps this is why it is surprising (because one does not clearly distinguish the two notions) that on the conception of collective omniscience introduced at the opening of the paper, this is not a possibility.[21]

NOTES

1. The logical features of omniscience are discussed in the opening paragraphs of Prior [1968]. The later discussion focusses on problems connecting omniscience with tense and indeterminacy which will not concern us here.

2. On the concepts introduced here, see Chellas [1980]. The regularity property for an operator O may be stated as a rule of proof for a modal logic amongst whose operators O is, thus: for $n \geq 1$ from $\vdash (A_1 \wedge \ldots \wedge A_n) \rightarrow B$ to $\vdash (OA_1 \wedge \ldots \wedge OA_n) \rightarrow OB$, and monotonicity as the restricted form of this rule for n = 1. The '$\rightarrow$' here and in the main text is for material implication. Sometimes epistemologists have something other than either of these notions in mind when they speak of closure: that those consequences of what is known which are known to be consequences are themselves known. Occasionally there is also talk of the subject's actually inferring the consequences from what is presumed already known. Of these

various closure conditions, it should be remarked that certain philosophers have denied even the weakest of them to be appropriate even as principles pertaining to the most ideally rational of knowers; Dretske and Nozick are examples (see pp. 203*ff*. of Nozick [1981].).

3. Adapting Segerberg's use of the prefix 'quasi', one would describe the system in question here as quasi-monotonic; see Segerberg [1971].

4. Some of the issues here have received an airing in Harman [1972] and in Kuhn [1981]).

5. Of course, if there are only finitely many (pairwise non-equivalent) true statements in the language (assumed to be equipped with a conjunction operator with the usual logical properties) then one of these will be equivalent to the conjunction of all of them and the fact that if each is known by someone, someone knows all of them is immediate from a consideration of the individual who knows this conjunctive truth.

6. In his well-known study of these matters, Hintikka [1962] assumes not only regularity, but also normality, i.e., the rule schematically indicated in Note 2 but with '$n \geq 0$' rather than '$n \geq 1$' as side condition. (However, charges of unreasonable idealization on the part of those not familiar with Hintikka's treatment should be held back until they have seen how the formal material is related to the data of propositional attitude ascription *via* the concept of indefensibility.)

7. James [1896]; Axinn [1966].

8. A closer examination of the crucial rôle these connectives (conjunction, disjunction) play in our two deductions (1–6, 1′–6′ respectively) is undertaken in the Appendix to this paper. [[The Appendix has not been reproduced for this chapter.]]

9. Parry [1968].

10. I here express agreement with the central paragraph of p. 211 in Routley and Routley [1975].

11. See Jackson [1984] for an account of wanting which has this upshot; the point is made explicit in Jackson [1985]. [[The latter was described as "Jackson (fortchoming)" in the originally published version of this chapter.]]

12. Note that this reformulation also circumvents any inclusion-of-concepts objections (like those mentioned for belief above, and quite separate from the Jackson objection) to Monotonicity for 'W'.

13. Lewis [1973].

14. In Routley [1979]. Even the more mathematically sophisticated social choice texts exhibit the logical confusion of which Routley complains, such as that – not mentioned by him – by Kelly [1978], which, at pp. 10–11 incorrectly characterizes non-dictatorship in terms of a notion of decisiveness defined in an intra-profile manner, where an appropriate characterization would require inter-profile relations (quantification over alternative combinations of individual preferences, that is).

15. Personally, I am doubtful of the coherence of the notion of essential omnipotence: a being who is a fitting subject of knowledge-attributions seems thereby to be a being logically capable

of forming erroneous beliefs, even if the being never in fact forms such beliefs. Similarly with the capacity for mere ignorance (rather than false belief). There are strong arguments against this line, however, which I cannot take up here.

16. This talk of analogies and disanalogies between knowledge (or omniscience) and ability (or omnipotence) is intended to be non-committal with respect to the question of whether knowledge (and hence omniscience) is itself species of ability (such as the ability to answer, if only in thought, this or that question).

17. See Hilpinen [1969], [1977], and, for an explicit discussion of the point about intersections of accessibility relations, Hilpinen [1974].

18. This equivalence is exploited in dynamic logic, e.g., as axiom A9 at p. 147 of Goldblatt [1982].

19. A deontic example of a rather different type is given in the paper by Hilpinen cited last in Note 17; there the intersection taken is of the accessibility corresponding to the moral obligations of various different agents.

20. I do not have strong feelings as to whether the disjunctive or the collaborative conception of collective omniscience does better justice to the intuitive idea of collective omniscience, since we seem to have no particularly strong (or at least no single) intuitive idea here at all. In applications of the idea of collective omniscience as a goal for the scientific community, such as Campbell [1969], the collaborative conception works better. Campbell's metaphor—the title of the paper just cited being 'Ethnocentrism of disciplines and the fish-scale model of omniscience'—is that individual scientists' knowledge are like the partly overlapping scales covering a fish's body (in the ideal situation for science), this being analogous to the body of truths; here there will of course be parts which are either too large or too scattered to be covered by any single scale.

21. A variation on the Hilpinen-inspired definition of collective omniscience has been suggested to me by Frank Jackson. Call a set of true statements from which all true statements follow a comprehensive set of truths; then the suggestion is that a and b are collectively omniscient when there is some comprehensive set of truths each of which is known either by a or by b. Then, assuming that the knowers here satisfy the monotonicity condition and also that the logical consequence relation is finitary (in the sense of the Appendix to this paper), Jackson's characterization of collective omniscience is equivalent to the collaborative notion in the text. In addition to Jackson, I should like to thank Pamela Tate for suggesting the inclusion of Notes 5 and 12.

'The Formalities of Collective Omniscience': Updates and Afterthoughts

Some technical repercussions of the derivation (1′)–(6′) in this paper receive attention in §3 of Humberstone [1990*a*]; the discussion there attends to the adverse effects of 'boolean impoverishment' (omitting from a formal language for intensional logic certain truth-functional connectives as either primitive or definable) on completeness results—or in some cases just on

standard proofs of such results—for modal logics with respect to classes of Kripke frames. In the case of the section referred to, we have a failure of completeness rather than merely a hitch in the standard proofs of completeness – a failure which may reasonably be blamed on the absence of disjunction. Other technical aspects of the paper—in particular the idea of modal operators interpreted by the intersections of the accessibility operators interpreting two given operators—have been taken up by what is sometimes called the Bulgarian school (in propositional dynamic logic); see for example Passy and Tinchev [1991].

Our discussion of the (1′)–(6′) deduction with desire in place of belief raised the question of monotonicity for the 'W', whose plausibility – we remarked – had been questioned, *inter alia*, by Frank Jackson. The kind of objection he raises there he also raises for the case of the 'it ought to be' operator of deontic logic. I have attempted to reply to the latter objection on pp. 155–6 of Humberstone [1991].

From *Mind* **101** (1992), 59–83. (The OUP on-line archived version of this paper is at http://mind.oxfordjournals.org/content/vol101/issue401/index.dtl.) Notes begin on p. 107 below.

CHAPTER 6. DIRECTION OF FIT

1. Three Quotations – by way of introduction

In her seminal presentation of the distinction between what have since come widely to be called two 'directions of fit', Anscombe described a man going shopping with a shopping list while being tailed by a private detective listing the man's purchases, and asked what distinguishes the shopping list from the detective's list. She answered the question thus:

> It is precisely this: if the list and the things that the man actually buys do not agree, and if this and this alone constitutes a *mistake*, then the mistake is not in the list but in the man's performance (…) whereas if the detective's record and what the man actually buys do not agree, then the mistake is in the record. (Anscombe [1957], p. 56)

The 'direction of fit' terminology actually antedates Anscombe's monograph, as we'll see in a moment, though not for marking quite the distinction to which she drew attention; for a nice example of its use in this capacity, we may quote from Mark Platts

> The distinction is in terms of the *direction of fit* of mental states to the world. Beliefs aim at being true, and their being true is their fitting the world; falsity is a decisive failing in a belief, and false beliefs should be discarded; beliefs should be changed to fit with the world, not vice versa. Desires aim at realization, and their realization is the world fitting with them; the fact that the indicative content of a desire is not realised in the world is not yet a failing *in the desire*, and not yet any reason to discard the desire; the world, crudely, should be changed to fit with our desires, not vice versa. (Platts [1979], p. 257)

Having himself quoted the above passage from Platts, Michael Smith adds the following further reflections:

> For the difference between beliefs and desires in terms of direction of fit comes down to a difference between the counterfactual dependence of a belief and a desire that p, on a perception that *not* p: roughly, a belief that p is a state that tends to go out of existence in the presence of a perception that *not* p, whereas a desire that p is a state that tends to endure, disposing a subject in that state to bring it about that p. Thus, we may say, attributions of beliefs and desires require

> that different *kinds* of counterfactuals are true of the subjects to whom they are attributed. We may say that this is what a difference in their direction of fit *is*. (Smith [1987], p. 54)

These quotations help us in their different ways to light upon a single and apparently significant distinction, which for the moment we may take to be understood and indeed familiar. In the following section, its significance will be illustrated by examples of its application. After that, we will return to the passages quoted above and assess their characterizations of the distinction (§3), as well as commenting on some other suggestions. In §4 we present tentatively a positive proposal, distilled out of criticisms of these suggestions, with the aid of an idea taken from J. O. Urmson. The interest throughout is on saying, clearly and non-metaphorically, what direction of fit is, and in what difference in respect of direction of fit consists.

It will be useful to have some terminology to mark the distinction. Searle [1979] applies the distinction to effect a taxonomy of speech acts. He distinguishes the two directions of fit, for this application, as the *words-to-world* direction (statements, predictions, etc.) and the *world-to-words* direction (commands, promises, etc.).[1] Since we will be interested mainly in the distinction as it applies to propositional attitudes, or mental states, rather than to (putative) expressions thereof, this terminology is too specifically linguistic. Obviously one could reach for some familiar dichotomy and press it into service – for example, theoretical *vs*. practical, or cognitive *vs*. conative – but each such contrast comes with its own potentially distracting associations, and it seems preferable to start afresh. Accordingly my proposed terminology will distinguish the *thetic* and *telic* directions of fit, as generalizations of the linguistic words-to-world and world-to-words directions, respectively.[2]

It was, incidentally, *à propos* of the linguistic arena that the phrase 'direction of fit' was first used: namely in Austin's study [1953] of speech acts involving predication. His use of a contrast in respect of direction of fit was quite different from those described as linguistic applications above, since it pertained to putatively fact-stating speech acts, all of them, in our terms, having the thetic direction of fit.[3] The use of the direction-of-fit terminology to mark the present, telic/thetic distinction, as illustrated by our second and third opening quotations, seems to have gained currency in the first ten years after Anscombe's monograph appeared, sufficiently so that by 1966 Bernard Williams was writing of "the line between discourse which (to use a now familiar formula) has to fit the world, and discourse which the world has to fit" (Williams [1966]). As already remarked, our interest will not be so much in types of discourse as in the typology of the attitudes such discourse expresses.

2. Applications

Let us consider—by way of reminder—some ways in which the thetic/telic distinction comes to our attention. The examples mentioned are drawn from these areas: the theory of motivation, the ethics of belief, the analysis of knowledge, and the doctrine of double effect.

2.1 *Theory of Motivation.*

Platts and Smith were addressing, in the discussions from which the quotations in §1 were drawn, the question of whether the mere possession of a belief could motivate its possessor to action. A negative answer to this question forms the core of Smith's 'Humean' theory of motivation, and his main argument for that answer turns on the observation that to be motivated to procure this or that outcome is to have a certain *goal*, which is precisely, in our terminology, to be in a state with the telic direction of fit. Thus the only way the mere holding of a belief could motivate its holder to action would be if that state also had a thetic direction of fit. But Smith argues that no state could have both directions of fit, in view of the difference between the characteristic features of the two directions (as outlined in the quotation in §1).

It would not be appropriate here to enter the debate between those who, like Smith, support (what he calls) the Humean view and those who, like Platts, have opposed it. Our main concern is after all to draw attention to some of the roles the thetic/telic distinction has played in recent philosophical discussion, so that the importance of clarifying the distinction will be evident. However, it is germane to this more general project to register an observation made *à propos* of the motivation debate by Philip Pettit [1987], and developed by Huw Price [1989]. This observation is that one may concede to Smith the claim that no propositional attitude can have opposite directions of fit in respect of its propositional object, without being forced to deny, for example, that desires are beliefs.[4] For the concession means only, for this case, that a desire that *p* isn't a belief that *p*; not that, for example, a desire that *p* is not a belief that *q*, for some distinct proposition *q* (as it might be: the proposition that it is, in some suitably attenuated sense, desirable that *p*).[5]

2.2 *Ethics of Belief.*

This heading is intended to recall such discussions as that of Williams [1970] of the propriety of deciding—for whatever reasons (not bearing on its truth)—to adopt a belief – an obvious place to invoke the thetic/telic distinction. The sentiment that such a procedure is manifestly improper is articulated by drawing attention to the fact that a belief adopted because, say, of its comforting effect on the believer, is a belief adopted in flagrant disrespect for the idea that, in belief, the subject's state of mind is supposed to match the way the world is in the respect of the subject matter of the belief. (§4 below, some effort will be made, however, to distance direction-of-fit considerations from those in play in much of the ethics of belief literature.)

2.3 *Analysis of Knowledge.*

One early response to Gettier's counterexamples to the Justified-True-Belief analysis of knowledge was the idea of explicitly excluding the case in which the justified believer comes by a belief that in the circumstances "just happens" to be true, the justification notwithstanding. While the recent history of epistemology abounds with analyses attempting to rule out such cases without ruling out in addition cases of the genuine article, our purposes here are served by citing the very direct response Peter Unger once suggested. Simply say, by way of analysis, that one knows that *p* iff it is not at all accidental that one is right about its being the case that *p*.[6]

Consider the case of a subject, *S*, whose beliefs about the future are monitored by a supernatural being who, taking (for whatever reason) a special interest in minimizing falsity amongst *S*'s beliefs, intervenes in the course of history so as to make these future-oriented beliefs of *S* true. Note that we do not suppose that *S* has the slightest inkling that this is what is going on. It does not seem correct to say that *S*, who believes, for example, that Islam will be the state religion of a United Europe by the year 2100, *knows* this to be the case, even though it is not at all accidental that *S*'s belief here is true. The trouble is that the nonaccidentality pertains to a matching of the world to *S*'s mental state rather than in the converse direction that befits a thetic attitude.[7]

2.4 *Doctrine of Double Effect.*

According to the proponents of this doctrine, the badness of intended consequences of actions (whether intended simpliciter, or intended as means to some further end) can make an action wrong in circumstances in which a similar badness of consequences merely foreseen and not intended would not. Opponents of the doctrine are sceptical about how this difference in respect of an agent's propositional attitudes could possibly have the repercussions claimed for the morality of that agent's conduct. The attitudes concerned differ in respect of direction of fit, since foresight is thetically directed and intention is telically directed. From this perspective, we can put one aspect of the doctrine's appeal by saying that what is especially problematic about using evil means to achieve good results as against foreseeing that bad consequences will ensue either as side-effects or as after-effects of one's conduct, is that the telic direction of fit is inherently subject to moral constraints in a way that the thetic direction is not. The badness of a situation provides no reason whatever for not getting one's beliefs to fit the world in respect of its obtaining, whereas the badness of a situation provides every reason for not fitting the world to one's desire that it obtain.

3. Attempts to Characterize the Distinction

3.1 *Smith's Characterization.*

Recall that according to Smith a belief that *p* is (roughly) "a state that tends to go out of existence in the presence of a perception that not *p*, whereas a desire that *p* is a state that tends to endure, disposing a subject in that state to bring it about that *p*".[8] There are two ways of interpreting the "perception that" locution here. I will argue that if this phrase is read (as it most naturally would be) as meaning "belief that", the characterization is vulnerable to a certain objection, to be called the *mutatis mutandis* objection, whereas if the phrase is read non-doxastically, the characterization does not meet a plausible requirement of universality.

On the first interpretation, the problem with the characterization, as an attempt to explicate the distinction between the two directions of fit, arises over its employment, in what is said about belief, of the concept of perception. Since it would be too restrictive to count the reference to a "perception that *not p*" as specifically to sensory perception, we should understand by this phrase: a coming to believe that *not p*. Or perhaps, since "perceive that" is a factive construction: a coming to know that *not p*. Whichever reading is chosen, it is clear that we are

here explicating the thetic direction of fit by reference to states with the thetic direction. Indeed, even if we took the "perceive" in "perceive that" to allude to genuine (i.e. specifically sensory) perception, the point would remain that this is a thetic propositional attitude.

It is true that, whereas the explicandum is a belief that *p*, the explication is in terms of the hypothetical adoption of a belief that not *p*, but this makes no difference to the present point. No asymmetry has been characterized by the observation that whereas a *belief* that *p* tends to disappear on the acquisition of a belief that not *p*, a *desire* that *p* does tend instead to persist on the acquisition of a *belief* that not *p*. To disclose an asymmetry, all the relevant *mutanda* must be mutated. The quoted passage itself conceals its vulnerability to this objection by the use of the phrase "perception that not *p*" which may convey the (deceptive) appearance of being neutral with respect to the distinction between directions of fit, and thus legitimately available as a fixed ('unmutated') constituent in the explication of the asymmetry.

It is important not to misconstrue the dissatisfaction with Smith's characterization here expressed. Obviously the point has nothing to do with the truth of (his form of) the claim that whereas beliefs…, desires --- ; nor, perhaps less obviously, is it a circularity objection that is being urged. The worry is not that some would-be analysis of the concept of belief fails in virtue of employing, in disguise, that very concept; for clearly no such analysis was being offered.[9] The point is rather that you cannot informatively characterize a fundamental disanalogy between the ways in which beliefs and desires relate to their objects by contrasting them in a respect itself specified by reference to one of those two ways. It's as if one were to suggest that there is the following deep asymmetry between men and women as regards sexuality: whereas a heterosexual man will not be sexually attracted to males, a heterosexual woman will be. A type of example raising similar objections will be familiar to many readers, viz., alleged asymmetries between space and time. For instance, it is claimed that an object can be in the same place at two different times but not at two different places at the same time. It is then replied that there is a tacit 'wholly' understood in the latter case, since an object can certainly be partly in one place and partly in another at a given time, and that when we look at the temporal analogue of this particular spatial 'wholly', we find the asymmetry disappears (see Garson [1971]). Again, a premature disanalogy-claim resulted from an unnoticed *mutandum* (we say a little more about this *mutatis mutandis* style of objection in §3.2).

We turn to the second interpretation of Smith's talk of perception, taking the subject's perceiving that *not p* to amount to no more than its perceptually appearing to the subject that *not p*, where this is related to any belief that may or may nor be formed as the ground of that belief.[10] This corresponds to the narrow version of the first reading. More broadly, the second reading would count as a perceiving that *not p*, any case in which it seems to the subject that *not p*, where this phrase is so understood as not to entail that a belief is formed on the basis of its so seeming. On this second reading, perceptions are merely appearances and seemings, and are not associated specifically with the thetic direction of fit in the unwanted way of actually themselves possessing that direction. Thus the *mutatis mutandis* objection lapses.

On the current interpretation, the need to talk of talk of states tending to go out of existence or to endure on a perception that not *p* arises from the feature of the present interpretation that makes the *mutatis mutandis* objection no longer applicable. That difference is that we are now

reading talk of perception non-doxastically: how things seem or appear to *S*, on the current weak understanding of those verbs, need not be how *S* takes things to be. *S* may well take appearances at face value, and come to believe on the basis of those appearances, that that is how things are. But *S* may suspend belief, perhaps out of suspicion that the current circumstances are not conducive to reliable belief-formation, or indeed retain or form a belief that things are not how they seem in the respect in question.[11] The talk of a tendency for beliefs that *p* to disappear on perceptions that not *p* is a way of registering the default status of taking appearances at face value and the special or unusual nature of the circumstances prompting the other (suspension, disbelief) responses.

Can we, however, rest content with an account of that in which direction of fit consists which speaks in this way of mere tendencies? It is not just that belief, as a type of propositional attitude, has a certain characteristic direction of fit—one that we have baptized 'thetic'—which it might be deemed to have on the basis of how its instances typically behave, the existence of a minority of instances behaving otherwise notwithstanding. It seems rather that every individual case of believing is a case of attitudinizing which itself has the thetic direction of fit. (Likewise with wanting, and the telic direction.) If this is right, we should look for an account of direction of fit which is *universal* in the sense that it addresses something shared by all cases of belief (or all cases of desire) and locates the theticity of belief (the telicity of desire) in that shared feature.

3.2 *Platts' and Anscombe's characterizations.*

The objection in §3.1 to the first interpretation of Smith's formulation, as an articulation of the difference in respect of direction of fit between beliefs and desires, was that it pointed to a difference between the way thetic attitudes (beliefs that *p*) related to other thetic attitudes ('perceptions' that not *p*) and the way telic attitudes related to those other thetic (again) attitudes. An obviously cruder way of making this mistake would be to say that the big difference between beliefs and desires is that whereas a belief that *p* is a belief that *p*, a desire that *p* is not a belief that *p*. We would like the – as I put it – unmutated second reference to beliefs to be either mutated or else replaced by something neutral with respect to the terms of the disanalogy. If in Smith's case it is the constancy of this allusion to the thetic direction which thwarts the project, one's first reaction to Platts' discussion is that here the problem is in the constancy of the telic direction. According to the passage quoted in §1, "false beliefs should be discarded; beliefs should be changed to fit with the world", whereas, "the world, crudely, should be changed to fit with our desires"[12] The constant element across the contrast is the 'should', and while this does not make a reference to a telic propositional attitude the upshot of the whole assertion is to (purport to) express such an attitude. (We can say this without commitment to a non-cognitivist—e.g., emotivist—meta-ethics.) Let us see if this particular kind of 'failure to mutate' is objectionable.

As just introduced, the envisaged objection surely cannot be correct. Consider any account of the disanalogy we are interested in, of the form: whereas telic attitudes ---, thetic attitudes…. the dashes and dots being filled purely descriptively (or non-evaluatively). It could hardly be an objection to such a proposal that both halves of the 'whereas' contrast are suitable for the expression of belief. So there cannot be such an objection when in each case a normative (or

evaluative) filling is provided, on the grounds that here both halves are suitable for the expression of a *desire*. Let us imagine a variation on Platts' account into which the concept of desire explicitly enters, and then return to his own account in terms of 'ought'.

One *wants*—so the suggestion would run—one's false beliefs to be abandoned, whereas one does not want one's unsatisfied desires to be abandoned: rather one wants them to be satisfied. Now this suggestion (not Platts', recall) may initially appear to be vulnerable to the 'failure-to-mutate' objection considered above. The fully mutated analogue of *wanting* not to have false beliefs would be *believing* that one's desires are satisfied, and there is neither any kind of general tendency for this to be so, nor any condition of rationality which demands it. It is anything but clear, however, that this attempt to re-run the *mutatis mutandis* objection can work. For there is exposed here a difference, and one that deserves to be regarded as a difference in respect of direction of fit, between *what we want from* our desires and *what we want from* our beliefs. The 'incomplete mutation' objection does not touch this formulation, the retained telic element notwithstanding, since the asymmetry is being claimed as one between the *contents* of our (higher-order) telic propositional attitudes rather than as one between the conditions for our *having* certain (lower-order) attitudes. (The positive proposal sketched in §4 below, in terms of controlling background intentions, will be of this general form.) In more detail, the contrast just drawn is as follows. The current proposal is that we distinguish an attitude *toward* believings from an attitude (having the same direction of fit) *toward* our desirings. This is to be distinguished from saying that having a belief requires meeting some condition, while for a desire some different condition is required. The *mutatis mutandis* objection sets in for an account of the latter form when this difference is itself just a by-product of the thetic/telic contrast it was supposed to illuminate.[13]

Similarly, returning to Platts' own formulation in terms of what should be changed to match what, we have a constant telic ingredient in the normative vocabulary. (Platts would not himself agree with this, since the line he is defending is that beliefs about what should be are as straightforwardly thetic in respect of direction of fit; but then we would have a constant thetic ingredient.) There is obviously a problem about all these 'shoulds', since the most straightforward way of interpreting them is as moral vocabulary. Do we really agree, under this interpretation, that, as far as any desire is concerned, e.g., Hitler's desire to reduce the Jewish population, the world *should* be changed to fit the desire? (Need even the individual whose desire is at issue hold that the world ought to be brought into line with it? Is there even a *prima facie* moral 'should'?) We would prefer a normative but non-moral interpretation.

The non-moral normativity involved here is very clearly brought out by Anscombe's use, in the passage quoted in §1, of the concept of a *mistake*. For – concentrating on the thetic direction – we can see that the concept of a mistaken belief is not just the concept of a false belief. There is another element involved here. Suppose someone said: "I know the belief I held yesterday about the combination of the safe in the office was false, but what makes you say it was a *mistake*? – You are supposing I intended only to have such beliefs as were true". There is a kind of provisional intelligibility to this remark which should persuade us that one only makes a mistake when one thwarts one's own goals. The goal here is that of having only true beliefs; it is not one which every believer (possessing the concept of belief) must share. Its endorsement is again a substantive position in the ethics of belief. Notoriously, we must be careful in the way

we make room for its non-endorsement, however. We need to avoid the incoherence, highlighted by Moore, of holding the conceded falsity of a *particular* belief of one's to be of no concern, since one never intended to have only beliefs that weren't false.[14] That is why in the combination-lock case just given, the belief now conceded to be false was taken as one held on the previous day. Even without such a temporal shift, however, we see that the incoherence only arises for some particular belief cited as possessed in spite of its falsity. One might very well remain unconcerned at the thought that *some or other* of one's present beliefs were false, and indeed the prospect of devoting oneself to rooting out such false beliefs at all costs rather than getting on with the rest of one's life seems itself to be more than faintly irrational.

These considerations, however, do not perhaps serve to rebut the charge that in holding a false belief one does make a mistake. Set aside the general aim—not, if the above is correct, obligatorily possessed—of believing only truths, and consider a believer's aims in respect of the single belief in question. Are these aims not that this state of mind should be responsive to the way the world is? Is this, at least, not required proposition-by-proposition, for beliefs We have returned to the point from the end of our discussion of Platts, that the thetic direction of fit is specifically that we want (or 'aim at') this responsiveness. But must this goal always take precedence over others? Suppose that last week I stole from the office safe and was due last night to be given a lie-detector examination. Accordingly, I underwent a few days ago a course of hypnosis to instill in me the belief that the combination was 10–24–39, as it had been in fact two years before, when I was employed in a capacity that made me legitimately privy to such information. Last night when the question came up in the test, I confidently and sincerely gave this incorrect answer to the crucial question, and removed myself from the list of suspects. And you are trying to tell me that in believing falsely on this occasion, I was making a *mistake*![15]

You will of course reply—and with great plausibility—that we need to distinguish the question of whether it was a mistake to inculcate the false belief (which, let's agree, it wasn't) from the question of whether having inculcated it, I had deliberately got myself into the position of *mistakenly believing* that the combination was 10–24–39. 1 surely had. And we can say all this without retracting the point that there are no mistakes in performance other than those of performance based on mistaken belief. I made no mistake in inculcating the erroneous belief because I acted on the perfectly correct (and of course quite distinct) belief that this procedure would save me from detection. It is, then, not possible after all to override the 'truth-tracking' aim of holding beliefs by an ulterior goal and avoid having the charge of 'mistaken' apply to false beliefs. Such a conclusion is compatible with deeming it sometimes to be reasonable to hold mistaken beliefs when morality or prudence dictates this. The point is just that what might be called the 'internal axiology' of belief continues to pass its own negative verdict ('mistake') even when its demands are reasonably overridden. As with Platts' characterization, we must conclude this treatment of Anscombe's suggestion with the observation that while it seems to be on the right track, the normative element we have just noted to be implicit in the suggestion remains somewhat mysterious.[16] An attempt to dispel the mystery—to ground this normativity—will be offered in §4.

3.3 *Naturalistic Characterizations.*

One common response to the philosophically mysterious is to reach for the biologically explanatory. In this vein, Dennett [1971] directs our attention to the fact that "the capacity to believe would have no survival value unless it were a capacity to believe truths" (p. 101); thus "In general, normally, more often than not, if *x* believes *p*, *p* is true" (p. 102). A non-statistical interpretation of 'normally' in this second remark best brings out the point of the first. It is a matter of what beliefs (or more accurately perhaps, belief-forming tendencies) are *for*. The function of beliefs is to represent the world the way it is: if they did not have this feature they would have no survival value for the creatures possessing them.[17] That is their 'proper function' in the terminology of Ruth Millikan, who has developed similar ideas at some length in her [1984]: just as hearts would not (now) exist if they didn't pump blood, which is why *pumping blood* counts as the function of the heart, so beliefs would not be here today if they didn't (under 'normal' circumstances) tend to represent the world correctly, which is why *being true* is the function of beliefs.[18] Similarly, if desires had no tendency to be satisfied—if say, a desire for food tended by contrast to reduce the chances of consuming food—then desires would simply not have evolved.[19] We should note in passing that this kind of 'evolutionary functionalism' (and its analogue for intentionally created artifacts) is not, even when applied to the illumination of psychological phenomena, at all the same as what goes under the name of functionalism in the philosophy of mind, since arbitrary black-box (input/output) descriptions of intentional states need make no special reference to this teleological dimension.[20]

So far, the evolutionary functionalist perspective may appear to throw little light on the distinction between the thetic and telic directions of fit. It does represent one spelling out of the 'aims at' terminology in descriptions (as in the quotation in §1 from Platts) of belief and desire as aiming at truth and realization, respectively; but this is by itself not to make a contrast in respect of direction of fit, since the 'aims at' (however explicated) is common to both descriptions, and talk of a desire's being realized (fulfilled, satisfied) is talk of the *same* relation of desire to the world as is involved in talk of a belief's being true: namely, the relation given by: *a* attitudinizes that *p*, and, in fact, *p* (where 'attitudinizes' holds a place for 'believes' or 'desires'.) Indeed, the writers with whom we are here concerned were not addressing the direction-of-fit question. But under magnification, the proposal does offer some promise, since the successful evolution of the belief-forming apparatus requires a causal dependence of the beliefs formed upon the states of the world about which beliefs are held, whereas in the case of the mechanism for desire, what is required is that the way the world is affected by the organism be causally dependent on the desires possessed. So here we have a explication of difference in direction of fit in terms of evolutionarily advantageous difference in direction of causal dependence.

Having elicited this causal directionality difference, one may wonder how important the evolutionary part of the story is after all. Recall its role as grounding the 'function of' talk, which was in turn suggested as the source of the normativity involved in saying what ought to fit what: we can say what beliefs ought to do in much the same way as we can say what hearts ought to do – subserve those functions their subserving which constitutes their *raison d'être*.[21] But we should ask whether evolutionary considerations belong in this discussion at all. It is conceivable that a selective advantage should occasionally, or even frequently, be conferred on those prone to form false beliefs.[22] But beliefs in the category in question would not differ from others in

respect of how they are 'supposed to' fit the world. (The desideratum of universality, again.) It is also conceivable that beliefs—and, for that matter, desires—should be possessed by a creature whose very existence is not to be explained evolutionarily (or 'artifactually'): for example, having emerged as a result of an explosion in an organic chemistry lab; or by a being (God, perhaps) who has always been in existence and always been capable of having beliefs. Such 'merely logical' possibilities will be dismissed as *recherché* by those who want to 'naturalize' epistemology and the philosophy of mind. But they cannot be ignored if what we want is an account of what the direction-of-fit contrast *consists in*, as opposed to a description of some of its contingent concomitants. Our present concern is not to explain the emergence of mental states exhibiting the telic and thetic directions of fit, but to explicate the distinction between those directions.

Even if we set aside as irrelevant the question of the evolution (or the design) of the capacities for believing and desiring, there remains, from the views canvassed here, the causal asymmetry. Beliefs get to be possessed because they are true, whereas desires get to be fulfilled because they are possessed. At least, that is how things go under ideal circumstances. Already the restriction to ideal circumstances involves a violation of the demand of universality from §3.1, which in the case of belief restricts attention to circumstances conducive to the formation of true beliefs. But the attempt to explicate direction of fit in terms of causal direction is in even more serious violation of the universality desideratum. Consider, for example, the case of beliefs of the 'self-fulfilling prophecy' type, such as the belief that one will indeed be successful in a certain venture, success in which causally requires—and will in the circumstances be ensured by—such confidence. These cases, in which "faith in a fact can help create the fact",[23] have the direction of causal influence going the wrong way-from mind to world though they involve beliefs with the same direction of fit as any other beliefs, being appraised for correctness (more on which in §4) in terms of how well their content matches how things are with their subject matter. Our conclusion must be that in noting the typical causal asymmetries between telic and thetic states, we have not got to the bottom of what distinguishes their directions of fit.

4. A Positive Suggestion

We have noted that, from a logical point of view, the relation of a belief to its propositional object, which has to obtain for the belief to be true, is the same as the relation a desire must bear to its propositional object for the desire to be satisfied: if the object (or 'content') is the proposition that *p* then, in either case, for the attitude to have the property in question is for it to be the case that *p*. Yet while we find it perfectly natural to describe a belief with a true propositional object as a true belief, there is no tendency toward a similar transference in the case of desire. A somewhat more widely applicable term here is 'correct'. Not only can beliefs be described as correct (as an alternative to 'true') when their propositional objects are true, but also expectations, answers, and various other things it would be somewhat strained to call *true*. Yet, for all its greater applicability in this respect, there is no lessening of our resistance to applying it to desires.[24] Isn't the reason for this simply that 'correct' is a term of favourable evaluation for various propositional attitudes and speech acts, and while it is a merit in a belief

(or an expectation, or an answer) to have a propositional object which is true, this is no kind of merit in desire? As Platts put it (in the passage quoted in §1) "falsity is a decisive failing in a belief" whereas "the fact that the indicative content of a desire is not realised in the world is not yet *a failing in the desire*"; correlatively, the fact that a desire is realised in the world is not a creditable feature of the desire. To summarize the intuitions in play here, we can say that although in the case of thetic and telic attitudes alike we have a sense of 'things going right', when it comes to focussing this favourable evaluation more specifically, in the former case, the evaluation comes to settle on the attitude itself, in the latter case, it settles, if anywhere, on the state of affairs in virtue of which the content of the attitude is a true proposition. Of course, this is only another way of putting the favourable side of the coin whose unfavourable side was expressed by Anscombe in terms of where, if a mistake was to be located anywhere, the mistake should be located. As intimated in §3.3, to make a mistake is to jeopardize one's success in a achieving some goal one has (at some level). Thus to trace the source of the normativity in talk of direction of fit for the thetic case, we need to see what makes believing *truly* believing *successfully*. With luck, what we shall see will suggest a suitable treatment for the telic direction.

What, then, makes truth the mark of success for beliefs? To answer this question, it will help to recall the troubles besetting traditional empiricist accounts of memory (Hume, Russell, …), as diagnosed in Urmson [1967]. A recurrent theme in such accounts is a contrast between memory and imagination, with discussion tending to ask what grounds this contrast. Urmson points out that two such contrasts are confusedly in play: that between recollections which are genuine memories and recollections which are not, on the one hand, and that between recollections (veridical or otherwise) and mere free imaginings which do not even purport to represent the past as it was, on the other. It is this latter distinction which is relevant here. As Urmson puts it, what distinguishes the cases is not the presence of any special features in such mental images as may accompany the recollecting or imagining. Rather (p. 87), "All we have to do is know what criteria of success are applicable, and that is a question which depends on our own intentions". He has been considering a case in which one's images are as of conducting the defence in a criminal trial, and the question is whether one is (correctly or otherwise) recollecting, or instead merely imagining, having conducted the defence. "We are recollecting", Urmson continues, "not if we did conduct the defence in the trial but if it matters whether we did. We are imagining if some such criteria of success as general verisimilitude, or interestingness, are the relevant ones".

Now, apparent memories, or recollections as Urmson calls them, are only a special case of beliefs. To the more general question of what distinguishes believing that something is the case from imagining it to be, the reply that it is a matter of which criteria of success one has chosen for one's mental activity seems equally applicable. If the criterion of success is *truth*, then the propositional attitude is *belief*. This is not to say that for every propositional attitude there is some such associated criterion for success, and we need not even endorse Urmson's suggestion that such criteria exist for imagining.[25] The upshot of our earlier observations on the absence of a notion of correctness for desires is that such criteria do not exist in their case. The present point is simply that unless one takes there to be a criterion of success in the case of an attitude towards the proposition that *p*, and, further, takes that criterion to be truth, then whatever else it

may be, the attitude in question is not that of belief. So unless the attitude-holder has what we might call a controlling background intention that his or her attitudinizing is successful only if its propositional content is true, then the attitude taken is not that of belief.

This way of explicating the thetic direction of fit presents beliefs having that direction as a matter of a constitutive principle rather than a regulative principle. It's not – as it might be in the case of some ethics-of-belief-inspired proposal – that one should regulate one's cognitive life by the principle: make every effort to believe only truths; rather, it's that unless one counts one's (current) intentions in φ-ing that *p* as thwarted if it is not true that *p*, one's φ-ing that *p* does not constitute believing that *p*. Thus the very concept of belief imports its own criterion of success, or, as it was put in §3.2, has its own 'internal axiology'. We have here a contrast also with the evolutionary functionalist suggestion of an externally supplied criterion of success in terms of general adaptivity, or perhaps instead, case-by-case pragmatic utility. It is of course useful—to understate the point considerably—to have a faculty which is sensitive to the way the world is, since without such a faculty there would be no chance of acting so as to satisfy one's needs in a world which is that way, but the prior concept is the concept of sensitivity deployed in making this point. (One sometimes encounters the view—it is presented sympathetically, if not finally endorsed, for example, in §§1,2 of Chapter 3 in Adams [1975]—that the reason we want our beliefs to be true lies in the instrumental value of true belief to the project of satisfying our more mundane desires.[26] This wrongly suggests it to be a perfectly straightforward conceptual possibility—which the prudent agent will refrain from exploiting—for belief-formation to occur with complete indifference to any requirement that the belief in question be possessed only if it is true.[27]

If we treat the thetic direction of fit in accordance with the above proposal, requiring for an attitude to count as thetic that it be subject to a background intention of only being possessed should the world be a certain way,[28] then the normative aspects of talk of correctness and of "what ought to fit what" are straightforwardly given by compliance with the intention. And since, on the proposed account, the intention in question plays a constitutive rather than a regulative role, it is not as though the normativity can be eluded by retaining the attitude while disowning the intention. (Hence the fact, noted in discussion of the lie-detector example in §3, that in believing that *p* when it is false that *p* one is *mistakenly* believing that *p*, whatever extraneous motives one may have had in getting into that state and however well those motives are served by being in it.[29]) This leaves the telic attitudes, and here the natural suggestion is that we tell a similar story, with, this time, a certain background intention to the effect that the telic attitude, which may be a desire or may itself be an intention, should be fulfilled. In the case where the specific attitude is an intention, say, to cross the road, a question arises about the status of the intention that that intention be fulfilled. Is this the same intention over again, or is it a different—perhaps 'higher-order'—intention? One might say that it's nothing but the same intention all over again.[30] On this supposition, the telic direction of fit requires little by way of comment: as the point was put above, the 'ought' in 'what ought to fit what' is given as a matter of compliance with this very intention. On the other hand, a more direct parallel with the thetic case is provided by taking the higher-order route, and the contrast between the two directions of fit is made visible in a formulation of what does the conditioning and what is conditioned, in certain conditional intentions, as we shall now see.

The controlling background intention in the case of belief is a conditional intention. Suppose that the propositional object of a belief is the proposition that *p*. Then this intention can be described as the intention not to be believe that *p*, given that (or: in the circumstance that) *not p*. Let us represent this by: *Intend*(¬B*p*/¬*p*). Recall the justification for postulating this intention (derived from Urmson): unless a piece of attitudinizing is thought of as controlled by such an intention, there is no reason to think of it as an instance of believing that *p*, as opposed to imagining that *p*, entertaining the proposition that *p*, supposing that *p*, desiring that *p*,… and so on through the range of non-thetic attitudes. On the 'higher-order' proposal for controlling intentions in the telic case, the intention is that it be the case that *p*, given the telic attitude toward *p*: intention, desire, or whatever. Actually, as we shall note in the final paragraph below, it may not be that intention is quite the right higher-order attitude to invoke here, but we gloss over these worries for the sake of presenting a definite proposal; a more cautious presentation of the current positive suggestion would treat our italicized '*Intend*(α/β)' as schematic for some attitude akin to intention proper. Abbreviatively, let us write W*p* in this case (for: the subject *wants* that *p*): *Intend*(*p*/W*p*).

So far, all we have is some suggestive notation, contrasting (1) with (2):

(1) *Intend*(¬B*p*/¬*p*)

(2) *Intend*(*p*/W*p*)

The notation is meant to recall that of dyadic deontic logic, of course, and in particular to discourage – as in that context – the idea that such ascriptions of conditional intention can be contraposed. We can remain neutral on certain further features of their logical behaviour, as long as we hold firm to this one. It can be explained by either of two accounts of the construction in (1) and (2), between which it is not necessary here to choose. The first, taking its lead from the conditional obligation literature (e.g. Lewis [1974]), would reach for a notion of preferability underlying claims of the form *Intend*(α/β), according to which this says that its being the case that $\alpha \wedge \beta$ is preferable (from the intender's point of view) to its being the case that $\neg\alpha \wedge \beta$, glossing this in possible worlds terms thus: some worlds in which the former is true are higher-ranked than any in which the latter is true. Give or take certain complications not to the point here, this amounts to saying that if we restrict attention to the β-verifying worlds, we find that those in which α is true are ranked higher (by the preference ordering, this time conceived as comparing worlds) than those in which α is false.[31] It is clear that contraposition – the inference from *Intend*(α/β) to *Intend*(¬β/¬α) – fails on this account, and illuminating to contrast (1) and (2) in the terms provided by the account. For (1), we look at the worlds in which the object of the belief is false, and declare our preference for those in which the belief is not possessed. For (2) we look at the worlds in which the desire is possessed, and declare our preference for those in which the object of the desire is true. So what is held fixed in the two cases is different: in the former case, facts about the object of the attitude, but in the latter, facts about its possession. This is a direct formal—or *structural*—rendering of the metaphor of difference in respect of direction of fit: the thing held fixed is that to which what is not left fixed is to be 'fitted'. The thetic/telic difference is a difference in the structure of a controlling conditional intention, a difference over what does the conditioning and what is conditioned.

This general description—cashing direction of fit in terms of the logical structure of conditional intentions—also applies in the case of the second way, alluded to above, of interpreting (1) and (2); for this second way, we are to think of the dyadic construction they feature as really monadic after all, with *Intend*(α/β) being an alternative notation for

(3) *Intend*($\beta \rightarrow S\alpha$)

in which the arrow symbolizes (let's say) material implication, and the 'S' is a subjunctivizing operator.[32] We can think of (3) as describing the attitude of a subject who would endorse the following: either it is not the case that β or else *let it be* the case that α. Such a subject, on learning that β, will (in the absence of a change of heart) come to intend *tout court* that it be the case that α ('*Intend*($S\alpha$)' in the present notation). We need not concern ourselves with the semantics of this language here,[33] pausing only to note the failure of contraposition, even when S is taken to commute with negation (as in Humberstone [1982]. The passage from *Intend*(α/β) to *Intend*($\neg\beta/\neg\alpha$), on the present approach, amounts to a passage from (4) to (5):

(4) *Intend*($\beta \rightarrow S\alpha$)

(5) *Intend*($\neg\alpha \rightarrow S\neg\beta$).

But even with the above commutation property and also the assumption that equivalents may be substituted within the scope of '*Intend*', the closest we come to (5) on the basis of (4) is

(5′) *Intend*($S\neg\alpha \rightarrow \neg\beta$)

which has the subjunctivity in the wrong place.

Alike in their resistance to contraposition, the two suggested understandings of (1) and (2) differ in respect of other inferential properties. Conspicuously, the inference-pattern sometimes called 'Strengthening the Antecedent', from *Intend*(α/β) to Intend ($\alpha/\beta \wedge$ gamma) is valid on the second interpretation but invalid on the first. Why is (non-) contraposition so important by contrast with these other features, on the basis of which we shall not even attempt to reach a comparative verdict as between the two interpretations?[34] As already noted, the failure of contraposition is crucial to seeing the structural contrast between (1) and (2) as formally registering the distinction between the thetic and telic directions of fit: what (1) [= *Intend*($\neg Bp/\neg p$)] represents as (conditionally) intended is a matter of mental state, whereas material pertaining to mental states in (2) [= *Intend*(p/Wp)] is relegated to the role of a conditioning factor. This is the reason for the disparities observed earlier: fulfilled desires, unlike true beliefs, are not thought of as correct, and an unfulfilled desire, unlike a false belief, does not constitute a mistake. The normativity, either way, attaches to the belief rather than to the desire, because the controlling intention is the intention that one's beliefs be a certain way in the thetic case (namely, as (1) says, that they not be false) but that the world be a certain way in the telic case. If such representations as (1) and (2) could be contraposed, precisely this much-needed asymmetry would be lost.

The presence of the two negations in (1) echoes (or is echoed by) the 'only' in our earlier talk of a background intention in the thetic case to believe only that which is true; one would not with equal naturalness speak of intending to desire only what is in fact the case.[35] This difference would of course also be nullified by wanton contraposition. Such formulations call for an additional comment to ward off an inappropriate 'generality' construal. The intention whose

possession is said here to be needed by one purporting to believe that *p* is the intention only to have that attitude on the condition that *p*: that is, not to be believing that *p* unless in fact *p*. A *general* intention to believe only truths is not to the point, and if the moral of our lie-detector example is accepted, this general intention may well be not possessed even rationally so – by someone who still has various particular beliefs. (Note that in that example, the present proposal still dictates that the individual who has deliberately inculcated a false belief currently intends-should-that-belief-be-false, *not* to have that belief.[36]) A similar point applies in respect of desires; see Kenny [1966*a*], p. 95*ff*.

This matter of generality is one of several which distinguishes the present efforts from forays into the ethics of belief in the tradition of William James. Norms issuing from that tradition have included not only the generalized version of (1) just mentioned—the norm of avoiding false beliefs—but also its mirror image, that of acquiring true beliefs. A corresponding particularized intention, represented by (1) with the negation signs deleted, appears to lie outside of what we have called the internal axiology of belief: that is, no such intention seems needed in order correctly to be said to believe that *p* or to believe that not *p*. Of course, *given an interest in the question of whether or not p*, attitudinizing controlled by intention (1) will indirectly be guided to settle on the true belief, since the only alternative (the no-belief-either-way option being excluded) will violate that intention. But no such interest is mandatory.[37]

Apart from the two respects just mentioned—the question of generality and the norm of gaining true beliefs—another consideration divides the present attempt to base an account of the thetic direction of fit on controlling intentions rather than on the ethics of belief: the role of evidence. The view (opposed by James) according to which it is wrong to believe other than on the preponderance of evidence might suggest that the appropriate controlling intention in the thetic case is not: only to believe what is true, but rather: only to believe what is supported by one's evidence. It appears, though, that the role of evidence in the present connection is derivative: it is hard to see how one could attitudinize with the aim of not believing something unless it was true other than by taking advantage of whatever evidence came one's way. A more direct role for evidence seems ruled out, also, by the observation that a true belief retained in spite of overwhelming contrary (but, as it turns out, misleading) evidence is not thought of as mistaken, even if the subject is criticized as irrational, or held to have made a mistake elsewhere (e.g., as to the strength of the evidence, or its bearing on the case in hand). Finally, a simple reason for not allowing the notion of evidence the role here contemplated comes from the fact that believing something on 'blind faith' is still a matter of believing something, so the universality desideratum dictates that evidence-sensitivity cannot play this role.[38]

Intending to believe only that which is true and intending that what one wants to be the case should be the case, may appear to involve one in conflict of intentions where no such conflict seems plausibly attributed – telling against the suggested account of direction of fit. One might well desire that *q* but believe that *not q*. How, in this case, the objection proceeds to ask, can one consistently have both the intentions recorded in (1) and (2)? Now if (2) had been given instead as

(2′) *Intend*(*p*/B*p*)

then from this, with ¬*q* for *p*, together with (1) there would indeed follow:

(6) $Intend(q/\mathrm{W}q) \wedge Intend\,(\neg q/\mathrm{B}\neg q)$

which gives our subject who believes that p but desires that not p contradictory intentions, each relative to some condition which actually obtains, exactly as the objector supposes. But (2) was not given as (2′), as the above discussion of the failure of contraposition for the construction in question emphasized: the whole difference between thetic and telic direction of fit comes down, on the present account, to the difference between what is intended conditionally on what. Making the above moves with (2) in place of (2′) leads to the rather different conclusion:

(7) $Intend(q/\mathrm{W}q) \wedge Intend(\neg\mathrm{B}\neg q/q)$

in which the conditional intentions are not similarly contradictory.

Having disposed of the spurious worry about the present account attributing irrationality to our envisaged subject, there remains a genuine, though not rationally criticizable, tension in that subject's state of mind. We can bring this out by calling a conditional intention that α given β *violated* in any situation in which it is true that $\beta \wedge \neg\alpha$. (These are the 'dispreferred' situations on the dyadic-deontic-logic-inspired account of conditional intention sketched above.) In the case of our subject who believes that not q but desires that q, we can argue by cases thus, to bring out the respect in which our subject is indeed 'intentionally conflicted':

> Suppose q. Then the intention recorded in the second conjunct of (7) is violated, since $\mathrm{B}\neg q$.
>
> Suppose $\neg q$. Then the intention recorded in the first conjunct of (7) is violated, since $\mathrm{W}q$.

Thus we can see in a way which is *a priori* relative to whether or not q, that the subject's conditional intentions cannot all be fulfilled (i.e., at least one must be violated). This, as I say, can be regarded as exhibiting a tension rather than an inconsistency in those attitudes: though not irrational, the subject's state is unfortunate. This result is not untoward: any subject with the attitudes in question will regret the putative fact of the unsatisfied desire that q. However, that verdict does not fully exploit the materials at our disposal, drawing only on (i) $\mathrm{B}\neg q$ and (ii) $\mathrm{W}q$. According to the line of objection we now suppose resurrected after conceding that the premature conflation of (2) and (2′) was not to the point, any such regret is due to the putatively unfulfilled desire that q, and arguably to the violation of the intention (conditional upon that desire) that q, and not at all to the violation of the theticity-characterizing intention only to believe that not q should it be the case that *not* q.[39] Suppose that the subject later discovers that in fact q, so the earlier belief was false but the desire is satisfied. According to (7), this should come as a bit of good news and a bit of bad news: the intention attributed in its first conjunct has been fulfilled, while that in its second conjunct has been violated. But in fact—the objection proceeds—isn't the news really all good news? Believing your daughter will be convicted while wanting her to be acquitted, how seriously *disappointed* will you be on hearing of the acquittal that your belief was false? Whenever you desire that q but believe that $\neg q$, don't you in fact hope that you're wrong, and won't you be undilutedly *delighted* to find that you were? Hardly the sort of response one would expect of an intention violated!

One might attempt a reply to the above objection which stressed the relative seriousness of different intentions, holding that any potential disappointment at having had a false belief is

outweighed by relief that, for example, your daughter has been acquitted. But this seems wrong – and to belong more to a puritanical strand in the ethics-of-belief tradition ("Forgive me, for I have sinned against the injunction never to believe falsely") than to the account suggested in this section. It seems wrong because, for example, the allegedly outweighed concern with believing only the truth leaves no disappointment even when there is nothing to outweigh it when for instance one is (effectively) indifferent towards the proposition one has just learnt one believed falsely. Noting that this is a situation one can only be in after a change of belief, we should consider what happens after a change of desire. There is no disappointment consequent on the non-satisfaction of a discarded desire, a fact which I take it does nothing to undermine the claim that one has a higher-order pro-attitude toward the fulfillment of any *current* desire.[40] (For various reasons, 'intention' may not be quite the right word for this attitude; one is that the subject may be alienated from some desires, for which talk of an intention that they be fulfilled sounds too committed; another is that talk of intention can carry the suggestion of intending to *make something the case*, which is not appropriate for the propositional objects of many lower-order telic attitudes.) Thus, in (1), the 'Intend' and the 'W' should be understood as temporally indexed to the same time. Likewise in (2), for 'Intend' and 'B'. But whereas the past belief has been given up, it might be claimed that one still endorses the past intention to believe only what is true, and that it therefore remains unclear why there is no affective trace of this intention's having been frustrated. One might of course be disappointed at having been 'taken in' by what one thought of as good evidence for the belief now abandoned: but any such sentiments would be echoes of an ethics of belief approach, and, in any case, the belief now abandoned might have been precisely appropriate in the light of the evidence then available. The only vestige of the frustrated theticity-characterizing intention is the concession that one was mistaken; but if what was said above is correct, there can be no mistake without a thwarted intention. (Thus, in this case, 'intention' is exactly the right word.) This concludes our discussion of the 'no disappointment' objection[41] and with it, the presentation of the case for seeing the difference in direction of fit between telic and thetic attitudes as constituted by a difference in the direction of conditionality of the controlling conditional intentions which make those attitudes the attitudes they are .[42]

NOTES

1. Compare also the direction-of-fit based explication of the subjunctive/indicative contrast in James [1986]; subjunctivity will make a brief appearance below in §4.

2. In Searle [1983] the appellations 'world-to-mind' and 'mind-to-world' are used to mark the present, non-linguistic, version of the distinction. 'Thetic' and 'telic' are preferred here partly on grounds of brevity; these adjectives may be found, incidentally, in the OED, with senses not rendering too inappropriate their proposed technical usage here (though my own preference would be for a pronunciation, in both cases, with 'e' as in 'be' rather than—the OED's recommendation—as in 'bed'). Actually, brevity aside, I have another reason for not following Searle's terminological proposals, namely a difficulty in remembering always that 'mind-to-world' abbreviates 'mind-to-fit-world', rather than alluding to the characteristic direction of

causation 'from-mind-to-world', which is how I naturally interpret the phrase, but which picks out typically (though not invariably – see §3.3) precisely the reverse direction of fit – as noted in Searle [1979], pp. 97, 122.

3. Indeed, the present distinction between directions of fit is as close to Austin's distinction in respect of what he calls 'onus of match', which distinction he contrasts with the distinction in respect of (what he calls) direction of fit. (In fact it is grasping the distinction between these distinctions which makes Austin's paper so challenging.)

4. Whether this concession should be made, the present *ad hominem* context aside, is another matter. What about that attitude defined by saying that one O's that *p* iff one both believes and desires that *p* (compare: being glad that *p*).

5. On the need for attenuation here, see Humberstone [1987]; by ignoring it, it is possible to make spurious trouble for the present suggestion – see for example note 44 of Smith [1987], which is appended to a discussion anticipating the Pettit-Price objection.

6. See Unger [1968]. By way of clarification, Unger adds (p. 159) "In my analysis of human factual knowledge, a complete absence of the accidental is claimed, not regarding the occurrence or existence of the fact known nor regarding the existence or abilities of the main who knows, but only as regards a certain relation concerning the man and the fact".

7. Of course by cashing out non-accidentality in causal terms, as in Goldman [1967], for example, attention to the direction of causation can eliminate the problem noted here. In §3.3 we will express dissatisfaction with the idea of explicating direction of fit in terms of causal direction. [[See 'Updates and Afterthoughts' for a point on Unger's discussion.]]

8. Compare E. S. Russell, as quoted in the final footnote of Braithwaite [1947]: "If the goal is not reached, action usually persists".

9. Contrast Mayo's criticism of Schiller in Mayo [1967], p. 148.

10. 1 benefit, here and in some formulations to follow, from Michael Smith's own explanations to me of what he actually had in mind in his [1987]; thus this second interpretation is the authorially correct interpretation of the passage in question.

11. In the perceptual case, these matters are elegantly treated in Peacocke [1983] with the concept of representational content; see his example (p. 6) of the *trompe l'oeil* violin painting.

12. Similar formulations appear in Searle [1983], p. 8, and Smith [1988], p. 244.

13. A more precise formulation would certainly be desirable than that given here of what makes a particular account vulnerable to the kind of *mutatis mutandis* objection under consideration; I hope the remarks offered will be sufficiently suggestive for present purposes. (There is a similar difficulty in saying precisely what makes an analysis vulnerable to circularity objections.)

14. See Sorensen [1988], Chapter 1, for a contemporary study; the above example is what Sorensen calls *commissively* – rather than *omissively* – Moore-paradoxical. (This terminology marks a distinction noted in J. N. Williams [1979].)

15. For the record: the experimental evidence on the success of hypnosis as a polygraph countermeasure is not particularly favourable. It should be noted, though, that effort is usually

directed at reducing the anxiety associated with lying rather than with changing one's beliefs so that lying is not necessary. References to the relevant literature may be found in Barland and Raskin [1973], pp. 462–5.

16. We note that a further problem with Smith's characterization, on either of the interpretations considered in §3.1, is that this normative element is not so much left mysterious as, rather, left out of account altogether. (Compare the objections in Kripke [1982], 23*ff*., 34*ff*., to dispositional accounts of rule-following.) Similarly, though for reasons of space this is not undertaken here, one can assess Anscombe's characterization with respect to the *mutatis mutandis* objection urged against one interpretation of Smith. One way of thinking of her deployment, in Anscombe [1957], of the distinction between mistakes in performance and mistakes in judgment is as protection against that objection; see, however, Houlgate [1966].

17. According to another strand of thought prominent in Dennett's discussion, but not in those parts of it brought in here, we should refer, not, *à propos* of beliefs, to the creatures possessing them, but rather to the creatures to which we ascribe them. This other strand concentrates not on the utility to believers and desirers of their beliefs and desires, but of the utility to us of making sense of agents (animate or otherwise) by ascribing to them beliefs and desires.

18. 1 say "similar ideas" because although the view about desires which follows is Millikan's, she does not make quite the present suggestion concerning beliefs. Beliefs being true is not what she would call their proper function, but rather a Normal Condition for the proper function of desires. (*Cf.* Millikan [1984], p. 99.) This point was clarified for me by James Hopkins.

19. As Millikan puts it ([1984], p. 140): "It is the focused proper function of an explicit desire to produce a state of affairs onto which it maps in accordance with certain mapping rules. (This is not to say that explicit desires are usually fulfilled. Many desires, like sperm, emerge in a world that does not permit their proper functions to be performed. But surely desires proliferated in part because they were *sometimes* fulfilled. What use to have them otherwise?)"

20. This point is emphasized by Millikan ([1984], p. 139), and it is also the main theme of Sober [1990].

21. The thetic part of the above proposal has been aired by Stalnaker ([1984], p. 18) in terms he regards as naturalistic, though without explicit mention of evolutionary processes.

22. For example, it is reported that women who, having recovered from breast cancer, believe (falsely) that they were wrongly diagnosed and never had the disease, are less likely than those who acknowledge the fact, to suffer a recurrence.

23. In the words of James [1891]. (Compare also the example in §2.3 above of the person with the 'Islamic Europe' belief.)

24. Stalnaker notices that we appraise true beliefs, but not satisfied desires, as 'correct' ([1984], p. 80) and proposes to use this as a sign of membership in a broader class—of which belief is the paradigmatic member—of what he calls *acceptance* attitudes. This is conspicuously *not* to raise the question of *why* we describe beliefs but not desires as correct. Perhaps sensing that without an answer to this question, the division into acceptance attitudes and others seems founded on nothing very solid, he goes on to propose (p. 82) a formal characterization of attitudes of acceptance, as those which are (for rational agents) consistent and held towards any

proposition which follows from other propositions toward which they are held. This suggestion doesn't work, because the conditions are satisfied by "The proposition that *p* follows from propositions *a* intends should be the case", which is not to ascribe an acceptance attitude to *a* in respect of *p*.

25. Such a non-endorsement accords well with the fact that imagination has no direction of fit in the sense of the present paper; there is nothing 'wrong' with a subject's mental state or with the world from that subject's point of view if what is imagined to be the case is not the case (or if it is the case). J. D. Velleman has developed a distinction somewhat similar to that between telic and thetic directions of fit, in Velleman [1992], between propositional contents regarded as true and those regarded as to be made true. The former locution is not intended to suggest 'believed', but, more broadly, to include regarding something imaginatively or hypothetically or… as true. Thus imagination falls on the same side as belief for Velleman's dichotomy, though it has no direction of fit in the sense of the present paper.

26. Such a view is mentioned, though not endorsed, at p. 121 of Stich [1990]. Stich's discussion in the chapter in question is of the truth of sentences in a language of thought whose interpretation is 'up for grabs', whereas we have been presuming a definite propositional content for a belief which has to be true for the belief to be true. This distinction is emphasized in §2 of Harman [1991], at p. 196 of which Harman also makes the remark "More generally, I often have the desire that I believe that P only if P". (What worries me about this is of course the 'often'.)

27. This talk of a *requirement* is of course not meant to suggest any externally imposed restriction on one's mental life: there is nothing wrong with supposing, hoping, imagining, etc. – it's just that what you're doing won't count as *believing* unless this is a requirement you aim to respect in any given case.

28. We don't want to say, specifically, "should their propositional content be true" since of course *disbelief* is also a thetic attitude.

29. Compare the distinction—often encountered in discussions of the criminal law—between the intention with which, and the motive from which, an action is performed; see, for example, Fitzgerald [1962], p. 120.

30. Cf. Loar [1981], p. 198: "Isn't desiring *s* and desiring that one's desire that *s* should be *T* virtually the same thing?" (Think of '*T*' here as a predicate of truth or of fulfillment.)

31. Reading the first 'those' as 'some of those' and the second as 'all of those'.

32. This suggestion is adapted from some work of Castañeda's, massaged into roughly the present form in Humberstone [1982]. (I have some hesitation in referring the reader to this paper, in view of the damage done to intelligibility by the many typographical errors marring the published version: considerable reconstruction is needed at some points, though the general idea will be clear enough. [[*Note*: this paper appears, typographically corrected, as Chapter 1 above]].)

33. The semantic treatment in possible worlds terms of this apparatus in Humberstone [1982] is not quite right for the present application, and would need adjusting along lines suggested by

Humberstone [1987]; we need not go into the nature of this adjustment here. [[These papers appear, respectively, as Chapters 1 and 2 in this thesis.]]

34. The existence of telic attitudes which are not, in the terms of Parfit [1984], p. 151, 'conditional on their own persistence' would appear to tell against the first interpretation. More accurately, what is relevant is the modal version of this temporal concept: the point is not so much to set aside desires (for example) which one only wants to be satisfied at later times so long as one has those desires at the later times, as to set aside desires which one only wants to be satisfied at worlds in which one has the desires. I learnt of this distinction in respect of (modal) possession-dependence from its application in unpublished work in the 1970's by André Gallois to the distinction between egoistic and altruistic desires: genuinely altruistic concern for another's welfare—so the suggestion went—is displayed by a desire that the other should prosper even if one did not want that. Justice is not done to such desires by citing (as on the second interpretation) a ranking of worlds in which what is (there) desired obtains above those in which it does not obtain.

35. Compare the difference between wanting to get along with everyone who is admitted to the club, and wanting only people one gets along with to be admitted. See also Castañeda [1975], p. 67, on 'only' in 'only if'; the contrast drawn on p. 160 of the same work between the intention expressed by "I shall press button A, unless I jump" and "Unless I press button A, I shall jump" is also relevant: below, use will be made, without explicit further comment, of *unless*-constructions.

36. According to the present suggestion, the point is not that it is morally wrong to believe falsely, but it is something, which, in respect of any given proposition, a believer must intend not to be doing, in order to qualify as believing. Beware of confusing the intention not to hold a belief which is false with the—vacuous—intention of not holding a belief which one believes to be false. (Compare the difference in respect of what counts as successful execution of the intention, between the intention to purchase a painting by Rubens and the intention to purchase a painting one believes to be by Rubens. [[In the published paper I mistakenly spelt this as "Reubens".]])

37. A point nicely emphasized in Heal [1988].

38. In this last sentence, I am indebted to David Velleman. A related worry—this one a worry for the positive account suggested here—concerns beliefs held not necessarily on faith but perhaps on evidence deemed sufficient to warrant them, yet to which the subject is so firmly committed as to lack the intention not to have them should they be false. I do not know how, short of backing off from the desideratum of universality, to reply to this objection. A similar backdown may also be the only response to a further difficulty, arising over cases in which the falsity of what is believed is – for one reason or another – inconceivable. For example, I believe that $24 + 19 = 43$; but do I have the intention not to believe this, conditional on its falsity? What sense can we make of this condition? Both of these difficulties for the present view were put to me by Rae Langton.

39. I am much indebted to J. A. Burgess for pressing this line of objection.

40. §61 of Parfit [1984] on 'mere past desires'.

41. This leaves the 'synchronic' form of the objection: how can you at one and the same time hold a belief which you hope is false, when your believing is subject to the intention not to believe the false? But this is just a matter of (say) believing one proposition and desiring that its negation is true, and the discussion surrounding (7), of the consistency of the controlling intentions, applies as above.

42. Comments from John (A.) Burgess on an earlier draft of this paper resulted in considerable improvements; I am also grateful to Rae Langton for criticisms. I am especially indebted to Michael Smith and to David Velleman for very full and extremely helpful suggestions on the material, and, in Velleman's case, for making available a copy of his [1992] before the final version of this paper was completed.

'Direction of Fit': Updates and Afterthoughts

At the end of Section 2.3, the following passage appeared:

> Consider the case of a subject, *S*, whose beliefs about the future are monitored by a supernatural being who, taking (for whatever reason) a special interest in minimizing falsity amongst *S*'s beliefs, intervenes in the course of history so as to make these future-oriented beliefs of *S* true. Note that we do not suppose that *S* has the slightest inkling that this is what is going on. It does not seem correct to say that *S*, who believes, for example, that Islam will be the state religion of a United Europe by the year 2100, *knows* this to be the case, even though it is not at all accidental that *S*'s belief here is true. The trouble is that the nonaccidentality pertains to a matching of the world to *S*'s mental state rather than in the converse direction that befits a thetic attitude.

At the time of writing, I had embarrasingly forgotten that something very like this example appears in Unger's own discussion, though the being who arranges for our subject's beliefs to be true is not supernatural – just a well-connected veterinarian who arranges for the subject's (ungrounded, dream-initiated) beliefs about which horses will win races to be true beliefs (by drugging the competition). Rather than suggesting that our reluctance to classify such beliefs as instances of knowledge because of the direction of fit issue, Unger argues that our initial intuitive reaction to the case needs to be re-thought, and on reflection, we see that when it is indeed *not at all* accidental that the beliefs are true, we are happy to see them as cases of knowledge.

A propos of note 36, Fodor [1980] asks us to consider the rules (a) and (b):

(a) Iff you it's the case that P, do such and such

(b) Iff you believe it's the case that P, do such and such.

and he notes that "the compliance conditions are quite different. In particular in the case in which P is *false but believed true*, compliance with (b) consists in doing such and such, whereas compliance with (a) consists in *not* doing it. But despite this difference in compliance conditions, there something *very* peculiar (perhaps *pragmatically* peculiar, whatever precisely that may mean), about supposing that an organism might have different ways of going about attempting to comply with (a) and (b)." Similarly, Stenius [1969], Section XXII discusses the proposal that rather than his own preferred rule of assertion, according to which one should only assert what is true, the rule should instead be formulated as requiring that one only assert what one believes to be true. Plausibly, he argues that any such proposed revision should be resisted.

Zangwill [1998] has criticised the main proposal 'Direction of Fit', as well as the proposal of Smith [1987] criticized in 3.1 thereof, in the course of defending a rival proposal of his own as to the distinction between the two directions of fit. The criticism of Smith's account runs as follows (Zangwill [1998], p. 179):

> The immediate worry is this. Even if Smith correctly describes the properties of beliefs and desires, the appeal to dispositionals does not and cannot capture the essence of the distinction. For we want to say that it is *because* my state is a desire for p that I am disposed to bring it about that p. (…) Intuitively it is the difference in the nature of beliefs and desires which *explains* their typical causal relations to other mental states and to inputs and outputs. The essence of a mental state explains its causal powers; it is not constituted by them. Smith gets the direction of explanation the wrong way round.

It is not clear to me what Zangwill's precise complaint against Smith is in this passage. Why should Smith disagree with the claim that – clearing up Zangwill's formulation in the interests of literacy – it is *because* my state is a desire that p, that I am disposed to bring it about that p? Nor is the distinction clear between the essence of a mental state's explaining its causal powers, on the one hand, and its being constituted by them, on the other: according to accounts like Smith's, the concept of belief is the concept of a mental state having a certain characteristic causal role (on an understanding of this phrase generous enough to subsume patterns of persistence and disappearance in the light of various courses of experience), while the concept of desire is the concept of a mental state having a different characteristic causal role.

We can perhaps see something of what Zangwill objects to in Smith's account by comparing it with a modification of Smith's proposal (as presented in the third of the quotations in our opening section), a modification which is consonant with Zangwill's preferred "normative functionalist" account:

> The belief that p would be *pro tanto* rational if it were caused to go out of existence in the presence of the perceptual experience as of not-p. The desire that p would *pro tanto* be rational if it endured given the perceptual experience as of not-p; and if the desire that p caused the attempt to bring about that p then the attempt would be *pro tanto* rational. (Zangwill [1998], p. 196.)

It is the presence of the word "rational" in this proposal that signals the normative element in Zangwill's normative functionalism, an element whose absence he sees as the fatal flaw in the

remarks about direction of fit in both Smith [1987] and Humberstone [1992] (i.e., the main body of this chapter). After quoting the latter's suggestion (from Section 4) that "unless the attitude-holder has what we might call a controlling background intention that his or her attitudinizing is successful only if its propositional content is true, then the attitude taken is not that of belief" Zangwill ([1998], p. 180) writes:

> Someone will probably criticize the details of Humberstone's account. There are obvious difficulties. He appeals to a notion of truth which needs to be distinguished from satisfaction. He does not make it plausible that we do always have such second-order attitudes. He gives no account of what gives second-order intentions their direction of fit. And he does nothing to reassure us against the threat of a regress of higher-order attitudes. But even if we put aside these and other worries, Humberstone's appeal to second-order background intentions has the same immediate problem as Smith's account. It may be true that we tend to have different controlling intentions towards beliefs and desires in something like the way that Humberstone describes. But there is a strong intuition that this is explained by the different natures of those states. The fundamental difference between beliefs and desires – whatever it is – explains why we have difference second-order intentions towards them. The direction of explanation flows the other way. Like Smith, Humberstone describes the epiphenomenon, not the phenomenon itself.

Zangwill then proceeds to say that an account extracted from Hume suffers from this same defect as he finds in Smith and in Humberstone (so we are at least in fairly respectable company).

> The normative functionalist idea is that beliefs and desires have certain distinctive normative liaisons. (…) The asymmetry we are considering, then, is that the desire that p, by itself, rationalizes the intention to bring it about that p, but the belief that p does not. And the perceptual experience that p, by itself, rationalizes the belief that p, but not the desire that p. This rationalization is only *pro tanto*, and it can be outweighed by other rational considerations. (Zangwill [1998], p. 195)

On this view, then, the fact that the dispositions to which Smith's account alludes and the controlling intentions of my own account, are characteristic of the psychology of rational subjects is to be explained by reference to their rationality and Zangwill's preferred account of rational links (normative liaisons). I am not sure what to say about this account of Zangwill's. Smith's version of the Humean account of motivation would certainly make him unhappy with what Zangwill says, since a desire by itself can never rationalize (make rational) an intention to act in a certain way without the aid of a belief. But a few words can still be said in connection with Zangwill's earlier objections – the above listed "obvious difficulties" – to the proposal of 'Direction of Fit'.

The first difficulty was, apparently, that I "appeal to a notion of truth which needs to be distinguished from satisfaction." Not so at all. Section 4 began with these words:

> We have noted that, from a logical point of view, the relation of a belief to its propositional object, which has to obtain for the belief to be true, is the same as the relation a desire must bear to its propositional object for the desire to be satisfied: if the object (or 'content') is the proposition that *p* then, in either case, for the attitude to have the property in question is for it to be the case that *p*.

The intention here was precisely to deny that any weight is to be placed on the distinction between truth and satisfiedness, since the same logical relation between the attitude ascription and the propositional object of the attitude obtains in the two cases. In one version – the conditional intention version – of the way the distinction between telic and thetic attitudes was drawn in the paper we have: attitude A has the telic direction of fit if the controlling background intention is *Intend*(*p*/A*p*) (i.e., one intends that *p* given that one As that *p*) and the thetic direction of fit if the intention is instead *Intends*(¬A*p*/¬*p*) (i.e., one intends that one not A that *p* given that not *p*). There is no use made here of any distinction between being true and being satisfied. So we are free to go on to observe that in colloquial English we tend to calling an instance of A-ing ("attitudinizing") that p *true* or *satisfied* when it is the case that p according as A is a thetic or a telic attitude.

The second difficulty was that "he does not make it plausible that we do always have such second-order attitudes". In the case of the thetic attitudes like belief, I took it that Urmson [1967] had already made this plausible, once we extend his point about recollecting to cover the thetic attitudes generally. Perhaps more should have been said on this score in the case of the telic attitudes like desire. Thirdly: "(h)e gives no account of what gives second-order intentions their direction of fit." Well, there is nothing special about second-order attitudes: one just applies the general schema What may be more problematic is the specific second-order attitude of conditional intention when this is represented by a primitive dyadic operator. (Recall that this was one of two options canvassed in our final section.) It is true that the account of the distinction cannot be applied in this case, since we are not dealing with an attitude ascription of the form A*p*. Certainly something should be said about this, though it is not clear exactly what. Finally, "he does nothing to reassure us against the threat of a regress of higher-order attitudes". This point, which is related to the previous one, also seems reasonable. Assuming the problem of extending the telic/thetic distinction to the dyadic case is solved, there will remain infinitely many higher order intentions. This is not *obviously* a problem, however – threatening the simultaneous thinking of infinitely many thoughts, for instance – because we should not think of intending as consciously and occurrently engaging in mental activity. Still Zangwill is right to complain out that the point was not even addressed in 'Direction of Fit'.

After all these minor complaints, Zangwill raises the main objection which as we have seen he holds up against Michael Smith's account too. The having of the right controlling intentions for belief and desire should, he thinks, be explained by something else, and not taken as the essence of the belief/desire distinction. In fact, on his preferred explanation, in terms of what makes what rational, these controlling intentions may be absent in an irrational individual. This is incompatible with the story in 'Direction of Fit', so we have a more fundamental difference than any about the relatively subtle and indeed confusing business about direction of explanation in connection with direction of fit. According to that story, no-one, however irrational, can believe that p unless that individual has, with respect to his or her attitude to the proposition that p, an intention not to

have that attitude given the falsity of the proposition. The point here is again Urmson's, though put in terms of intention rather than what we take as the criterion of success (and applied to belief rather than recollecting): without such an intention, there is nothing to make the attitude one of belief as opposed to, e.g., merely entertaining the thought that p, or imagining that p, or desiring that p… There is nothing in this suggestion – which, to be sure, Zangwill complained ("the second difficulty", above) had not been made fully plausible – which would exempt irrational believers. To the extent that we deem them irrational it may well have to be because we first identify some of their beliefs as beliefs (rather than imaginings, hopes, etc.) and then worry about how they came to have such beliefs or what they come to do on the basis of them. But implicit in such an identification, according to the Urmson line, is precisely what they themselves count as the criterion of success; or, as this is transmuted in 'Direction of Fit', what it is that they themselves would see as contra-intended (namely, falsity).

From *Philosophy and Phenomenological Research* **57** (1997), 249–280. Notes for this chapter begin on p. 140.

CHAPTER 7. TWO TYPES OF CIRCULARITY

1. Introducing the Distinction

An account purporting to give conditions necessary and sufficient for the application of some concept may be vulnerable to a charge of circularity on two different sorts of grounds. The distinction between the two types of circularity charge corresponds, as we shall see, to a distinction between two ways of construing the account on offer. Although it has some potentially distracting complications, the 'psychological continuity' account (or family of accounts) of personal identity provides a convenient way of introducing the distinction. We recall Butler's objection to accounts along these lines (such as Locke's) which stress the role of *memory*: "And one should really think it self-evident that consciousness of personal identity presupposes, and therefore cannot constitute, personal identity; any more than knowledge, in any other case, can constitute truth, which it presupposes."[1] It was in order to circumvent any such charge of circularity that Shoemaker introduced the concept of *quasi*-memory, which does not require that the person who quasi-remembers φ-ing at some earlier time should actually have been the same person who did φ at that time, as long as *someone* then φ-ed as a result of which (*via* whatever type of causal route in characteristic of the laying down of experiential memories) our current subject is having the apparent memory as of having φ-ed.[2]

We can now introduce, in schematic terms, the distinction between types of circularity charge, and see which of the two types this manoevre is designed to respond to. The general form of an account of the application of a concept *K* we are concerned with says that the concept *K* applies if and only if certain conditions $C_1,\ldots,C_n$ obtain. We further take the claim that an account of the application conditions of *K* is thereby provided to involve at least the claim that this biconditional holds *a priori* (and one may or may not wish also to require that it is free from at least one of the two kinds of circularity to be distinguished here.) Now, in the first place, we may be thinking of such an account as a putative *analysis* of the concept *K*, in which case the analysis is circular if that concept is (overtly or covertly) employed in specifying the conditions $C_1,\ldots,C_n$. We will call this *analytical* circularity. Alternatively, we may just be thinking of the conditions as together providing us with an *a priori* route to the conclusion that the concept *K* applies. Since the formulation involved not just 'if' but also 'only if', we are also being told that unless the cited conditions are satisfied, the concept does not apply: we still do not need to think of such a bidirectional claim as purporting to provide an *analysis* of the concept *K*. For the moment, however, our interest is in the 'if' direction of the claim. The kind of circularity pertinent here will not be like the circularity of a putative analysis or definition, but rather, like the circularity of an argument: in particular, an argument whose premisses are the statements that (in

some given case) condition C_1 obtains, that C_2 obtains,…, that condition C_n obtains, and whose conclusion is that (in the case in question) the concept *K* applies. Let us distinguish this from analytical circularity by calling it *inferential* circularity. That is, an account of the application of some concept is inferentially circular when any argument or inference from premises claiming the various conditions provided by the account to obtain, to the conclusion that the concept applies, is itself circular in whatever sense an inference or argument can be circular. It would be nice one day to have general agreement on precisely what that sense is, but in the meantime we simply help ourselves to this unexplicated notion of the circularity of an argument.[3] There is of course a similar difficulty in saying exactly what makes an analysis circular, skirted over in our formulation above by the parenthetical 'overt or covert'. We can, in this case too, restrict our attention, when convenient, to the 'if' direction of the account, which is now naturally interpreted as a 'partial analysis' of the concept *K*: it says that the obtaining of the cited conditions is at least partially constitutive of that concept's applying. The mereological metaphor of constitution is of apiece with the idea that analytical circularity is a fatal flaw even in this partial and unidirectional form of would-be analysis: a whole cannot be composed of parts at least one of which has the original whole as a part. But, rather than *via* any such speculative involvement in the mereology of concepts, the more usual way of explaining why circularity is a flaw in an a putative analysis adverts to the role analysis is supposed to play for thinkers: if a concept is being explained, the explanation should not be one intelligible only to those already possessing the concept. This is just a formulation at the level of concepts of the usual grounds for objecting to circularity in the definition of *terms*: we are supposed to be explaining how the term defined is to be used in terms already familiar to one for whom the explanation is necessary. Analytical circularity is a fault, then, when and because it obstructs the transfer of understanding an account of the application conditions of a concept may be designed to effect: from understanding of the terms in which the account is couched to understanding the concept being analysed. Inferential circularity, on the other hand, is a fault to the extent that what is obstructed is the transfer, not of understanding, but of knowledge. Here we envisage using the account of the concept's application conditions not so much as a way of getting the concept across to someone not familiar with it, but as a recipe for telling us when it applies: if we had already to know about this before we could employ the account to that purpose, the recipe would not yield the desired knowledge. To avoid confusion, we stress that when we speak of inferential circularity we are not referring to the circularity of an inference, but to the kind of circularity in the account of a concept's application-conditions which is related to circularity of inference in the way described above when the account is employed to make an inference. (A more generous description may in the end do better justice to the phenomena, as we shall note in Section 5. This rough and ready characterization will do in the meantime.)

Let us return to the case of Butler's circularity objection to an account of personal identity which invoked *inter alia* a condition involving memory. We recall again Butler's own words, given as a gloss on his claim that memory (= 'consciousness of personal identity') presupposes personal identity, namely the addendum "any more than knowledge, in any other case, can constitute truth, which it presupposes." It is clear from this that the kind of circularity in which he is mainly interested is what we have called inferential circularity: what is objectionable is that in order to apply any criterion to decide a question of personal identity, one would have already to know whether the present would-be rememberer and the past experiencer of what is

remembered were the same person.[4] (However, we also notice the presence of the word 'constitute', suggestive of the defect we have baptized as analytical circularity.) Similarly, to react to this circularity charge by passing from a formulation in terms of memory to one in terms of quasi-memory is to take the charge as one of inferential rather than of analytical circularity. For if an objection is mounted to the deployment of the concept of memory in the analysis of personal identity on the grounds that the latter concept is itself implicit in (experiential) memory attributions – remembering having φ-ed amounting to remembering oneself as having φ-ed – then such deployment is not avoided by passage to quasi-memory since we have a similar (and similarly elucidatory) paraphrase in this case too, with 'remembering' replaced by 'quasi-remembering'. It is here that the potentially distracting feature, alluded to above, of the present example surfaces: for it is a matter of some contention whether the 'oneself' (and similar constructions) in such paraphrases does indeed count as a deployment of the concept of identity.[5] Rather than getting entangled in that debate, let us change the example to something less antecedently familiar which raises the same issue of inferential *vs.* analytical circularity without also raising this side issue about what constitutes a manifestation of the concept of identity.[6]

For our new simplified illustration, consider the concept, not of personal identity, but of personhood *simpliciter*, and the following proposed account of its application conditions: for any individual x, x is a person if and only if x knows that x is a person.[7] As before, we are not interested in the plausibility of the account — only in the details of its vulnerability to circularity objections. Butler's 'knowledge presupposes truth' objection arises in the same form as before, and a response along somewhat similar lines to that of the quasi-memory theorist might replace the 'knows' in the proposed account by 'believes'. This may remove the circularity if the circularity in question is (as with Butler) inferential circularity—though there is a question as to whether we *could* establish that a belief if possessed without first establishing that the putative believer is a person—but it certainly does not remove the evident analytical circularity occasioned by the use of the concept of a person in the account of the application conditions for that concept: all that has changed is that the concept now appears in a belief-ascription rather than a knowledge-ascription.

Suppose that we consider a biconditional of the type we are interested in but without reference to whether it is intended to serve as an analysis or not. We can still ask about whether it exhibits what we have called analytical circularity, though from this neutral standpoint we cannot say whether this constitutes a fault. Likewise with inferential circularity. Neutralized from their suggestion of failure at some appointed task, we can compare the two types of circularity. We find that any case of inferential circularity is a case of analytical circularity (from which it does not follow—*cf.* the above examples—that any *criticism* on grounds of inferential circularity is a criticism on grounds of analytical circularity), while the converse does not hold. Roughly, as we see from the cases already reviewed, the analytical circularity may fall short of inferential circularity if the concept concerned appears in a suitably protective embedding. In Section 3, we look more closely into precisely what suitability might amount to here; for the moment we notice only that embedding under 'knows' (to put the point linguistically) is not suitably protective—because of the logical relations between knowledge and truth—while embedding under 'believes' is; Section 4 brings these distinctions to bear on some meta-ethical discussions, and

Section 5 returns to the relation between unsuitably protected occurrences and inferential circularity. Before that, Section 2 considers a few questions raised by the way the distinction between inferential and analytical circularity has been introduced so far, and illustrates the recent interest in analytically circular accounts with some quotations which will lead us on naturally to the matter of protective contexts. Appendix A serves as a historical postscript to Section 2, and Appendix B fills out some of the logical details of the picture sketched in Section 3.

We close the present introduction, however, with a word about why analytical circularity might not always be a bad thing; or, to put it another way, why there is room for interest in an account of the application-conditions of some concept which does not purport to be an analysis of that concept. The considerations above concerning the transfer of understanding, on the one hand, and of knowledge, on the other, can give us some appreciation of the potential utility of accounts which are analytically circular, by seeing them in the light of some remarks made by Dummett on the circularity of an attempt at justification.[8] The latter is thought of here as an argument to the conclusion that some practice—in the case of interest to Dummett, the practice of deductive inference—is justified. So the circularity in question is circularity of argument. There is an obvious apparent threat of circularity in the case of such an argument to the conclusion that the practice of deductive inference is justified, since giving such an argument will be participating in the very practice whose justification is at issue. Dummett's response is to distinguish between a *suasive* argument and an *explanatory* argument. An argument of the former type aims at persuading someone of the truth of its conclusion (or more generally, of *establishing* its conclusion), an aim which would (or should) be frustrated by the circularity of the argument, while those of the later type aim at explaining why the conclusion – which we may suppose not to be in doubt for the intended audience – is true, perhaps in spite of some considerations which seem to preclude its truth. Dummett writes (p. 296):

> Characteristically, in an explanation, the conclusion of the argument is given in advance; and it may well be that our only reason for believing the premisses of the explanatory argument is that they provide the most plausible explanation for the truth of the conclusion. Hence the charge of circularity or of begging the question is not applicable to an explanatory argument in the way that it is to a suasive argument. A philosopher who asks for a justification of the process of deductive reasoning is not seeking to be persuaded of its justifiability, but to be given an explanation of it.

Now, a suasively intended argument is what was described earlier as one intended for the transfer of knowledge, and as Dummett notes, this is not the mandatory epistemic direction for an argument, since an explanatory argument may be useful when the conclusion is not in doubt, and may indeed (though it need not) operate with the reverse epistemic direction: it any rate may place the conclusion and the premises in an illuminating perspective by showing how they cohere. An explanatory argument whose conclusion is not in doubt and which, if suasively intended would be castigated as circular, is analogous to an account of the application conditions of some concept which utilises that very concept, and accordingly may play a similar cohesion-inducing role for those who are familiar with that concept. In the one case the conclusion is already

accepted, and there is no question of establishing it by means of the argument in question; in the other the concept is already possessed, and there is no question of introducing it by means of the account in question. If it can be useful to present arguments other than for transferring knowledge from premises to conclusion, it can be similarly useful to present accounts of concepts other than for the purpose of transferring understanding from the *analysans* to the *analysandum*. These last two are not, to be sure, optimal terms, since we reserve the title of 'analysis' for the non-circular case: more on these terminological matters in the following section, where we shall also see examples of several writers making rather more specific claims on behalf of circular accounts.[9] To clear up a possible misapprehension before closing, it should be emphasized that the analogy has been between circular arguments and analytically circular accounts: we have not said anything, the first term of this analogy notwithstanding, about inferentially circular accounts. Indeed we shall see reasons—if the above illustrations (concerning personal identity and personhood) have not already provided such reasons—for being worried about inferential circularity in an account of the application conditions of a concept, reasons which would not tell against analytical circularity.

2. The Flight from 'Reductive' Analysis in Recent Philosophical Literature

Our presentation of the distinction between analytical and inferential circularity in the preceding section took the former to be a fault specifically in what purported to be conceptual analyses or, in a more linguistic vein, definitions. Not everything which gets called a definition, however, is a definition of the kind for which analytical circularity is a defect. Consider *legal* definitions, for example. There is nothing objectionably circular about a statute which specifies that amongst the conditions sufficient for an action to count as rape, it should be perceived as rape by the victim, or about a stipulation that for the purposes of such-and-such a legal instrument an aboriginal is deemed to be anyone accepted by any of several specified communities as an aboriginal. As Andrew Markus puts it, the inset quotation here being from the *1973 Yearbook of the Commonwealth of Australia*, "Since 1973 the Commonwealth has employed a definition for administrative purposes in which self-definition forms a constituent element. Eligibility is limited to:

> A person of 'Aboriginal' or 'Torres Strait Islander' descent who identifies as an Aboriginal Islander and is accepted as such by the community with which he is associated."[10]

Legal definitions like these do not purport to analyse pre-existing concepts so much as to refine them into something more suitable (*e.g.* because more precise) for legal purposes. Indeed it may not be accurate to think of concepts as being at issue at all here. The sense in which a class of people, for example, is 'defined' in such cases seems to have more in common with the sense in which a perimeter fence 'defines' (= *delimits*) an area of land, than that in which the concept some word expresses is given by a definition of that word. The topic of definition in the law is not something with whose difficulties we need grapple here.[11] It is worth noting that the present feature is shared by all such special purpose stipulations and 'deemings', legal or otherwise. We meet it, for example, in Mary Lassiter's discussion of 'unusual names',[12] with the remark (p. 45)

that "In my own survey, I defined names as unusual only if the bearers deemed them so." (The site of the circularity here is of course the pro-form 'so'. The author is explaining the basis on which she decided to count a name as unusual, rather than saying what 'unusual' strictly meant in this context, so there is no question of a circular definition in the *semantic* sense of 'definition'.)

Before getting to the phenomenon alluded to in the title of this section, we should mention another (and even more recent) development in the formal theory of definition: the tendency for theorists to embrace circular definitions. Since we shall not make further reference to these proposals (except in passing), it will be sufficient here just to mention two candidate approaches: that of Gupta and Belnap, on the one hand, and that of Yablo on the other.[13] On the former view a circular definition embodies "not a rule of application but a rule of revision"[14], so that we apparently have here something different from the enterprise introduced in our opening section, according to which a circular account does indeed purport to state the application conditions of the concept concerned. (Gupta and Belnap see the role of a circular definition as telling us how to revise a provisional hypothesis about the extension of the *definiendum*: we interpret the *definiendum* in its occurrence(s) on the right as having this extension to compute, with the aid of the rest of the *definiens*, the revised extension, whereupon the process of revision begins again.) Yablo's conception of definitions—circular or otherwise—is, by contrast, more in keeping with the idea of a specification of the application conditions a concept, since it at least requires the (material) biconditional connecting the *definiendum* (in Yablo's notation: $\mathrm{P}x$) and *definiens* (Yablo: $\varphi(x)$) to be true. He writes (p. 150*f.*), concerning the issue of how a definition determines the extension of its *definiendum*:

> By far the most usual theory has P *inheriting* its meaning from φ. Seen from this perspective, circular definitions are indeed objectionable. Before it can bestow its meaning on P, φ must already *have* a meaning: and how can it if it contains P as apart? That said, there may be other ways for a definition to fix P's meaning than by furnishing a chunk of language to which that meaning already attaches.

Yablo's paper may be consulted for a survey of the possibilities heralded by the final sentence of this quotation, and all we need to notice here is that the talk of fixing P's meaning amounts on those various possibilities to specifying that meaning *uniquely*. While this seems to justify the use of the term 'definition', the circularity notwithstanding, the provision of analytically circular accounts of the application conditions of a concept does not in general carry with it any such uniqueness commitment, as will be apparent from the (far weaker) notions of non-triviality the proponents of such accounts have had in mind, mentioned below. Thus again we appear, with theories of circular definition, to be in somewhat different territory from that under consideration in our discussion.

Although we have used the term 'analysis' for something in which what we have called analytical circularity would be a fault, and reserved the term 'account' for the more general notion, another way of marking the same distinction is to use 'analysis' as the wider term, and mark out the narrower notion by some explicit modifier. The commonest such modification, at least in the 1980's — a period which saw a rather widespread reaction against what was perhaps perceived as excessive 'cyclophobia' — involved talk of *reductive* analysis, often with the present understanding of 'analysis' intended, with the term 'reductive' serving to emphasize

the feature we are focussing on (non-circularity). This is Mark Johnston's practice in Johnston [1989]:

> Notice that nothing in what I have said by way of characterizing response-dependent concepts implies that such concepts admit of a reductive definition or *analysis* in terms of subjects' responses. (…) That is, it may be that sometimes the biconditional of the relevant form which shows the concept to be response-dependent is strictly speaking circular. Circularity would be a vice if our aim were reductive definition. However, our aim is not reductive definition but the exhibition of conceptual connections. In such an endeavour, circularity is a defect only if it implies the triviality of the biconditional.

A well-known passage from McGinn [1983] [16] had voiced a similar sentiment, though in rather different terms, when what he begins by calling the dispositional thesis turns into the 'dispositional analysis'[17]:

> It is sometimes supposed that the dispositional thesis about (say) colour involves a circularity the exposure of which undermines any significant distinction between primary and secondary qualities. The circularity is supposed to be this: to specify which kinds of experience a red object is disposed to produce we need to *use* the word 'red'—but then the dispositional analysis of being red in terms of how things visually seem is circular. (…) Note, first, that the alleged circularity is of a peculiar kind, since it is not *generally* true that an object's looking Q entails that it is Q; this means that a claim of logical equivalence between 'is Q' and 'seems Q' will not be trivial, since not all values of 'Q' will sustain such an equivalence.

McGinn is oversimplifying the relevant proposal, of course, in suggesting that looking red entails being red, since one needs qualifications about normal perceivers and standard conditions. In discussing particularly the latter notion, John Burgess put the worry about circularity worries like this:[18]

> What I hope to have eliminated is the sense of viciousness, not the circularity. Anybody who thinks that the notions of seeming and being are indeed connected in something like the way I have outlined, but who still feels that my failure to eliminate the circularity renders the exercise philosophically futile, is, I believe, committed to an unjustifiably restrictive conception of the role of conceptual clarification. If the connections are there, it is our duty to make them. Any reductive conceptual breakdown we may achieve in the process is a welcome, but inessential, bonus.

Both Johnston and McGinn, then, reply to the charge of circularity in the accounts they are considering by stressing the non-triviality of those accounts, with McGinn further spelling out the relevant notion of triviality in terms of the correctness of analogous biconditionals for *all* concepts.[19] One can read a similar concern into Burgess's remark that if the connections are there, it is our duty to make them: for if the corresponding connections obtained automatically in all cases, then we should have no special duty to allude to them in the given case — and it would, indeed, be misleading to do so.

Circularity can threaten an analysis along dispositional lines—where the *analysandum* is held to apply just in case normal perceivers in standard conditions are disposed to respond or react in such and such a way—in three possible places: in the spelling out of normality for perceivers, in the spelling out of standardness for conditions, and in specifying the nature of the response/reaction (filling out the 'such and such way'), and it is specifically this third locus of circularity that will concern us. Circularity in this third position makes for what, adapting Richard Holton's taxonomy, we may call an *echo* account.[20] The first two, which, as just remarked were those Burgess had particularly in mind, have been extensively discussed by Crispin Wright, who worries about trivializing dispositional accounts by spelling out the notions concerned in what he calls a 'whatever-it-takes' way. (See the references given in notes 22 and 23, as well as those given in Holton [1991] — where some difficulties for Wright's programme are aired.) It is partly in order to steer clear of these problems that we will be concentrating on the specification of the response. Also, as we shall see from a some of Wright's remarks, it is the question of circularity in this specification on which the distinction between analytical and inferential circularity promises most directly to bear. Before passing to those remarks, some background is called for.

What Wright has in mind as a 'basic equation' is an account of a response-dependent concept *K* to the effect that K applies just in case if certain conditions C were to obtain (where these 'C-conditions'—no relation to the conditions schematically represented as $C_1,\ldots,C_n$ in Section 1 above—pertain to the circumstances and responsiveness of a subject S) then S would judge that *K* applied. Neither he nor Mark Johnston is satisfied with this 'counterfactual on the right-hand side' treatment of the particular kind of response-dependent concepts at issue here (called 'judgment-dispositional' concepts by Johnston, and 'extension-determining' concepts by Wright), for reasons we need not go into here. Wright favours instead what he calls *provisional* (or 'provisoed') biconditionals to the effect that *if* the C-conditions obtain, *then* it would be the case that (*K* applied if and only if S would judge that *K* applied).[21] For the quotations below, think of 'P' as standing for 'concept *K* applies'; the remarks quoted can be read with either the basic equations (for simplicity) or the provisional biconditionals in mind. Here, then, is what Wright has to say:

> So far, we have merely required that the formulation be "substantial", and have suggested no very definite account of this beyond its exclusion of offering of the "whatever-it-takes" line of goods. But what kind of formulations should count as substantial? Well, because we are not in the business of trying to provide reductive *analyses* of the truth conditions of the judgments in the substitution class of "P", there may seem no immediate barrier to characterizing the C-conditions, or the relevant response, in ways involving the very concepts which distinctively feature in those judgments. At least there is no barrier erected by the threat of circularity, as that philosophical vice is usually conceived. If the project is not to analyse the concepts in question but to exhibit the implication, in the truth conditions of judgments involving those concepts, of facts about idealised human response, circularity of that kind need surely be no objection to what we produce. However, it does not follow from the fact that there is no objection from circularity that there is no objection at all—that we run no risks of spoiling things if we make unfettered use of the concepts in question in the specification of the C-

> conditions and the appropriate response. On the present proposal, the Euthyphronistic thesis about a class of judgements is to the effect that the extension of the truth predicate among at least a proper subclass of them is determined as a function of best opinion. Suppose this thesis is advanced for judgements of colour. Then using colour concepts in the specification of the C-conditions—of what makes an opinion "best"—raises a (somewhat subtle) worry not about circularity but about making implicit demands on the extension of colour concepts which are inconsistent with the very Euthyphronism which we would be trying sharply to formulate. If what it needs to make the relevant opinions "best" under certain circumstances is represented as depending on facts about the actual extension of colour concepts in those same circumstances, how do we answer a critic who questions whether, in that those facts are fully determinate, we are implicitly presupposing some mode of constitution of colour facts which is conceptually unconstrained by best opinion and hence potentially at odds with the Euthyphronist's central claim?[22]

Wright sums up this not wholly transparent passage with the following words ([1992], p. 122); the reference to 'Notes on Basic Equations' in the sentence following the summary is to an earlier privately circulated draft:

> To stress: a Euthyphronist who is content to implicate the distinctive concepts of a discourse in his formulation of the C-conditions, in such a way that the satisfaction of the C-conditions turns on details of the actual extension of those concepts, has to expect the following challenge: "Show that the way that you have implicated those concepts is consistent with your overall thesis, that their extension is, at least partially, constrained by best opinion."
>
> The "Notes on Basic Equations, etc." offered a somewhat conservative response to this difficult matter. Evidently there can be no well-founded worry along the lines described, provided an *Independence condition* is satisfied: the relevant concepts are to be involved in the formulation of the C-conditions only in ways which allow the satisfaction of those conditions to be logically independent of the details of the extensions of those concepts. There are, roughly, two ways in which that might be accomplished. One is, obviously, to avoid the use of the targeted concepts altogether. The other is to allow the concepts to feature within the specification of the C-conditions only *inside the scope of intensional operators*—as, of course already occurs in the description of the relevant response, since "S judges that P" places the vocabulary figuring in P in *oratio obliqua*. Provided one or the other course is followed, the Euthyphronist should be safe from the sort of concern outlined.
>
> The motivation for the condition is thus that it *suffices* to preempt a certain kind of complaint. That is not to say that respecting it is necessary if the complaint is not to arise.

Now Wright's description of the issue as 'this difficult matter', and his talk of a 'somewhat subtle' worry ('not about circularity but…') suggests perhaps that he is not quite satisfied with his

own formulations of precisely what that issue is. Accordingly, I propose that we proceed, following up selectively on some of the remarks quoted, by modifying the claim that the worry is not about circularity to the claim that it is not analytical circularity that is worrisome—since, as Wright says, "the project is not to analyse the concepts in question"—but, picking up on the remarks about dependence (in the characterization of appropriate circumstances or suitable conditions) "on facts about the actual extension" of the concepts being treated, that the worry does concern what we distinguished from analytical circularity and called *inferential* circularity in our opening section. Although the characterization of inferential circularity given there was somewhat vague (or at any rate unclear, in view of its appeal to the notion of a circular argument) the possibility of an inference 'in the offing' in this context is explicitly raised in the following remarks by Mark Johnston (discussing an idea favoured by Wright to the effect that the 'basic equations' involve an asymmetry, their biconditionality notwithstanding, between the right- and left-hand sides, with the latter determining the former rather than conversely)[23]:

> Maybe the idea is this. Suppose you know the basic equation for *P* and you are told the extension of the predicate 'believes that *P*', that is, you are told who believes that *P*. You are also told under what conditions various people believe that *P*. Then you would be able to use this information and the basic equation to determine the extension [= truth-value] of *P*.

We shall take up this suggestion in the following section, for which we adopt as a provisional hypothesis the idea that the kind of circularity Wright wants to avoid is what we have called inferential circularity; in Section 5 we shall note some difficulties for this hypothesis. While these may make for some awkward-ness in continuing to use the 'inferential circularity' terminology, the interest of the observations of Sections 3 and 4 (and Appendix B) should survive any relabelling that awkwardness might prompt.

3. Independence and Intensionality

Johnston's example in the passage quoted just before the Postscript to the preceding section is simpler than the case of the secondary qualities, *inter alia* because it replaces a one-place predicate (such as 'is red') by a zero-place predicate (a sentence, that is: in Johnston's case represented by *P*), for which the appropriate extension is a truth-value rather than a class of individuals. In the further interests of simplicity, let us follow this lead not just for the case of the embedded expression but also for the embedding expression, replacing 'believes that' by '*a* believes that', so that we end up with something—*a* believes that *P*—again capable of being true or false *tout court*, rather than true or false of this or that individual. Note that *a* and *P* are here taken to be, respectively, a specific name and sentence: they are, in particular, not variables. Then Johnston's suggested offering to Wright becomes the following: one can use the imagined equivalence of (1) and (2)

(1) *P*

(2) *a* believes that *P*

to infer the truth-value of (1) from that of (2). But what has become of Wright's idea of independence? An equivalence is precisely a denial—one kind of denial—of independence. We can restore that idea by considering the general case. This time we *do* need a variable, and will use 'α' as a metalinguistic variable ranging over sentences. Whatever we say about specific cases, we can still say that, *in general*, (3) and (4) are independent:

(3) α

(4) *a* believes that α

If we think of (4) as representing a context into which any sentence can be spliced by replacing the 'α' with that sentence, then it seems correct to say—though exactly what we might mean by saying this will need further clarification (below)—that the context in question is intensional. This accords with Wright's restriction for securing independence—again a notion calling for further clarification—adumbrated in a passage quoted in the preceding section: that such use as there may be of 'the targeted concepts' must be confined to within the scope of intensional operators. To connect this up with the notion of inferential circularity introduced in our opening section, we should say that such operators provide contexts which are what were there called 'suitably protective': recall that the contrast was with the case of analytical circularity, against a charge of which there is simply no context in which to protect the 'targeted' concept (*K* in our schematic formulations). Any occurrence whatever, however deeply or shallowly embedded, of an expression with that concept as its sense, within the right-hand side of a biconditional, makes the biconditional unacceptably circular as a ('reductive') analysis of the concept. But our main interest in this section is in the notions of independence and intensionality in Wright's discussion, and we can begin by noting straight away that his restriction that occurrences of the targeted concept be confined to intensional contexts does not, after all, suffice to secure the desired independence. We recall what Wright said about this matter: first, that there was an independence condition to the effect that "the relevant concepts are to be involved in the formulation of the C-conditions only in ways which allow the satisfaction of those conditions to be logically independent of the details of the extensions of those concepts"; secondly, that if the concepts are to used at all, they (*sc.* terms and predicates expressing them) must occur only inside the scope of intensional operators; and thirdly, that the motivation for this restriction is "that it *suffices* to preempt a certain kind of complaint." Now we have already seen, with the example in Section 1 in which knowing oneself to be a person was replaced by believing oneself to be a person as a condition putatively necessary and sufficient for being a person, in order to avoid inferential circularity, that satisfying the 'intensional confinement' restriction by no means suffices for securing Wright's independence desideratum. The point is that any understanding of the notion of intensionality which rules '*a* believes that' to be intensional will make this same ruling in the case of '*a* knows that', whereas whatever can be said in favour of the ('in general') independence between (3) and (4) has no plausibility for the case of (3) and

(5) *a* knows that α

precisely, of course, because (3) follows from (5), the intensionality in (5) notwithstanding.

To throw some light on what is going on here, and to supply a corrected account of the relationship between independence and intensionality, at least for the simple setting—essentially that of sentential logic—into which we have transposed the discussion, we need to get clear

about what these two notions amount to. For that purpose, we need to make precise the idea of independence as complete combinatorial independence in respect of truth-value, and of intensionality as non-truth-functionality. (Another candidate understanding of intensionality will be mentioned in Appendix B.) And for that purpose we use the notion of a *valuation* as a map from the sentences of the language under consideration to the set {T,F} of truth-values. We have in mind, as a semantics for the language, a collection of such valuations, the 'admissible' valuations if you like, to reflect our inferential practices: what we count as (logically) following from what, what is consistent with what, and so on. As is well known, a class V of valuations gives rise to a consequence operation Cn_{V} *via* the definition:

$$Cn_{\mathsf{V}}(\Gamma) = \{\alpha \mid \forall v \in \mathsf{V}\ (v(\gamma) = \mathrm{T} \textit{ for all } \gamma \in \Gamma) \Rightarrow v(\alpha) = \mathrm{T}\}$$

which we may take as representing the logic induced by selecting V as the class of valuations we consider to be admissible. (Here 'Γ' ranges over sets of sentences, and 'α', 'γ', over individual sentences.) As is also well known, a single consequence operation can be so induced by more than one class of valuations, which is why we have chosen to *start* with a selection of what are to count as the admissible valuations.[24]

In order to define what it is for a set of sentences to be independent, we invoke the ancillary notion of what we shall call a 'ban' on certain truth-value assignments. Where each $x_i \in \{\mathrm{T},\mathrm{F}\}$ $(1 \leq i \leq n)$, we say that a class V of valuations *obeys a ban on* $\langle x_1,\ldots,x_n\rangle$ *for* $\langle\alpha_1,\ldots,\alpha_n\rangle$ just in case there is no $v \in \mathsf{V}$ with $v(\alpha_1) = x_1,\ldots, v(\alpha_n) = x_n$. Thus to say, for instance, that the class of admissible valuations obeys a ban on $\langle \mathrm{T},\mathrm{F},\mathrm{T}\rangle$ for $\langle\alpha,\beta,\gamma\rangle$ is to say that we are not prepared to countenance the (logical) possibility that α,β,γ should respectively be true, false, and true. We can now define sentences $\alpha_1,\ldots,\alpha_n$ to be *independent* with respect to a class V of valuations when for no $x_1,\ldots,x_n \in \{\mathrm{T},\mathrm{F}\}$ does V obey a ban on $\langle x_1,\ldots,x_n\rangle$ for $\langle\alpha_1,\ldots,\alpha_n\rangle$; the sentences are independent *tout court* when this holds with respect to the class V of all admissible valuations. In other words, we have simply made explicit the idea that no possible combined assignment of truth-values to the sentences in question is ruled out. For the remainder of the discussion, variables (possibly with subscripts) such as $\alpha,\beta,\ldots$ range over sentences, and those such as x, y,…over truth-values (over the set {T,F}, that is).

For an n-ary sentential operator #, and a class V of valuations, we say that V respects the *determinant* $\langle x_1,\ldots,x_n,x_{n+1}\rangle$ *for* #, just in case V obeys a ban on $\langle x_1,\ldots,x_n,\overline{x}_{n+1}\rangle$ for all $\langle\alpha_1,\ldots,\alpha_n,\#(\alpha_1,\ldots,\alpha_n)\rangle$, where $\overline{\mathrm{T}} = \mathrm{F}$ and $\overline{\mathrm{F}} = \mathrm{T}$.[26] Each row in a truth-table specification of the truth-function associated with a boolean connective is a determinant for the connective. Thus the class of 'boolean' valuations respects the determinants $\langle \mathrm{T},\mathrm{T},\mathrm{T}\rangle$, $\langle \mathrm{T},\mathrm{F},\mathrm{F}\rangle$, $\langle \mathrm{F},\mathrm{T},\mathrm{F}\rangle$ and $\langle \mathrm{F},\mathrm{F},\mathrm{F}\rangle$ for the binary connective $\wedge$ (conjunction) for example.[27] (We treat a sentence connective as a special case of an operation from sentences to sentences, not distinguishing the connective from the mapping from components to the compound formed by them with its aid.) An n-ary sentence operation # is, we shall say, *fully determined* (*partially determined*, *completely undetermined*) w.r.t. V iff for all (some, no) $x_1,\ldots,x_n$, V respects a determinant $\langle x_1,\ldots,x_n,x_{n+1}\rangle$ for #. Of course, an alternative and more familiar terminology exists for the 'fully determined' case: we say that # is truth-functional w.r.t. V; but it is helpful to have a form of words which makes explicit the contrast with the other cases. The connectives (more generally, contexts for sentences) which are called truth-functional *tout court* are those which

truth-functional (= fully determined) w.r.t. the class of all admissible valuations. Conjunction falls into this category since for every choice of a pair $\langle x,y\rangle$ of truth-values, some truth-value z appeared, on the list above, in a determinant $\langle x,y,z\rangle$ for $\wedge$. Indeed, making the usual identification of an n-ary function with a (certain kind of) $(n+1)$-ary relation, and of the latter with a set of ordered $(n+1)$-tuples, we can say that the set of all determinants V respects for # just is the truth-function associated over V with # (in the sense of note 27). (A complication arises in one special case of being fully determined—which we might call the case of being *overdetermined*—and that is when V respects two determinants for # which differ only in their final position. As this is only possible under pathological circumstances—in fact only when for all $v \in \mathsf{V}$, either $v(\alpha) = \mathrm{T}$ for every sentence α or else $v(\alpha) = \mathrm{F}$ for every sentence α—we shall ignore it below. The interested reader is left to verify that overdetermination, as defined here, does indeed imply full determination. The trouble these cases cause for our formulation is not there is no truth-function associated with # over V but that there isn't a unique such truth-function.)

For the sake of illustration, let us select for our class V of admissible valuations the class of characteristic functions of sets of sentences (formulas) which are maximal consistent w.r.t. Hintikka's favoured logic of knowledge and belief, these being represented by 1-ary connectives K and B, respectively.[28] The boolean connectives are all fully determined w.r.t. this choice of V, while B is completely undetermined, and K is partially but not fully determined, since V respects the determinant $\langle \mathrm{F},\mathrm{F}\rangle$ for K and no determinant of the form $\langle \mathrm{T},x\rangle$. The former determinant is respected simply because the logic in question sanctions the inference from $\mathrm{K}\alpha$ ('*a* knows that α') to α, for any α, so that if $v \in \mathsf{V}$ assigns F to α, v must assign F to $\mathrm{K}\alpha$. Recall that we are now exploring the hypothesis that intensionality is best understood as non-truth-functionality, and we have already noted that truth-functionality is a matter of being fully determined, so that (w.r.t. the chosen V) the expectation that K and B should count as intensional is fulfilled, neither of these connectives being fully determined. But, rephrasing the point about partial determination for K, we have V obeying a ban on $\langle \mathrm{F},\mathrm{T}\rangle$ for $\langle \alpha, \mathrm{K}\alpha\rangle$, for all α, so for no α are α and $\mathrm{K}\alpha$ independent. By contrast, for some choices of α, we do have α and $\mathrm{B}\alpha$ independent (*e.g.*, choose α as a sentence letter), so that the 'in general' independence of belief from truth is respected. Wright, as we saw, claimed that a sufficient condition for the independence of $\#\alpha$ and α was that # was intensional, which for the moment we are taking to mean 'non-truth-functional'; as the present working through of the example originally given in (1)–(5) above shows (taking # as K), we have here a case in which the condition is satisfied but what it is supposed to be sufficient for does not obtain. The mistake appears to be that of confusing not being fully determined ('non-truth-functionality') with the stronger property of being completely undetermined; it is the latter property that is linked with the independence phenomenon, not the former. We can trace this linkage explicitly by dropping the talk of 'in general' independence in favour of suitable quantification over sentences, as well as bringing into the open the fact that it is, specifically, the independence of a compound and its components that we are concerned with, by defining an n-ary sentence operation # to possess, w.r.t. a class V of valuations, the *compositional independence property* just in case there exist sentences $\alpha_1,\ldots,\alpha_n$, such that $\alpha_1,\ldots,\alpha_n$, and $\#(\alpha_1,\ldots,\alpha_n)$ are independent w.r.t. V. When we now compare this property with intensionality, as currently construed (an alternative construal being amongst the topics considered in Appendix B), we find that we can make the following:

CLAIM. *With respect to any class* V *of valuations:*

(i) *If an operation # has the compositional independence property then it is non-truth-functional.*

(ii) *The converse of (i) does not hold in general.*

In view of the above Claim, a proof of which may be found in Appendix B, the relationship between intensionality, construed as non-truth-functionality, and the compositional independence property, is that intensionality is necessary, though not, *pace* Wright, sufficient for the compositional independence property (w.r.t. an arbitrary V). We already knew the 'non-sufficiency' part of this conclusion, of course, from the example of knowledge (the 'K' operator of epistemic logic), but it is worth having a systematic analysis of the situation – for which, see Appendix B. We could, incidentally, have illustrated the point with either the notion of necessity or that of possibility (the $\Box$ and $\Diamond$ operators as they behave in any normal modal logic between **KT** and **S5**). These examples, like the original, are of singulary sentence operations; for an example in which $n = 2$, take the case of the counterfactual conditional, $\Box\!\!\rightarrow$, which is non-truth-functional w.r.t. the class of valuations treated as admissible for the counterfactual logic favoured by David Lewis,[29] though that class of valuations respects for $\Box\!\!\rightarrow$ the determinants $\langle$T,T,T$\rangle$ and $\langle$T,F,F$\rangle$, reflecting a failure of compositional independence. (In terms of the induced consequence operation *Cn*, such a conditional and its consequent are equivalent given the antecedent, in the sense that $Cn(\{\alpha, \alpha \,\Box\!\!\rightarrow\}) = Cn(\{\alpha, \beta\})$ for all α, β.)

4. Meta-Ethical Applications

In the preceding section, we stressed the incorrectness of equating, in the way the remarks quoted from Wright arguably do equate, non-truth-functionality with what we have called the compositional independence property: in particular, as we have seen, a context (such as that provided by knowledge ascriptions) may be non-truth-functional without affording compositional independence.[30] Whether or not Wright makes this mistake, it is certainly made in Arnold Johanson's discussion of Hume's is/ought barrier.[31] Johanson sets up some notation (p. 341) by supposing that "*B* is a deontic operator and *P* is a descriptive formula", with a view to replacing (in the formal language he is considering) occurrences of the latter which are in the immediate scope of the former, by new unstructured expressions, to be regarded as normative formulas, and he writes:

> These replacements should be regarded as merely technical devices. Thus we replace a given purely normative formula by one containing no descriptive predicate letters. It should be clear that this is possible since the non-truth-functional nature of deontic operators makes the truth of *B*(*P*) independent of the truth of the descriptive formula *P*.

The mistake arises, of course, in what follows the 'since': non-truth-functionality simply does *not* guarantee the claimed independence. No doubt we should hold it against any deontic logic that the distinctively deontic operators therein treated did not possess the compositional independence property (w.r.t. the valuations comprising the intended semantics): the point is simply that this is not a consequence of the failure of truth-functionality (with respect thereto).[32]

A second application of the ideas of Section 3 to meta-ethical theory arises in connexion with what has been variously called analytic, definitional, or semantic naturalism.[33] Here it is not the compositional independence property of deontic concepts themselves that matters, so much as the compositional independence property of belief (and certain other propositional attitude contexts). The way this comes to matter to the formulation of a naturalism of the present type (as opposed to what is variously called 'ontological', 'metaphysical', or 'property-identification' style naturalism) is over the precise range of concepts deemed admissible in the account of a moral concept for that account to qualify as naturalistic. We can take it from G. E. Moore's own discussion of naturalism and of the naturalistic fallacy that the straightforwardly moral concepts of being right, wrong, good, or bad are paradigmatically to be excluded from this range, while such concepts as conduciveness to survival, social organization, and the general happiness, are to be included. But what of the status of the concept of being approved of, or—what we may take to be the same thing—being believed right? It will be replied that if a putative account of rightness involved a reference to what was — perhaps by a suitably circumstanced subject (an 'ideal observer', for example) — believed to be right, then the concept of rightness itself would be involved *via* the involvement of the concept of being believed to be right. This, however, would be to overlook the precise *way* the concept of rightness was involved in the proposed account: in particular the occurrence of 'right' (to return to the formal mode) to which we are now attending is itself in the scope of 'believes', an operator whose possession of the compositional independence property means that, even though analytically circular, the account escapes the more serious charge of being circular in virtue of having the analysandum appear other than in the protective scope of such operators. In view of these considerations, one would have to think twice before endorsing the following line of thought, from Michael Smith:[34]

> Definitional naturalism is the view that we can define moral terms exclusively in terms apt for describing the subject matter of the natural and social sciences. The catch-cry of definitional naturalists is therefore not just analysis, but *reductive* analysis.

Since there is no reason to deny social psychologists and anthropologists, for example, recourse to descriptions in terms of approval, disapproval, shame, etc., and no reason to deny the apparent embeddedness of moral concepts in these, a philosophical account of the moral concepts in terms of such propositional attitudes and affective states should count as ('definitionally') naturalistic without being reductive (*i.e.*, analytically non-circular). There are then, two distinct kinds of 'analytic, definitional, or semantic' naturalism, since a proposed account of the application conditions of a moral concept may be wholly free of moral vocabulary

(the 'reductive' case), or it may contain 'protectively embedded' moral vocabulary. In the latter case we do not have, on the usage of Section 1 above, an *analysis*, because of the analytical circularity: but a biconditional's failure as an analytical proposal does not of course mean that it fails to be analytic (in the sense of the analytic/synthetic distinction), so that we need not refrain from speaking of analytic naturalism in this case.

5. Coda: Inferential Circularity and Compositional Independence

We have seen a relatively clear distinction between two ways a concept may appear within the right-hand side of a biconditional account of its own application conditions: either—let's put it—'safely', in the sense that each such occurrence is protected within the scope of an operator with the compositional independence property, or else 'unsafely', in the sense that at least one occurrence is not thus protected. But what of our provisional hypothesis that this latter, 'unsafe', case coincides with the phenomenon of inferential circularity as introduced in Section 1? Since that introduction was made in terms of the idea of the circularity of an argument or inference, a notion which gives every appearance of being a pragmatic-cum-epistemic notion, while compositional independence is a purely semantic matter, the chances of establishing the provisional hypothesis are not good. We close with a few inconclusive but relevant points bearing on this issue.

Inferential circularity was introduced rather casually in our opening section, describing it to afflict an account of the application conditions of some concept "when any argument or inference from premises claiming the various conditions provided by the account to obtain, to the conclusion that the concept applies, is itself circular in whatever sense an inference or argument can be circular." Now, whatever one may think of the circularity of an inference of the form α, *Therefore* α (*cf.* Sorensen [1991]), even such inferences of the contrasting form $\neg\alpha$, *Therefore* α as are valid would never be considered circular: they do anything but help themselves to their conclusions while listing their premises. Yet a circular account of a putative predicative concept *K* according to which anything is *K* if and only if it is not *K* (or indeed—to avoid worries about inconsistency—just the 'if' half of this account) certainly counts as inferentially circular by 'unprotected occurrence' test, since negation lacks the compositional independence property.[35] Thus quite apart from the above worry about semantic *vs.* pragmatic matters, the notion answering to that test is not going to be exactly what was introduced in terms of circular inference in Section 1. The question of how to react to this divergence is terminological. One could retain the original meaning for 'inferentially circular account', and introduce a more general category for accounts in which the concept under discussion appears unprotected: we could call such accounts *blatantly* circular, for example, and put the point we have extracted from Wright's discussion by saying that not all analytical circularity is, in this technical sense, blatant circularity. Alternatively, one could retain the term 'inferential circularity' for this category, in honour of the motivating examples of Section 1, without trying to deny that the interesting property thereby exemplified goes beyond the tight connection with circular inference they themselves most immediately suggest.

A circular argument is at least a valid argument, and if we concentrate on this aspect of circularity as it impinges on the notion of inferential circularity as introduced in Section 1, we find the following contrast with the idea of circularity-*via*-unprotected-occurrences.[36] That introduction spoke only of circularity in an argument from premisses to the effect that the conditions provided (by a given account) were satisfied to a conclusion stating that the concept in question applied, which is to say: a circularity of inference when the account on offer was employed from right to left. In the simplest case (exhibiting circularity), with a monadic predicative concept P for which the account

$$\forall x(Px \leftrightarrow O(Px))$$

is proposed, 'O' being some singulary sentence operation, we are dealing with the validity of arguments of the form $O(Pa)$, *Therefore* Pa. McGinn's non-triviality criterion (from Section 2) will require that while such arguments are valid, the same does not hold when 'P' is replaced (at both occurrences) by an arbitrary expression of the same syntactic category — in this case, that of a monadic predicate. If we fill out the picture by making the same comments about the corresponding left-to-right arguments, we arrive at something arguably deserving to be called a kind of independence: incomparability with respect to entailment. A stronger notion of independence sometimes employed would require (to illustrate with the present instance) not only that neither Pa nor $O(Pa)$ imply the other, but also that neither should imply the negation of the other. But the complete combinatorial independence involved in what we have been calling the compositional independence property (for 'O' in the present case) is stronger still, since for this we should also require that the negation of Pa does not imply $O(Pa)$.[37] (We assume that negation behaves classically as far as the admissible valuations are concerned: that such valuations assign the value T to a sentence just in case they assign F to its negation.) It is a failure of this 'non-subcontrariety' condition that the previous paragraph illustrated by, in effect, taking O as $\neg$ itself. One measure, then, of the distance of the informally introduced notion of inferential circularity from that of 'unsafe' or 'blatant' circularity, is given by exactly how much is packed into the condition of independence which is violated by a manifestation of such circularity.[38]

APPENDIX A: Historical Postscript

The suggestion, in Section 2, that it was particularly the 1980's that saw a reaction against 'knee-jerk' circularity objections, would need to be qualified in a historically more thorough discussion of these matters — and more so than is done in the following remarks, which trace such sentiments back to the previous decade. We mention the work of Christopher Peacocke and that of C. Mason Myers.[39]

Peacocke's 1975 discussion of what makes a language, considered as semantically individuated, the actual language of a given population, contains clear anticipations of some of the defence of circular accounts we have seen in later publications on the part of others.[40] He criticizes

> the view that a definition or biconditional can be of philosophical interest only if one side of it contains only predicates that can be applied in any given case in advance of any application of the predicate being explained on the other side. This will be false in any case in which biconditionals serve to connect one concept with

> another and thereby exhibit connexions without a specification of which any account of that concept will be incomplete, without *ipso facto* effecting a reduction of that concept.

As it happens, when addressing specifically the topic of dispositional style accounts of secondary quality concepts, he does not share with the writers whose views we have been citing an inclination to be similarly liberal about (analytical) circularity. In a famous discussion,[41] Peacocke used, on the right-hand side of the biconditional for *red*, the predicate *red′* (a predicate applying to regions of the visual field), rather than risking analytical circularity with the semantically structured 'looks red'. His idea that while this novel predicate could not be understood other in terms of the familiar *red*, such 'cognitive priority' need not work in the same direction as 'definitional priority', one concept being definitionally prior to another if the latter can be illuminatingly defined in terms of (*inter alia*) the former.[42] Unfortunately, this ambitious distinction between two notions of priority amongst concepts turned out not to do the circularity-avoiding work it was introduced for,[43] and in his more recent treatment of the topic, Peacocke approaches the matter in a different way, addressing himself to the possession conditions, rather than (directly) to the application conditions, of the concept *red*.[44]

It remains to acknowledge the overlap between the considerations presented above on the part of McGinn, Burgess and Johnston (and, earlier, Peacocke) and those to be found in Myers [1978].[45] Myers considers several illustrations of what, if presented as analyses, would be vulnerable to objection on grounds of circularity, one of which (from p. 3) we quote here:

> *X* knows that *P* if and only if *P* is true, *X* believes *P*, there is no false proposition *Q* such that *X* would not believe that P if he did not believe that *Q*, *X* has adequate evidence for *P*, and *X* is aware that this evidence is adequate for *P*.

Before saying a few words about this proposed account of the application conditions of the concept of knowledge, which Myers labels '(1)', we cite the relevant methodological remarks:

> Since (1) has a *prima facie* resemblance to a classical analysis but is not reductive it may be appropriately called a *quasi-analysis*. If (1) is true and the purpose of stating it is not reduction but simply elucidation through revealing the logical relation of the analysandum to certain other concepts, then (1) can function effectively. Prior to formulating (1) the analyzer was so unclear about the logical relation of the analysandum to certain other concepts that he mistakenly identified knowledge with true opinion, but by formulating (1) he corrected that error and gained considerable clarity as he listed necessary conditions that had not previously occurred to him. If (1) is true the analyzer has gained insights into logical relations and the reasons for the inadequacy of his previous formulations, and the circularity of (1) need not conflict with these insights.

Myers does not of course claim that (1) provides a correct 'quasi-analysis': the point is simply that its circularity does not of itself undermine its potential to illuminate the concept of knowledge.[46] More interesting are the grounds Myers gives for claiming it to be circular. One of those is the uncontentious point that the final clause mentions the subject's *awareness* that the evidence in question is adequate: a blatant appearance, under a different name perhaps, of the concept whose application conditions are at issue. Indeed, unlike the cases reviewed in the main body of this section, that concept appears 'unprotected' by embedding in whatever kind of

context (the topic of the following section) might be deemed to provide for analytical circularity without inferential circularity. But another of Myers' grounds is worth looking at more closely:

> If *X* believes that *P*, then *X* must *know* that he is committed to certain things, for without a knowledge of anything implied by *P*, *X* has no understanding of *P* and cannot properly be said to believe it.

The interest of this remark is of course that it would force us to hold any putative analysis of the concept of knowledge in terms, *inter alia*, of belief to be circular. But should we go along with this? Suppose we grant that no-one can believe anything unless certain appropriately related things are known. Does it follow that the concept of belief cannot but be understood in terms of the concept of knowledge, and that the latter concept is therefore covertly presupposed at every deployment of the former concept? Presumably not. Let us introduce the term 'clasps' with the special sense given by: one *clasps* that *P* just in case one either believes that *P* or desires that *P*. Then, in the manner of the above remark on Myers' part, we can say: if *X* believes that *P*, then *X* must clasp that *P* (for unless *X* clasps that *P*, *X* can neither believe nor desire that *P*). It does not seem a plausible claim, however, to say that the concept of belief cannot but be understood in terms of the concept of clasping, or that the latter is presupposed by the former.[47]

APPENDIX B: A Closer Look at the Logical Picture

We begin with a proof of the main claim made in Section 3, repeated here for convenience:

CLAIM. *With respect to any class* V *of valuations:*

(i) *If an operation # has the compositional independence property then it is non-truth-functional.*

(ii) *The converse of (i) does not hold in general.*

The Claim holds without restriction on the 'arity' of #, but we shall give the proof, to avoid a clutter of subscripts, for the case of $n = 1$.

To say that (1-ary) # has the compositional independence property is to say that there exists a sentence α with α and $\#\alpha$ independent w.r.t. the chosen V, in the sense that V obeys no ban on $\langle \alpha, \#\alpha \rangle$. Thus to say that # *lacks* this property is to say that for all α there exist x,y such that V obeys a ban on $\langle x,y \rangle$ for $\langle \alpha, \#\alpha \rangle$. This in turn amounts to the following (for which we continue the numbering scheme from where we left off in Section 3):

(6) $\forall\alpha\exists x,y\forall v \in \mathsf{V}\ [v(\alpha) = x \Rightarrow v(\#\alpha) = \overline{y}]$

Now since we may obtain (6) from (7) by moving a universal quantifier leftward over an existential quantifier, (6) follows from (7):

(7) $\exists x,y\forall\alpha\forall v \in \mathsf{V}\ [v(\alpha) = x \Rightarrow v(\#\alpha) = \overline{y}]$

Notice that (6) and (7) are respectively equivalent to the results of removing the overlining from 'y' in these conditions, and that (7) simply says that V respects some determinant for #. Since (7) implies (6), being partially determined implies lacking the compositional independence property. Contraposing, then, we conclude

(8) *If # has the compositional independence property w.r.t. a class of valuations, # is completely undetermined w.r.t. that class.*

Since being completely undetermined implies not being fully determined w.r.t. a given class of valuations, and the latter is equivalent to being non-truth-functional w.r.t. the class, (8) gives us part (i) of the Claim above. As for part (ii), suppose that non-truth-functionality implied the compositional independence property (w.r.t. an arbitrary V). Then, by (8), non-truth-functionality would imply being completely undetermined. But non-truth-functionality is equivalent to not being fully determined, so we should then have the (clearly false) conclusion that not being completely determined implied being fully undetermined.

We turn to the task of describing more fully the relationship between the concepts figuring in the above proof, beginning with that describing more fully the relationship between the concepts figuring in the above proof, beginning with that between the compositional independence property and the property of being completely undetermined. (8) above established a one-way implication between these properties (w.r.t. a given class of valuations); in most familiar logical

settings, this implication can be reversed, so that the two notions are equivalent (as we have noted is the case, without any qualifications, for the corresponding weaker notions of non-truth-functionality and the property of being not fully determined). The key to establishing the converse of (8) is the following condition on an n-ary sentence operation #:

(9) $\exists\beta_1\ldots\beta_n\forall x_1\ldots x_n y$ [$\vee$ *obeys a ban on* $\langle x_1,\ldots,x_n,y\rangle$ *for* $\langle\beta_1,\ldots,\beta_n,\#(\beta_1,\ldots,\beta_n)\rangle$
$\Rightarrow \forall\gamma_1\ldots\gamma_n$: $\vee$ *obeys a ban on* $\langle x_1,\ldots,x_n,y\rangle$ *for* $\langle\gamma_1,\ldots,\gamma_n,\#(\gamma_1,\ldots,\gamma_n)\rangle$

What follows the '$\Rightarrow$' here says that $\vee$ respects the determinant $\langle x_1,\ldots,x_n,\overline{y}\ \rangle$ for #, but we have preferred the more long-winded formulation to bring out the relationship with closure under Uniform Substitution: if the language in question has a class of sentence letters (propositional variables), the idea is that whatever the logic induced by $\vee$ says holds for various sentences should continue to hold when the sentence letters therein are replaced uniformly by arbitrary sentences.[48] The propositional variables are 'generic' for the logic, that is. (9) is a special case of the condition of closure under Uniform Substitution, since it simply requires that for the given #, there should exist n sentences (the β_i) which are generic (in view of the universally quantified γ_i) for any ban involving them as components along with the #-compound formed from them.

With the aid of (9) as an additional assumption on $\vee$ and (as before, for simplicity, 1-ary) # we can establish the converse of (8), by showing that (6) implies (7). Suppose, accordingly that (6) is satisfied: for all α there exist x,y with $\vee$ obeying a ban on $\langle x,y\rangle$ for $\langle\alpha, \#\alpha\rangle$. Let β (= β_1) be as promised for # by (9), and let this be our α. Since for β we have some x,y with $\vee$ obeying a ban on $\langle x,y\rangle$ for $\langle\beta, \#\beta\rangle$, by (9), for all sentences γ, $\vee$ obeys a ban on $\langle x,y\rangle$ for $\langle\gamma, \#\gamma\rangle$: but this—relettered—is what (7) says. Thus with the (typically available) additional condition (9) in the background, we have a very simple picture of the relationships between our four concepts: not only are non-truth-functionality and non-full-determination equivalent, but so are the (in general) stronger concepts of having the compositional independence property and complete undetermination. It is in this setting that our earlier diagnosis of the mistake of conflating non-truth-functionality with the compositional independence property as one of 'confusing not being fully determined ('non-truth-functionality') with the stronger property of being completely undetermined' is at its most accurate.

The discussion in Section 3 of Wright's remarks, on which we have just been elaborating, was premised on the provisional identification of intensionality with non-truth-functionality, which, as we have seen, was too weak ('not fully determined' as opposed to 'completely undetermined') to support the claim that intensionality sufficed for the compositional independence property. This suggests that we re-open investigation of that claim with a stronger notion of intensionality to hand. Such a stronger notion is, as it happens, made available by taking intensionality as non-extensionality, for an independently familiar notion of extensionality which is readily seen to deserve that name, and which is weaker than truth-functionality. This notion—extensionality in sentence position, we might call it for greater explicitness—is the sentential analogue of extensionality as understood in areas outside the confines of sentential logic, and, as in the earlier discussion, our precise definition involves a relativity to classes of valuations, droppable when the class in question comprises precisely the admissible valuations.

(Recall that, as a technical convenience, we take the latter concept as primitive.) Say that an n-ary sentence operation # is *extensional* w.r.t. V when for all $v \in \mathsf{V}$ and all $\alpha_1,...,\alpha_n, \beta_1,...,\beta_n$:

$v(\alpha_i) = v(\beta_i)$ *for all* i $(1 \leq i \leq n)$, *implies* $v(\#(\alpha_1,...,\alpha_n)) = v(\#(\beta_1,...,\beta_n)$

Every # which is truth-functional w.r.t. a given V is extensional w.r.t. V, though not conversely: so this notion is, as promised, weaker than truth-functionality. By way of illustration of the 'not conversely', consider the language of classical sentential logic and the operation # defined by: $\#\alpha = \alpha \wedge q$, where q is some (fixed) propositional variable (sentence letter). This 1-ary # is clearly not truth-functional (w.r.t. the class of all boolean valuations) since there is no singulary truth-function f for which, whatever boolean valuation v we select, we have, for all α, $v(\#\alpha) = f(v(\alpha))$, but it is certainly extensional, because any boolean valuations agreeing on the truth-value of α will agree on the truth-value of $\#\alpha$. For a second example, take V as the class of characteristic functions of sets of formulas which are maximal consistent w.r.t. the modal logic **K**, and $\#\alpha$ as $\alpha \leftrightarrow \Box\bot$; in possible worlds terms, this formula has the same truth-value as α at worlds with no worlds accessible to them, and the opposite value (to α's) at other worlds: again it is easy to see that at any worlds at which α and β have the same truth-value, so do $\#\alpha$ and $\#\beta$, even though there is no fixed truth-function usable for computing a #-compound's truth-value from that of its component.[49]

The relationship between truth-functionality and extensionality is most easily displayed by adapting the definition in note 27 of what it is for a truth-function to be associated with a sentence operations over a class of valuations; inserting an appropriate existential quantifier, and understanding the variable 'f' to range over n-ary truth-functions, we have (n-ary) # truth-functional w.r.t. V just in case:

(10) $\exists f \forall v \in \mathsf{V}\ \forall \alpha_1,...,\alpha_n: v(\#(\alpha_1,...,\alpha_n) = f(v(\alpha_1),...,v(\alpha_n))$

whereas it is not hard to verify that # is extensional w.r.t. V just in case the following weaker condition (moving the '$\exists$' rightward) is satisfied:

(11) $\forall v \in \mathsf{V}\ \exists f \forall \alpha_1,...,\alpha_n: v(\#(\alpha_1,...,\alpha_n) = f(v(\alpha_1),...,v(\alpha_n))$

So although not necessarily truth-functional w.r.t. a class of valuations in the sense of being associated with some one truth-function over the whole class, an operation which is extensional w.r.t. that class # is associated, over each singleton subset thereof, a truth-function: extensionality is what we might call a 'variable' or 'local' version of truth-functionality.[50]

Does replacing non-truth-functionality with non-extensionality as our construal of intensionality improve the prospects for the claim that intensionality suffices for compositional independence? We can see with the same example as before—that of the epistemic operator 'K'—that the claim remains false. For, w.r.t. the class of valuations invoked for assessing that claim on its earlier construal, this connective is not only non-truth-functional, but non-extensional as well, its lack of the compositional independence property notwithstanding. In view of this speedy negative verdict, the introduction of the current construal of intensionality in terms of the notion of extensionality in sentence position may seem hardly to have been worth the trouble. But the contrast between (10) and (11) is in fact suggestive of another issue which deserves our

attention, namely: whether the compositional independence property with which we have been working isn't too restrictively defined. To make visible the possibility of a global/local contrast here, parallel to the (10)/(11) contrast, it helps to define the notion of an individual valuation's respecting a determinant (as opposed to the notion in play above, of a class of valuations respecting a determinant). We say that a valuation v respects a determinant d (as a determinant) for the sentence operation # just in case the class $\{v\}$ respects d for #. We could equally well have started with this valuation-by-valuation notion of respect, explaining it thus: v respects $d = \langle x_1,\ldots,x_n,x_{n+1}\rangle$ as a determinant for n-ary # ($1 \leq i \leq n + 1$), if and only if for all $\alpha_1,\ldots,\alpha_n$, if $v(\alpha_i) = x_i$ (all $1 \leq i \leq n$) then $v(\#(\alpha_1,\ldots,\alpha_n)) = x_{n+1}$. Then the earlier class-relative notion emerges *via*: V respects d for # just in case every $v \in \mathsf{V}$ respects d for #.

With the above apparatus to hand, what we have been calling partial determination of # w.r.t. V obtains when (using 'd' to range over determinants of length $n + 1$, where # is n-ary):

(12) $\exists d \forall v \in \mathsf{V}$ (v respects d for #)

and we can, as in the passage from (10) to (11), see the 'local' variant as emerging by interchanging the quantifiers — # being *locally partially determined* w.r.t V if and only if:

(13) $\forall v \in \mathsf{V}$ $\exists d$(v respects d for #)

For a simple example of this local version of partial determination, consider the smallest normal modal logic to contain all formulas of the form

(14) $(\Box\alpha \rightarrow \alpha) \vee (\beta \rightarrow \Box\beta)$

This logic is sound and complete with respect to the class of all Kripke models whose accessibility relations R satisfy the condition that for all points x, y in the model: if Rxy then Rxx.[51] (The easiest way to see the completeness half of this claim is to work instead—in a canonical model argument—with the condition that for all x, y: if Rxy and $x \neq y$, then Rxx. This is clearly equivalent to the original condition, which itself is more suggestive in connexion of an alternative syntactic description of the same modal logic: as the least normal modal logic containing every formula of the form $(\Box\alpha \rightarrow \alpha) \vee \Box\bot$.)

Letting V comprise the characteristic functions of sets of formulas maximally consistent w.r.t. this modal logic, it is clear that (12) is not, while (13) is, a correct description of the situation. Given any $v \in \mathsf{V}$ we have either $v(\Box\alpha \rightarrow \alpha) = \mathrm{T}$ for all formulas α, or else $v(\beta \rightarrow \Box\beta) = \mathrm{T}$ for all formulas β. In the former case, v respects the determinant $\langle \mathrm{F},\mathrm{F}\rangle$ for $\Box$, while in the latter case v respects $\langle \mathrm{T},\mathrm{T}\rangle$ for $\Box$. Of course, instead of saying "or else $v(\beta \rightarrow \Box\beta) = \mathrm{T}$ for all formulas β" we could equally well have said "or else $v(\alpha \rightarrow \Box\alpha) = \mathrm{T}$ for all formulas α": we only needed to use distinct schematic letters in (14) to block an unwanted identification as they appear within the same schema. (Such an identification would turn (14) into something whose instances were in turn substitution instances of a truth-functional tautology.) The instance of (14) which replaces the distinct schematic letters by distinct propositional variables (or 'sentence letters') shows that the present logic is Halldén-incomplete, since this would be a disjunction neither of whose disjuncts belongs to the logic, their lack of a common propositional variable notwithstanding.[52]

Now, in view of the failure of (12) for the present case (taking # as $\Box$), what we have called the compositional independence property for $\Box$ is secured (w.r.t. the indicated choice of V),

even though the fact that (13) is satisfied reveals that there is an analogous 'local' compositional independence property which □ conspicuously lacks.[53] We recall (again) that Crispin Wright proposed a restriction to the effect that "the relevant concepts are to be involved in the formulation of the C-conditions only in ways which allow the satisfaction of those conditions to be logically independent of the details of the extensions of those concepts", and that we are taking this to be a prohibition on embedding in the scope of anything but operators with the compositional independence property (including, of course, the 'null embedding') — or at least to generalize naturally to such a prohibition when the reference to C-conditions is abstracted from. The original compositional independence property seems appropriate as a rendering of the idea here, and having introduced the possibility of this local version of the property, we leave it for those who dissent on this score to explore the ramifications of the alternative rendering.

NOTES

1. This remark appears on p. 100 of Butler [1736].

2. Shoemaker [1970]; the idea appears in the terminology '*q*-memories' in Parfit [1971].

3. There is not even a consensus as to whether circularity, as a fault in arguments, is the same as begging the question: see Palmer [1981]; Palmer also provides (note 1) a useful bibliography of published discussions of these questions, which the interested reader may update by a perusal of the footnotes to Sorensen [1991].

4. Unlike Butler, Reid [1815] takes the notion of memory in Locke's account in the non-veridical ('apparent memory') sense, and accordingly makes a rather different objection, claiming that in this account "personal identity is confounded with the evidence which we have of our personal identity" (p. 359). I mention this to correct the impression given by Reid, who writes (p. 356) that Locke's "doctrine on this subject has been censured by Bishop Butler (…) with whose sentiments I perfectly agree".

5. A discussion of this question may be found in Humberstone and Townsend [1994]. Note that the suspicion of an implicit use of the concept of identity (which the paper just cited attempts to dispel) arises from the use of reflexive pronouns or repeated names—in a formalized version, of repeated individual variables or constants—and not from the suggestion that what is remembered is that the individual remembering is identical with the original experiencer. (The latter suggestion is laboriously contested in Palma [1964].)

6. Another potentially troublesome source of analytical circularity is the parenthetical gloss in our characterization of quasi-memory "*via* whatever type of causal route in characteristic of the laying down of memories" which deploys the very concept (that of memory) whose involvement with identity was the reason for using the 'quasi-' variant. This matter is addressed in Perry [1975].

7. A further advantage, over and above that mentioned in the preceding paragraph, of this change of example is that we can by-pass the question of whether there is such a thing as analysing (or defining) the concept of personal identity. As Lombard [1986] puts the point,

giving a negative answer to this question (at p. 43): "no such definitions are offered by a criterion of identity. At best, a criterion of identity leads (…) to a definition of what it is to be a φ." (Lombard is discussing, in particular, circularity objections to Davidson's original criterion of identity for events.)

8. Dummett [1973]. The distinction introduced by Dummett here was later popularized in Chapter 1 of Nozick [1981] in terms of 'proof' *versus* 'explanation'.

9. Blackburn [1993] acknowledges that whether or not the analytical circularity of an account is a defect does in principle depend on what the account aims to provide, but remains unconvinced of the philosophical significance of (extant) accounts which, in view of their circularity, cannot be regarded as analyses.

10. Markus [1988].

11. A review of some of the relevant (though far from edifying) literature may be found in Bayles [1991].

12. In Lassiter [1983].

13. Gupta and Belnap [1993]; Yablo [1993].

14. This quotation is actually from p. 462 of Gupta [1987].

15. The quotation which follows is from p. 147.

16. The quotation which follows is from p. 6.

17. Michael Smith, addressing these questions in Chapter 2 of Smith [1994], takes a more aggressive stand on the terminological question, arguing that a circular analysis is properly so called because of the traditional understanding of what conceptual analysis was supposed to be *for*. The idea is that there are 'platitudes' involving any given concept, recording those of our inferential and judgmental dispositions concerning that concept which have a *prima facie* claim to *a priori* status: "As the project of conceptual analysis is being thought of here, then, an analysis of a concept is successful just in case it gives us knowledge of all and only the platitudes which are such that, by coming to treat those platitudes as platitudinous, we come to have mastery of that concept.(…) [T]he dispositional analysis [of *red*] is perhaps best seen as an attempt to *encapsulate*, or to *summarize*, or to *systematize*, as well as can be done, the various remarks we come to treat as platitudinous in coming to master the term 'red'. In this admittedly vague sense, it can therefore lay some claim to giving us knowledge of all the platitudes. And, accordingly, is can therefore lay some claim to constitute an analysis." (p. 31*f*.) Concerning non-reductive analyses (*i.e.*, what we are calling analytically circular accounts), Smith writes: "For, as we have already seen, colour terms can be given a summary-style, non-reductive characterization in the style of the dispositional analysis. If the goal of an analysis is simply to give us knowledge of all the platitudes surrounding our use of the concept up for analysis, then the mere fact that such an analysis is summary-style and non-reductive is, of course, no objection to it." (p. 52)

18. The quotation is from note 13 of Burgess [1983]. Since Burgess actually takes himself to be addressing the assertibility-conditions rather than the truth-conditions of claims to the effect that an object is of a certain colour, he is not (strictly speaking) interested in what we are calling

the application-conditions of the colour concepts (but rather—shall we say?—in their "ascription-conditions").

19. Johnston himself spells out the notion of triviality slightly differently on p. 106 of his [1993]: "circularity of the indicated sort would be a defect only if it made the biconditionals and their associated identities empty." The preceding page of this paper houses, incidentally, a remark in the spirit of the above quotation from Burgess: "So we have in each of these philosophically interesting examples the assertion of an *interdependence* between the philosophically interesting concept and the relevant concept of our disposition to respond."

20. Specifically, we are concerned with what Holton [1991] calls echo concepts, and especially *judgment-dependent* echo concepts which are also *users'* concepts. I have adapted Holton's use to speak of an echo account rather than an echo concept, to avoid involvement with the issue of whether, if some account with this feature is correct for a concept, than every correct account for that concept must possess the feature. A similar adaptation is made in Blackburn [1993], p. 261.

21. Johnston himself favours a different non-counterfactual account of dispositions, and so is happy to remain with 'unprovisoed' biconditionals whose right-hand sides speak of the disposition to respond (appropriately) in the (relevant) circumstances. Edited highlights of Johnston's theory of dispositions may be found in Appendix 2 of Johnston [1993].

22. These remarks appear on p. 120*f*. of Wright [1992].

23. The details of this view (apparently once favoured by Johnston himself) do not concern us here; readers interested in the dialectics of the debate between Wright and Johnston are referred to the Appendix ('The Euthyphro Contrast: Order of Determination and Response-Dependence') to Chapter 3 of Wright [1992] and to Appendix 3 ('On Two Distinctions') of Johnston [1993]. The quotation which follows is from p. 124 of the latter work.

24. The observation here reported was emphasized especially by Rudolf Carnap in Sections 14–27 of Carnap [1943]. Instead of writing '$\alpha \in Cn_{\mathsf{V}}(\Gamma)$' we may write, using 'consequence relation' notation: '$\Gamma \vdash_{\mathsf{V}} \alpha$'; this has the advantage of conveniently generalizing to allow multiple or empty right-hand sides. Let us denote such generalized consequence relations (originally used by Gentzen, but—in effect—urged for present purposes by Carnap) by '$\Vdash$'. The generalized consequence relation induced by V is defined by: $\alpha_1,\ldots,\alpha_m \Vdash_{\mathsf{V}} \beta_1,\ldots,\beta_n$ iff V obeys a ban on $\langle \mathrm{T},\ldots,\mathrm{T},\mathrm{F},\ldots,\mathrm{F}\rangle$, where there are m occurrences of T and n of F, for $\langle \alpha_1,\ldots,\alpha_m,\beta_1,\ldots,\beta_n\rangle$. (The terminology of 'obeying a ban' is defined in the following paragraph.) The restriction to finitely many sentences on the left and right here is for ease of exposition, and is by no means essential: see Chapter 2 of Segerberg [1982].

25. Such prohibitions can be expressed in more 'positive' terms as conditional constraints: for example, in the case just given, as the constraint that if $v(\alpha) = \mathrm{T}$ and $v(\beta) = \mathrm{F}$, then $v(\gamma) = \mathrm{F}$: V obeys the ban in question just in case every $v \in \mathsf{V}$ satisfies this constraint. But we prefer the formulation in terms of bans because distinct such conditional constraints correspond in the manner just illustrated to the same ban. For instance, we could equivalently have cited the constraint that if $v(\alpha) = \mathrm{T}$ and $v(\gamma) = \mathrm{T}$, then $v(\beta) = \mathrm{T}$. Thus a representation of the data in terms of conditional constraints has to deal with a multiplicity of constraints and a suitable

relation of (inter)derivability amongst them, which, simple though the rules for such a 'calculus of constraints' would be, presents a distraction by contrast with the unique representation of the data in terms of bans.

26. This is a form of the 'conditional constraint' idea just disparaged — but it is too useful to do without, and in any case does not suffer from the multiplicity problem: the claim that V respects some determinant $\langle x_1,\ldots,x_n,x_{n+1}\rangle$ for a given n-ary # and the claim that V respects the determinant $\langle z_1,\ldots,z_n,z_{n+1}\rangle$ are equivalent claims (if and) only if $x_1 = z_1$, and … and $x_{n+1} = z_{n+1}$.

27. We say that an n-ary truth-function f is 'associated with' # over V just in case for all $v \in \mathsf{V}$, for all $\alpha_1,\ldots,\alpha_n$, $v(\#(\alpha_1,\ldots,\alpha_n)) = f(v(\alpha_1),\ldots,v(\alpha_n))$; the boolean connectives are those with which there is a traditional such function associated over the class of valuations used in evaluating validity in classical sentential logic; a boolean valuation is a valuation v such that each boolean connective is associated over $\{v\}$ with the traditional truth-function in question. The remarks in the text about determinants (which are members of truth-functions, on the usual identification of functions with functional relations and of relations with sets of tuples) are an adaptation of the theory of type determination in §3.3 of Segerberg [1982].

28. Hintikka [1962].

29. Lewis [1973].

30. D. H. Sanford also shows the need for caution in view of 'knows that' contexts in one's account of truth-functionality, in Sanford [1970]. [[*Cf.* also the remarks quoted from Kenny on Chisholm in 'Updates and Afterthoughts' below – last sentence of the inset quotation.]]

31. Johanson [1973].

32. It is easily verified that in the normal modal logic **KD** often regarded as a basic system of deontic logic the modal operators have the compositional independence property, and that this is so *a fortiori* for the smallest normal modal logic **K**. This point was made by J. R. Lucas on p. 24 of Lucas [1982], in the words "[In these systems] we have no connexion between modalised and ordinary unmodalised formulae". For the logics mentioned, this is a special case of a 'lack of connexion' between occurrences of the same non-logical vocabulary at different levels of modal embedding.

33. For distinctions amongst naturalisms, see Pigden [1991]; Jackson [1992]; and Chapter 2 of Smith [1994].

34. Smith [1994], p. 35.

35. This example appears as (13) on p. 273 of Gupta and Belnap [1993], as a circular definition which is perfectly legitimate according to their theory of definition, even though "it is admittedly strange and doubtless defines a notion that is completely useless".

36. One of the referees has pointed out that on a conception of circularity of argument according to which an argument (as propounded on a certain occasion) is circular if one of the premisses is accepted (by the proponent) only because the conclusion is accepted, it does *not* follow that circular arguments are all of them valid: so this is a point on which our discussion is not neutral as between competing conceptions of circular argument.

37. This and the preceding notion of independence are distinguished as 'modal' and 'deductive' independence, respectively, Wolniewicz [1970].

38. For their assistance in the preparation of this material, I am grateful to John (A.) Burgess, Michael Smith, and Andrew Markus. In addition, several stylistic and organizational improvements are due to the referees for *Philosophy and Phenomenological Research*.

39. A referee has drawn my attention to a similar anticipation in Shoemaker [1979]; see especially the paragraph on p. 333 beginning, "I hasten to point out that this account is circular".

40. Peacocke [1974]; the quoted passage appears on p. 162.

41. Peacocke [1983].

42. Peacocke's official definition of cognitive priority is: "concept A is cognitively prior to B iff no one could possess the concept B without possessing the concept A" (Peacocke [1983], p. 42).

43. See Smith [1986].

44. Peacocke [1992], pp. 8, 27–9.

45. The quotations from Myers [1978] which follow are reproduced verbatim, with no attempt to correct the inconsistent use of '*P*', '*Q*', which sometimes occupy the place of sentences, and sometimes the place of names (of propositions).

46. Felicia Ackerman likewise uses the term 'quasi-analysis' (as well as 'partial analysis') for analytically circular accounts on p. 328 of Ackerman [1992], and like Myers, illustrates their interest with the example of the concept of knowledge.

47. I also have a difficulty with Myers' main example of circular explication, which arises from a consideration of Josiah Royce's idea of a self-representing system. Myers (quasi-) defines a geographical map to be '*Royceian* if and only if it is in the region it maps and contains an accurately placed and detailed part a Royceian map of the same region", but is the circularity necessary (or even appropriate)? Why not say "if and only if it accurately maps some region in which it is properly located"? From the supposition that there is a Royceian map *M* of a given region, it will then follow that *M* is located within the region and depicts a Royceian map (namely *M* itself, at the very least) of the region: but there is no need to try and pack all such consequences of a definition—or, more accurately, of the supposition that the defined term applies—into the definition itself.

48. The generalized consequence relation $\Vdash_{\mathsf{V}}$ (from note 24) is substitution-invariant, that is.

49. This second example is adapted from one given in Humberstone [1986], where a general discussion of the issues involved may be found (in a slightly different terminology). For more on the first example, see Section 3 of Humberstone [1993].

50. The notions of truth-functionality and (sentence position) extensionality are frequently confused, or at least discussed in confusion-inviting terminology; for example Gupta and Belnap ([1993], p. 249) describe as truth-functional the context 'John says something truth-functionally equivalent to the proposition that ___'.

51. Such soundness and completeness results are often expressed by talk of the logic's being determined by a given class of models (or, if preferred, of frames). Since this would invite confusion with the present notions of complete (partial, etc.) determination, we avoid the locution here.

52. An informative discussion of this topic may be found in Lemmon [1966]. The term 'logic' in the sentence to which this note is appended refers to $Cn_{\mathsf{V}}(\varnothing)$.

53. Lacking this property w.r.t. V amounts to its being the case that:

$$\forall v \in \mathsf{V}\ \exists x,y\ \forall \alpha.\ \{v\} \text{ obeys a ban on } \langle x,y\rangle \text{ for } \langle \alpha, \Box\alpha\rangle.$$

'Two Types of Circularity': Updates and Afterthoughts

The reference to Humberstone [1986] in note 49 should be supplemented by one to a later paper on the same topic, Humberstone [1997*a*], which had not appeared when this chapter was originally written. There is no similar excuse for the lack of any reference to Kenny [1963], on p. 201 of which the following passage appears, in a discussion of the account of intensionality in Chisholm [1957]:

> Chisholm offers another criterion of intensionality which is of assistance here. He says that any non-compound sentence Q which contains a propositional clause P is intensional provided that neither Q nor not-Q imply either P or not-P. By this criterion, "it can be the case that *p*" is not intensional, even though "it can be the case that …" is not a truth-functional operator; for "it cannot be the case that *p*" implies "not-*p*". But once again, Chisholm has given a sufficient, but not a necessary, condition of intensionality. By this criterion, "John knows that Queen Anne is dead" is not intensional, since it implies that Queen Anne is dead.

This is essentially our point about the need to distinguish non-truth-functionality from compositional independence.

Though it is too early for any published feedback on the content of 'Two Types of Circularity' to have appeared, John A. Burgess – the author of Burgess [1983], referred to in the paper (not to be confused with the American philosopher and logician John P. Burgess) – has been working on a refinement of some of the ideas in the paper. A work-in-progress version of his ideas was delivered as a seminar paper, listed in our bibliography as Burgess [1999]. The main point is that the original paper concentrates on issues that can be spelt out at the level of propositional (or sentential) logic, and in particular embeddings under this or that sentence operator: but this is an oversimplification, since we might want to say that a sentence in which the internal subject-predicate structure is made explicit, say as *Fa*, is susceptible of an analytically but not inferentially circular semantic treatment, when we are told that necessary and sufficient for its truth is the truth of, for example, *Fb*, where *b* is the name of some object other than (that named by) *a*. This itself oversimplifies Burgess' concerns somewhat, but conveys the essential drift. In particular, Burgess wants to think of defining the predicate *F* rather than a particular predication *Fa*, so whereas *b* here should be thought of as a proper name, *a* is merely a parameter so that we do not have a free variable such as *x*: really what we are defining is the

general predication Fx and the definition should be explicitly stated in the form $\forall x(Fx \leftrightarrow Fb)$. (In fact the example Burgess mentions is a bit more complicated than this one, namely:

$$\forall x(Fx \leftrightarrow ((Fx \wedge Gx) \vee Fb))$$

which he regards as only selectively circular, being potentially informative in its application to a (as x) when $a \neq b$. This raises an issue at the purely propositional level to be discussed in the following paragraph.) This would be an analytically circular proposal if it is thought of as part of delimiting the extension of the predicate F, since the condition cited employs this very predicate. But although the occurrence of the predicate in question occurs in the scope of no protective sentential operator on the right-hand side (i.e., one with the compositional independence property), since it occurs completely unembedded, the condition allows us to fix a truth-value for the whole predication Fa without already settled on such a truth-value, in every case except the case in which $a = b$. For this reason, Burgess regards the definition as selectively non-viciously circular. A somewhat similar situation may be held to occur with inductive definitions, which give the application conditions of a predicate to one object in terms of its application to other – 'earlier' in terms of the ordering underlying the induction – objects. The latter type of circularity was known to me at the time of writing the paper – it comes up in the references cited by Yablo, Belnap and Gupta, for instance – and I can only welcome the fact that Burgess is working on how to extend the basic framework of 'Two Types of Circularity' to bring these refinements into the fold.

It should also be mentioned that even at the level of propositional logic, Burgess's proposed criterion of non-viciousness differs from my own formalized version of inferential non-circularity in terms of the restriction that all occurrences of the term being defined which appeared in the definition appeared there in the ("protective") scope of an operator with the compositional independence property. Burgess has a different idea, expressed at the level of propositional logic by saying that a circular definition, for (e.g.) a monadic predicate F will have the form (tacitly understood as universally quantified)

$$Fx \leftrightarrow (\_\ \_\ \_)$$

with "Fx" recurring somewhere in the *definiens* (represented here by the blanks), and that this was benignly (non-viciously) just in case there was neither a verification loop nor a falsification loop involved. The latter concepts are clarified like this. A *verification loop* is involved when the finished (or 'completed' – i.e., every rule which can be applied has been applied) tree – as in Jeffrey [1967] – for $\_\ \_\ \_$ has 'Fx' on every open path, and a *falsification loop* is involved when the finished tree on $\neg(\_\ \_\ \_)$ has '$\neg Fx$' on every open path. Since, essentially as noted above, we should not really be thinking of open formulas like these occurring in a tree, we may the above provisional gloss to have its references to 'x' replaced by references to some constant 'a' not occurring in the definition

$$\forall x(Fx \leftrightarrow (\_\ \_\ \_)).$$

For brevity we may replace the analogue of 'Fa" or 'Fx" with "p". With this replacement, Burgess mentions in the chapter (developing the ideas of Burgess [1999]) on circular definitions in a book manuscript in progress, that a definition of p as $p \wedge q$ involves a verification loop, while defining p as $p \vee q$ would instead give a falsification loop. The intuitive idea is that to verify

p using the definition, we are instructed to verify the *definiens* $p \wedge q$, but since this requires verifying (*inter alia*) p, the procedure has gone into a loop. *Mutatis mutandis* for the case of $p \vee q$ with falsification in place of verification. And the formal criterion delivers the desired results since every open path in the finished tree on $p \wedge q$ (resp. $\neg(p \vee q)$) has p (resp. $\neg p$) on it. In these two cases his verdicts coincide with those of 'Two Types of Circularity': while he says in these cases the circularity is not benign because one or other type of loop is involved, I say the circularity is not just analytic but inferential because there is no protective embedding.

We can combine the some aspects of the conjunctive and disjunctive cases just reviewed by making p a conjunct of a disjunct: consider the definition of p as $(p \wedge q) \vee r$, and in this case Burgess's and own my verdicts differ. (Think of this as shorthand for the case in which a predicate F – assumed monadic for simplicity – is defined by the universal closure of $Fx \leftrightarrow ((Fx \wedge Gx) \vee Hx)$.) Since there is still no protective embedding, I say we have inferential circularity, the 'bad kind' of circularity, whereas since there is no verification loop (the *definiens* does not require the truth of p for its truth – for a justification of which gloss, see the following paragraph) or falsification loop (the *definiens* does not require the falsity of p for its falsity) involved, he rules the case as one of benign circularity. This is perhaps not so much a matter of disagreement as of the emergence of more distinctions needing to be drawn than were initially evident. My negative verdict is due to the fact that under some circumstances, e.g., if q is true and r is false, being told that p has the same truth value as $(p \wedge q) \vee r$ does not determine a truth-value for p, whereas his favourable verdict reflects the fact that under some circumstances, e.g., if r is true, or if q and r are both false, a truth-value for p is determined by the equivalence of p with $(p \wedge q) \vee r$. So while I demand a determination for *all* cases, he demands only that such a determination be forthcoming in *at least one* case. (This kind of characterization of the difference between our positions is due to Burgess. It does not completely capture the difference between the positions, however, because of what we call 'twisted' loops below.) There are two perfectly well-defined notions here, and whether one is better suited than the other to explicating the informal idea of non-viciously circular definition must await further investigation of the applications of that idea. The most interesting outcome of such an investigation would be that for some such applications the one notion is appropriate while for others, the other is. But whether this turns out to be so must await the results of the investigation.

Say that a formula B is a *consequence by decomposition* of a set Γ of formulas if every open path in the Jeffrey tree constructed on the basis of Γ has the formula B on it. In general, B can be a consequence (according to classical propositional or predicate logic) of Γ without being a consequence of Γ by decomposition; for example $\Gamma = \{p \wedge q\}$ has the formula $q \wedge p$ as a consequence but not as a consequence by decomposition. However, by a simple argument not included here, we can show that whenever B is either an atomic or a negated atomic formula, B is a consequence of a set Γ of formulas if and only if B is a consequence of Γ by decomposition. Thus Burgess's idea of a verification loop being involved in a definition with *definiens* D for some *definiendum* E amounts to this: D has E as a logical consequence. Similarly, a falsification loop amounts to $\neg$D having $\neg$E as a logical consequence, and thus to D's being a logical consequence of E. So freedom from loops of either kind, which is the mark of benign circularity, is actually a special case of a certain notion of independence. We might call this bilateral independence, meaning simply mutual non-entailment. Having extracted this from

Burgess's suggestion, we can make a comparison with that of 'Two Types of Circularity' for an especially simple kind of analytically circular definition, as illustrated by the example from Section 1 of that paper: (for all x) x is a person if and only if x believes that x is a person. The schematic form of such cases is: $Pa \leftrightarrow \mathrm{O}(Pa)$, in which the variable has been replaced by a parameter a, and P and O are respectively a one-place predicate and a one-place sentence operator. For circular definitions of this particularly simple form, there is a simple constrast between Burgess's criterion and my own. As the coincidence between consequence and consequence by decomposition for atomic and negated atomic formulas (here Pa and $\neg Pa$) shows, Burgess's looplessness criterion amounts to requiring a certain bilateral independence relation to obtain, where as the compositional indepedence property makes a stronger demand: Pa and $\mathrm{O}(Pa)$ should not just be bilaterally independent, they should be completely logically independent – which adds to bilateral independence the further requirement that both can be true together and both can be false together.

To obtain a clearer view of the relation between Burgess's proposal and my own, the following more fine-grained terminology will be helpful. We take the issue at the level of propositional logic, so that what is in question is an analytically circular definition of the form $\mathrm{p} \leftrightarrow \mathrm{A(p)}$, where the notation is as explained above. Let us say that such a definition involves an *inevitable verification loop* if p occurs on every open path through the tree constructed on A(p), an *inevitable falsification loop* if ¬p occurs on every open path in the tree constructed on ¬A(p), a *potential verification loop* if p occurs on some open path in the tree constructed on A(p), and a *potential falsification loop* if ¬p occurs on some open path in the tree constructed on ¬A(p). The definition is *Burgess-benign*, abbreviated below to "B-benign", the analytical circularity notwithstanding, if it involves neither an inevitable verification loop nor an inevitable falsification loop. JB calls this benignness "inferential groundedness", recalling without exactly reproducing the terminology of inferential non-circularity (or absence of inferential circularity) in my paper. He suggests that my precise characterization of the latter notion, in terms of the compositional independence property, amounts to the following, which I'll call Humberstone-benignness: an analytically circular definition is *Humberstone-benign* (abbreviated below to "H-benign") just in case it involves neither a potential verification loop nor a potential falsification loop. Thus the difference between being B-benign and being H-benign is that the former requires the absence of *inevitable* looping of either the verification sort or the falsification sort, while the latter requires the absence of even *potential* looping of either sort. Note that provided A(p) is consistent, so that there is at least one open path in the tree on A(p), inevitable verification (resp. falsification) looping implies potential verification (resp. falsification) looping, though not conversely, so setting aside the case of inconsistent A(p), H-benignness is a stricter requirement that B-benignness.

Now, while H-benignness in the above sense might be a reasonable thing to understand by inferential non-circularity, and indeed the remarks in the opening section of 'Two Types of Circularity', but it does not quite coincide with the precise explication of the latter notion (in terms of the compositional independence property) in the later sections of that paper. To the extent that the tree apparatus can serve to capture this notion – which is something of a moot point pending a consideration (not supplied here) as to how to adapt it to more than truth-functional and quantificational modes of (de)composition – we need to add further prohibitions

to the above bans on potential verification and falsification loops. Call loops of these kinds *straight* loops: we are worried about the occurrence of the *definiendum* (i.e., p) in open paths of the tree based on the *definiens* (our A(p)), or of the negated *definiendum* in open paths of the tree based on the negated *definiens*. Contrast the idea of a *twisted* loop: occurrences of the negated *definiendum* in open paths of the tree based on the (unnegated) *definiens*, or of the (unnegated) *definiendum* in open paths of the tree based on the negated *definiens*. We can consider the strengthening of the notion of H-benignness – no potential straight loops involved – to the following notion: no potential straight *or twisted* loops involved. (In a full discussion, it would be helpful, in the interests of providing a complete inventory of the available notions of benignness, to start by considering instead the idea of banning just inevitable loops – straight or twisted). We take over the potential/inevitable contrast here from its earlier use in connection with what we are now calling straight loops: just read the earlier definitions only understanding "loop" as we now understand it, to cover twisted as well as straight loops. The actual benignness notion – inferential non-circularity – from 'Two Types' is not the absence of potential straight looping but the absence of any potential looping, straight or twisted. There are no examples of this when only the boolean connectives are used, in fact, which is not surprising since none of these have the compositional independence property.

The fact that twisted loops are being excluded may seem to make the proposal as it is finally formlized, in terms of compositional independence, a long way from the initial idea of inferential circularity as corresponding to the circularity of an argument from *definiens* to *definiendum*, since while an argument making essential use of the conclusion as a premiss in establishing that conclusion is normally regarded as circular, the same cannot be said for an argument which makes use instead of the negated conclusion. Such an argument is pictured better as involving a Möbius strip than a circle. Whether this extension of the original idea is appropriate remains to be seen. Certainly, our earlier remark to the effect that "There are two perfectly well-defined notions here, and whether one is better suited than the other to explicating the informal idea of non-viciously circular definition must await further investigation of the applications of that idea" is an oversimplification, since we have at least the following three benignness notions in play: Burgess-benignness (no inevitable straight looping), Humberstone-benignness (no potential straight-looping), and – what comes closest to the 'Two Types' proposal – no potential straight or twisted looping. I say "what comes closest" because to bring into the discussion intensional operators of the kind considered in the paper, one would need to clarify the notions of looping so as to apply to them, which would require an extension of the tree method to beyong the confines of extensional logic. That is not an exercise we can undertake here.

Keefe [2002] has some points critical of the material in this chapter some of which are similar to those from Burgess; others are different, however, and I hope to extend these comments to touch on them at some stage. *A propos* of note 31, where it is said that in the logic **KD**, "modal operators have the compositional independence property, and that this is so *a fortiori* for the smallest normal modal logic **K**. This point was made by J. R. Lucas on p. 24 of Lucas [1982], in the words '[In these systems] we have no connexion between modalised and ordinary unmodalised formulae,' I can now add a reference to Zolin [2000], in which such logics (i.e., those making no non-trivial modalized-to-unmodalized connections) are called *modalized logics*.

LIST OF ALL REFERENCES CITED

The chapters in which the work in question is cited appear in angle brackets at the end of the entry; a superscripted '+' means that the work is cited in the 'Updates and Afterthoughts' postscript to the chapter in question but not in the main body of that chapter.

Ackerman, F. [1992] 'Paradoxes of Analysis' (pp. 326–329 of J. Dancy and E. Sosa (eds.) *A Companion to Epistemology*, Blackwell, Oxford. ⟨Ch.7⟩

Adams, E. W. [1975] *The Logic of Conditionals*, Reidel, Dordrecht. ⟨Ch.6⟩

Anderson, C. A., and J. Owens (eds.) [1990] *Propositional Attitudes: The Role of Content in Logic, Language, and Mind*, Center for the Study of Language and Information (Stanford), Lecture Notes #20. ⟨Ch.0⟩

Anscombe, G. E. M. [1957] *Intention*, Blackwell, Oxford. ⟨Chs.3,6⟩

Austin, J. L. [1953] 'How to Talk – Some Simple Ways', *Proceedings of the Aristotelian Society* **53**, 227–46. ⟨Ch.6⟩

Åqvist, L. [1973] 'Modal Logic with Subjunctive Conditionals and Dispositional Predicates', *Journal of Philosophical Logic* **2**, 1–76. ⟨Ch.1⟩

Axinn, S. [1966] 'Fallacy of the Single Risk', *Philosophy of Science* **33**, 154–162. ⟨Ch.5⟩

Barland, G. H., and D. C. Raskin [1973] 'Detection of Deception', pp. 417–77 in W. F. Prokasy and D. C. Raskin, eds., *Electrodermal Activity in Psychological Research*, Academic Press, New York. ⟨Ch.6⟩

Bayles, M. D. [1991] 'Definitions in Law', pp. 253–267 in J. H. Fetzer, D. Shatz and G. Schlesinger (eds.) *Definitions and Definability: Philosophical Perspectives*, Kluwer, Dordrecht. ⟨Ch.7⟩

van Benthem, J. F. A. K. [1978] 'Two Simple Incomplete Modal Logics', *Theoria* **44**, 25–37. ⟨Ch.4⟩

van Benthem, J. F. A. K. [1984] 'Correspondence Theory' in D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic, Vol. 2*, Reidel, Dordrecht. ⟨Ch.4⟩

Blackburn, S. [1993] 'Circles, Finks, Smells and Biconditionals', *Philosophical Perspectives* **7** (Language and Logic, ed. J. Tomberlin) 259–279. ⟨Ch.7⟩

Bostock, D. [1988] 'Necessary Truth and *A Priori* Truth', *Mind* **97**, 343–379. ⟨Ch.1^{+}⟩

Braithwaite, R. B. [1947] 'Teleological Explanation', *Proceedings of the Aristotelian Society* **47**, pp. i–xx. ⟨Ch.6⟩

Bratman, M. E. [1987] *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge, Mass. ⟨Ch.6⟩

Bricker, P. [1989] 'Quantified Modal Logic and the Plural *De Re*', pp. 372–394 in P. A. French et al. (eds.) *Midwest Studies in Philosophy, Vol. XIV: Contemporary Perspectives in the Philosophy of Language II*, University of Notre Dame Press, Notre Dame, Indiana. ⟨Ch.1^{+}⟩

Burgess, J. A. [1983] 'Perceptual Knowledge and Normal Conditions', *Artisan* 1984 (Issue 2), 25–38; the paper also appeared in *AATADE* (*Australian Association for Tertiary Art and Design Education Journal*) Vol. V (1983–4), 77–88. ⟨Ch.7⟩

Burgess, J. A. [1999] 'Vicious and Virtuous Circularity – What is the Difference?', paper presented to a staff seminar of the Philosophy Department, Monash University, June 1999. ⟨Ch.7^{+}⟩

Butler, J. [1736] 'Of Personal Identity', pp. 99–105 in Perry [1975]. (The discussion originally appeared in Butler's *The Analogy of Religion*, 1736.) ⟨Ch.7⟩

Campbell, D. T. [1969] 'Ethnocentrism of Disciplines and the Fish-Scale Model of Omniscience', in M. and C. Sherif (eds.) *Interdisciplinary Relationships in the Social Sciences*, Aldine. ⟨Ch.5⟩

Carnap, R. [1943] *Formalization of Logic*, conveniently accessible in Carnap, *Introduction to Semantics and Formalization of Logic*, Harvard University Press, Cambridge, Mass. 1961. ⟨Ch.7⟩

Castañeda, H.-N. [1966] '"He": A Study in the Logic of Self-Consciousness', *Ratio* **8**, 130–157. ⟨Ch.3⟩

Castañeda, H.-N. [1967] 'Actions, Imperatives, and Obligations', *Proceedings of the Aristotelian Society* **68** (1967/8), 25–48. ⟨Ch.1⟩

Castañeda, H.-N. [1967*a*] 'Acts, the Logic of Obligation, and Deontic Calculi', *Philosophical Studies* **19**, 13–26. ⟨Chs.1, 3⟩

Castañeda, H.-N. [1972] 'On the Semantics of the Ought-to-Do', pp. 675–694 in Davidson and Harman [1972]. ⟨Ch.1⟩

Castañeda, H.-N. [1975] *Thinking and Doing: The Philosophical Foundations of Institutions*, Reidel, Dordrecht. ⟨Chs.1, 6⟩

Chellas, B. F. [1980] *Modal Logic: An Introduction*, Cambridge. ⟨Ch.5⟩

Chisholm, R. [1957] *Perceiving: A Philosophical Study*, Cornell University Press, Ithaca. ⟨Ch.7^{+}⟩

Cresswell, M. J. [1990], *Entities and Indices*, Kluwer, Dordrecht. ⟨Ch.1^{+}⟩

Cresswell, M. J. [1990*a*] 'Anaphoric Attitudes', *Philosophical Papers* **19**, 1–18. ⟨Ch.2^{+}⟩

Collins, J. D. [1988] 'Belief, Desire, and Revision', *Mind* **97**, 333–342. ⟨Ch.2^{+}⟩

Crossley, J. N., and I. L. Humberstone [1977] 'The Logic of "Actually"', *Reports on Mathematical Logic* **8**, 11–29. ⟨Chs.1, 2⟩

Davidson, D. and G. Harman (eds.) [1972] *Semantics of Natural Language*, Reidel, Dordrecht. ⟨Cited as the source of several items in this bibliography⟩

Davies, M. K. [1976] *Truth, Quantification, and Modality.* D. Phil. Thesis, University of Oxford, 1976. ⟨Ch.1⟩

Davies, M. K. [1981] *Meaning, Quantification, Necessity: Themes in Philosophical Logic*, Routledge and Kegan Paul. ⟨Ch.1^{+}⟩

Davies, M. K., and I. L. Humberstone [1980] 'Two Notions of Necessity', *Philosophical Studies* **38**, 1–30. ⟨Ch.1^{+}⟩

Davis, W. [1981] 'A Theory of Happiness', *American Philosophical Quarterly* **18**, 111–120. ⟨Ch.3⟩

Dennett, D. C. [1968] 'Geach on Intentional Identity', *Journal of Philosophy* **65**, 335–41. ⟨Ch.2⟩

Dennett, D. C. [1971] 'Intentional Systems', *Journal of Philosophy* **68**, 87–106. ⟨Ch.6⟩

Dixon, R. M. W. [1991] *A New Approach to English Grammar, on Semantic Principles*, Clarendon Press, Oxford. ⟨Ch.3^{+}⟩

Dummett, M. [1973] 'The Justification of Deduction', pp. 290–318 in Dummett, *Truth and Other Enigmas*, Duckworth, London 1978. ⟨Ch.7⟩

Dummett, M. [1973*a*], *Frege: Philosophy of Language*, Duckworth, London. ⟨Ch.1^{+}⟩

Evans, G. [1977] 'Pronouns, Quantifiers, and Relative Clauses', *Canadian Journal of Philosophy* **7**, 467–536. ⟨Ch.2⟩

Fitzgerald, P. J. [1962] *Criminal Law and Punishment*, Clarendon Press, Oxford. ⟨Ch.6⟩

Fodor, J. A. [1980] 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology', *The Behavioral and Brain Sciences* **3**, 63–72. ⟨Ch.6^{+}⟩

Foot, P. [1958] 'Moral Beliefs', pp. 83–100 in P. Foot (ed.), *Theories of Ethics*, Oxford 1967. ⟨Ch.3⟩

Forbes, G. [1989] *Languages of Possibility: An Essay in Philosophical Logic*, Basil Blackwell, Oxford. ⟨Ch.1^{+}⟩

Gardner, M. [1977] 'The "Jump Proof" and its Similarity to the Toppling of a Row of Dominoes', Mathematical Games Column in *Scientific American* **236** (May 1977), 128–135. ⟨Ch.4⟩

Garson, J. [1971] 'Here and Now', pp. 145–53 in E. Freeman and W. Sellars (eds.), *Basic Issues in the Philosophy of Time*, Open Court, La Salle, Ill. ⟨Ch.6⟩

Geach, P. T. [1967] 'Intentional Identity', *Journal of Philosophy* **64**, 627–32. ⟨Ch.2⟩

Goldblatt, R. [1982], 'The Semantics of Hoare's Iteration Rule', *Studia Logica* **41**, 141–158. ⟨Ch.5⟩

Goldman, A. I. [1967] 'A Causal Theory of Knowing', *Journal of Philosophy* **64**, 357–72. ⟨Ch.6⟩

Goldman, A. I. [1976] 'Discrimination and Perceptual Knowledge', *Journal of Philosophy* **73**, 771–799. ⟨Ch.4⟩

Gosling, J. [1969] *Pleasure and Desire*, Clarendon Press, Oxford. ⟨Ch.2⟩

Grandy, R. [1973] 'Reference, Meaning, and Belief', *Journal of Philosophy* **70**, 439–52. ⟨Ch.2⟩

Greenspan, P. [1980] 'A Case of Mixed Feelings: Ambivalence and the Logic of Emotion', in Rorty [1980]. ⟨Ch.3⟩

Gupta, A. [1987] 'The Meaning of Truth', pp. 453–480 in E. Lepore (ed.), *New Directions in Semantics*, Academic Press, London. ⟨Ch.7⟩

Gupta, A. and N. D. Belnap [1993] *The Revision Theory of Truth*, MIT Press, Cambridge, Mass. ⟨Ch.7⟩

Halpern, J. and Y. Moses [1985] 'Towards a Theory of Knowledge and Ignorance', in K. R. Apt (ed.), *Logics and Models of Concurrent Systems*, Springer-Verlag. ⟨Ch.4⟩

Hare, R. M. [1971] 'Wanting: Some Pitfalls,' in R. Binkley *et al.* eds., *Agent, Action and Reason*, University of Toronto Press, Toronto. ⟨Ch.2⟩

Hare, R. M. [1981] *Moral Thinking*, Oxford. ⟨Ch.3⟩

Harman, G. [1972] 'Is Modal Logic Logic?', *Philosophia* **2**, 75–84. ⟨Ch.5⟩

Harman, G. [1991] 'Justification, Truth, Goals, and Pragmatism: Comments on Stich's *Fragmentation of Reason*', *Philosophy and Phenomenological Research* **51**, 191–5. ⟨Ch.6⟩

Hazen, A. P. [1976] 'Expressive Completeness in Modal Language', *Journal of Philosophical Logic* **5**, 25–46. ⟨Ch.1⟩

Hazen, A. P. [1978] 'Eliminability of the Actuality Operator in Propositional Modal Logic', *Notre Dame Journal of Formal Logic* **19**, 617–622. ⟨Ch.1⟩

Heal, J. [1988] 'The Disinterested Search for Truth', *Proceedings of the Aristotelian Society* **88**, 97–108. ⟨Ch.6⟩

Heny, F. W. [1970] *Semantic Operations on Base Structures*, UCLA Doctoral Dissertation, Photo-reproduction by University Microfilms Ltd., Tylers Green High Wycombe, England. ⟨Ch.1^{+}⟩

Hilpinen, R. [1969] 'An Analysis of Relativised Modalities', in J. W. Davis *et al.* (eds.), *Philosophical Logic*, Reidel, Dordrecht. ⟨Ch.5⟩

Hilpinen, R. [1974] 'On the Semantics of Personal Directives', in C. Heidrich (ed.) *Semantics and Communication*, North-Holland. ⟨Ch.5⟩

Hilpinen, R. [1977] 'Remarks on Personal and Impersonal Knowledge', *Canadian Journal of Philosophy* **7**, 1–9. ⟨Ch.5⟩

Hintikka, J. [1962] *Knowledge and Belief*, Cornell University Press, Ithaca. ⟨Chs.4 (and.4^{+}), 5, 7⟩

Holton, R. [1991] 'Intentions, Response-Dependence, and Immunity from Error', pp. 83–121 in P. Menzies (ed.), *Response-Dependent Concepts*, Working Papers in Philosophy, #1, RSSS, Australian National University. ⟨Ch.7⟩

Houlgate, L. D. [1966]: 'Mistake in Performance', *Mind* **75**, 257–61. ⟨Ch.6⟩

Humberstone, I. L. [1975], Review of Davidson and Harman [1972], *York Papers in Linguistics* **5**, 195–224 (Department of Language, University of York, England). ⟨Ch.1⟩

Humberstone, I. L. [1982] 'Scope and Subjunctivity,' *Philosophia* **12**, 99–126. ⟨Chs.2,3,6⟩ [= Chapter 1 of this Thesis.]

Humberstone, I. L. [1985] 'The Formalities of Collective Omniscience', *Philosophical Studies* **48**, 401–423. [= Chapter 5 of this Thesis.]

Humberstone, I. L. [1986] 'Extensionality in Sentence Position', *Journal of Philosophical Logic* **15**, 27–54 (and Correction, *ibid.* **17** (1988), 221–223). ⟨Ch.7⟩

Humberstone, I. L. [1987] 'Wanting as Believing', *Canadian Journal of Philosophy* **17**, 49–62. ⟨Ch.6⟩ [= Chapter 2 of this Thesis.]

Humberstone, I. L. [1988] 'Some Epistemic Capacities', *Dialectica* **42**, 183–200. [= Chapter 4 of this Thesis.]

Humberstone, I. L. [1990] 'Wanting, Getting, Having' *Philosophical Papers* **19**, 99–118. [= Chapter 3 of this Thesis.]

Humberstone, I. L. [1990*a*] 'Expressive Power and Semantic Completeness: Boolean Connectives in Modal Logic', *Studia Logica* **49**, 197–214. ⟨Ch.5^{+}⟩

Humberstone, I. L. [1991] 'Two Kinds of Agent-Relativity', *Philosophical Quarterly* **41**, 144–166. ⟨Ch.5^{+}⟩

Humberstone, I. L. [1992] 'Direction of Fit', *Mind* **101**, 59–83. [= Chapter 6 of this Thesis.]

Humberstone, I. L. [1993] 'Zero-Place Operations and Functional Completeness (and the Definition of New Connectives)', *History and Philosophy of Logic* **14**, 39–66. ⟨Ch.7⟩

Humberstone, I. L. [1997] 'Two Types of Circularity', *Philosophy and Phenomenological Research* **57**, 249–280. [= Chapter 7 of this Thesis.]

Humberstone, I. L. [1997*a*] 'Singulary Extensional Connectives: A Closer Look', *Journal of Philosophical Logic* **26**, 341–356.

Humberstone, I. L. [2000] 'The Revival of Rejective Negation', *Journal of Philosophical Logic* **29**, 331–381. ⟨Ch.1^{+}⟩

Humberstone, I. L. [2002] 'Invitation to Autoepistemology', *Theoria* 68, 13–51. ⟨Ch.4^{+}⟩

Humberstone, I. L. [2004] 'Two-Dimensional Adventures', *Philosophical Studies* **118**, 17–65. (Special issue with papers which, like this one, were presented to a conference on this two-dimensional semantics and its applications, held at the RSSS, Australian National University, in February 2001.) ⟨Ch.1^{+}⟩

Humberstone, I. L., and A. V. Townsend [1994] 'Co-Instantiation and Identity', *Philosophical Studies* **74**, 243–272. ⟨Ch.7⟩

Jackson, F. C. [1984] 'Weakness of Will', *Mind* **93**, 1–18. ⟨Ch.5⟩

Jackson, F. C. [1985] 'Internal Conflicts in Desires and Morals', *American Philosophical Quarterly* **22**, 105–114. ⟨Ch.5⟩

Jackson, F. C. [1992] Critical Notice of S. L. Hurley, *Natural Reasons*, *Australasian Journal of Philosophy* **70**, 475–488. ⟨Ch.7⟩

James, F. [1986] *Semantics of the English Subjunctive*, University of British Columbia Press, Vancouver. ⟨Ch.6⟩

James, W. [1891] 'The Will to Believe', in his *The Will to Believe and Other Essays in Popular Philosophy*, Longmans Green and Co., New York ⟨Ch.6⟩

James, W. [1896] 'The Will to Believe', in *The Will to Believe and Other Popular Essays*, Dover 1956. ⟨Ch.5⟩

Jeffrey, R. C. [1967] *Formal Logic: Its Scope and Limits*, McGraw-Hill, New York. ⟨Ch.7^{+}⟩

Johanson, A. [1973] 'A Proof of Hume's Separation Thesis Based on a Formal System for Descriptive and Normative Statements', *Theory and Decision* **3**, 339–350. ⟨Ch.7⟩

Johnston, M. [1989] 'Dispositional Theories of Value', *Aristotelian Society Supplementary Vol.* **62**, 139–174. ⟨Ch.7⟩

Johnston, M. [1993] 'Objectivity Refigured: Pragmatism Without Verificationism', pp. 85–130 in J. Haldane and C. Wright (eds.) *Reality, Representation, and Projection*, Oxford University Press, New York. ⟨Ch.7⟩

Kamp, J. A. W. [1971] 'Formal Properties of "Now"', *Theoria* **37**, 227–273. ⟨Ch.1⟩

Kamp, J. A. W. [1981] 'A Theory of Truth and Semantic Representation', pp. 277–322 in T. Groenendijk *et al.* (eds.) *Formal Methods in the Study of Language, Part 1*, Mathematical Centre Tracts #135, Maethematisch Centrum, Amsterdam. ⟨Ch.2⟩

Kaplan, D. [1973] 'Bob and Carol and Ted and Alice', in J. Hintikka (ed.), *Approaches to Natural Language*, Reidel, Dordrecht. ⟨Ch.2⟩

Keefe, R. [2002] 'When Does Circularity Matter?', Proceedings of the Aristotelian Soc. **102**, 275–292. ⟨Ch.7^{+}⟩

Kelly, J. S. [1978] *Arrow Impossibility Theorems*, Academic Press, New York. ⟨Ch.5⟩

Kenny, A. J. [1963] *Action, Emotion and Will*, Routledge and Kegan Paul, London. ⟨Chs.2, 7^{+}⟩

Kenny, A. J. [1966] 'Practical Inference', *Analysis* **26**, 65–75. ⟨Ch.2⟩

Kenny, A. J. [1966*a*] 'Happiness'. *Proceedings of the Aristotelian Society* **66**, 93–102. ⟨Ch.6⟩

Kripke, S. A. [1982] *Wittgenstein on Rules and Private Language*, Basil Blackwell, Oxford. ⟨Ch.6⟩

Kuhn, S. [1981] 'Logical Expressions, Constants, and Operator logic', *Journal of Philosophy* **78**, 487–499. ⟨Ch.5⟩

Lakoff, G. [1970] 'Linguistics and Natural Logic', pp. 545–665 in Davidson and Harman [1972]. ⟨Ch.3⟩

Lassiter, M. [1983] *Our Names, Our Selves: The Meaning of Names in Everyday Life*, Heinemann, London. ⟨Ch.7⟩

Lemmon, E. J. [1966] 'A Note on Halldén-Incompleteness', *Notre Dame Journal of Formal Logic* **7**, 296–300. ⟨Ch.7⟩

Lemmon, E. J. [1967] 'If I Know, Do I Know that I Know?', pp. 54–82 in A. Stroll (ed.), *Epistemology: New Essays in the Theory of Knowledge*, Harper and Row, New York. ⟨Ch.4^{+}⟩

Lenzen, W. [1978] *Recent Work in Epistemic Logic,* issue 1 *of Acta Philosophica Fennica* **30**. ⟨Ch.4⟩

Lewis, D. K. [1973] *Counterfactuals*, Basil Blackwell, Oxford 1973. ⟨Chs.1, 4, 5, 7⟩

Lewis, D. K. [1974] 'Semantic Analyses for Dyadic Deontic Logic', in S. Stenlund (ed.), *Logical Theory and Semantic Analysis*, Dordrecht, Reidel. ⟨Ch.6⟩

Lewis, D. K. [1979] 'Attitudes *de Dicto* and *de Se*', *Philosophical Review* **88**, 513–43. ⟨Ch.3⟩

Lewis, D. K. [1988] 'Desire as Belief', *Mind* **97**, 323–332. ⟨Ch.2^{+}⟩

Loar, B. [1981] *Mind and Meaning*, Cambridge: Cambridge University Press. ⟨Ch.6⟩

Lombard, B. [1986] *Events: A Metaphysical Study*, Routledge and Kegan Paul, London. ⟨Ch.7⟩

Lucas, J. R. [1982] *Moods and Tenses*, Merton College, Oxford 1982: this typescript is record of lectures given by Lucas on modal logic in Hilary Term of that year. ⟨Ch.7⟩

Markus, A. [1988] 'Australian Governments and the Concept of Race: An Historical Perspective', pp. 46–59 in M. de Lepervanche and G. Bottomley (eds.), *The Cultural Construction of Race*, Sydney Association for Studies in Society and Culture, University of Sydney. ⟨Ch.7⟩

Matthews, G. B., and S. M. Cohen [1967] 'Wants and Lacks', *Journal of Philosophy* **64**, 453–454. ⟨Ch.3⟩

Mayo, B. [1967] 'Belief and Constraint', pp. 147–61 in A. Phillips Griffiths (ed.), *Knowledge and Belief*, Oxford University Press. ⟨Ch.6⟩

McCawley, J. D. [1972] 'A Program for Logic', pp. 498–544 in Davidson and Harman [1972]. ⟨Ch.3⟩

McGinn, C. [1983] *The Subjective View*, Clarendon Press, Oxford. ⟨Ch.7⟩

Meredith, C. A., and A. N. Prior [1965] 'Modal Logic with Functorial Variables and a Contingent Constant', *Notre Dame Journal of Formal Logic* **6**, 99–109. ⟨Ch.1⟩

Millikan, R.G. [1984] *Language, Thought, and Other Biological Categories*, MIT Press, Cambridge, Mass. ⟨Ch.6⟩

Myers, C. M. [1978] 'Circular Explication', *Metaphilosophy* **9**, 1–13. ⟨Ch.7⟩

Nissenbaum, H. F. [1985] *Emotion and Focus*, Center for the Study of Language and Information (Stanford), Lecture Notes #2. ⟨Ch.3⟩

Nozick, R. [1981] *Philosophical Explanations*, Clarendon Press, Oxford, and Harvard University Press, Harvard . ⟨Chs.5, 7⟩

Palma, A. B. [1964] 'Memory and Personal Identity', *Australasian Journal of Philosophy* **42**, 53–68. ⟨Ch.7⟩

Palmer, H. [1981] 'Do Circular Arguments Beg the Question?', *Philosophy* **56**, 387–394. ⟨Ch.7⟩

Parfit, D. [1971] 'Personal Identity', *Philosophical Review* **80**, 3–27. ⟨Ch.7⟩

Parfit, D. [1984] *Reasons and Persons*, Clarendon Press, Oxford. ⟨Ch.6⟩

Parikh, R., and R. Ramanujan [1985] 'Distributed Processes and the Logic of Knowledge (Preliminary Report)', in R. Parikh (ed.) *Logics of Programs*, Lecture Notes in Computer Science #193, Springer-Verlag, Berlin. ⟨Ch.4⟩

Parry, W. T. [1968] 'The Logic of C. I. Lewis', in P. Schilpp (ed.), *The Philosophy of C. I. Lewis*, Cambridge. ⟨Ch.5⟩

Partee, B. H. [1972] 'Opacity, Coreference and Pronouns', pp. 415–441 in Davidson and Harman [1972].

Passy, S., and T. Tinchev [1991] 'An Essay in Combinatory Dynamic Logic', *Information and Computation* **93**, 263–332. ⟨Ch.5^{+}⟩

Peacocke, C. A. B. [1974] 'Finiteness and the Actual Language Relation', *Proceedings of the Aristotelian Society* **75** (1974–5), 147–165. ⟨Ch.7⟩

Peacocke, C. A. B. [1983] *Sense and Content*, Clarendon Press, Oxford. ⟨Chs.6,7⟩

Peacocke, C. A. B. [1992] *A Study of Concepts*, MIT Press, Cambridge, Mass. ⟨Ch.7⟩

Perry, J. (ed.) [1975] *Personal Identity*, University of California Press, Berkeley. ⟨Ch.7⟩

Perry, J. [1975*a*] 'Personal Identity, Memory, and the Problem of Circularity', pp. 135–155 in Perry [1975]. ⟨Ch.7⟩

Pettit, P. [1987] 'Humeans, Anti-Humeans, and Motivation', *Mind* **96**, 530–3. ⟨Ch.6⟩

Pigden, C. [1991] 'Naturalism', pp. 421–431 in P. Singer (ed.), *A Companion to Ethics*, Blackwell, Oxford. ⟨Ch.7⟩

Platts, M. [1979] *Ways of Meaning*, Routledge and Kegan Paul, London. ⟨Ch.6⟩

Porte, J. [1981] 'The Deducibilities of S5', *Journal of Philosophical Logic* **10**, 409–422. ⟨Ch.1^{+}⟩

Powers, L. [1967] 'Some Deontic Logicians', *Noûs* **1**, 381–400. ⟨Ch.1⟩

Price, H. [1989] 'Defending Desire-as-Belief', *Mind* **98**, 119–127. ⟨Ch.6⟩

Prior, A. N. [1961] *Formal Logic,* Second Edn., Clarendon Press, Oxford. ⟨Ch.4⟩

Prior, A. N. [1968] '"Now"', *Noûs* **2** (1968), 101–119. ⟨Ch.1⟩

Prior, A. N. [1968] 'The Formalities of Omniscience', paper III in Prior, *Papers on Time and Tense*, Oxford. ⟨Ch.5⟩

Prior, A. N. [1971] *Objects of Thought* (ed. P. T. Geach and A. J. Kenny), Clarendon Press, Oxford. ⟨Ch.1^{+}⟩

Quine, W. V. O. [1956] 'Quantifiers and Propositional Attitudes', *Journal of Philosophy* **53**, 177–187. ⟨Chs.0,1,2⟩

Reeves, A. [1975] 'Ambiguity and Indifference', *Australasian Journal of Philosophy* **53**, 220–237. ⟨Ch.2^{+}⟩

Reid, T. [1815] *Essays on the Intellectual Powers of Man*, M.I.T. Press, Cambridge, Mass., 1969 (orig. publ. 1813–1815). ⟨Ch.7⟩

Rorty, A. O. (ed.) [1980] *Explaining Emotions*, University of California Press. ⟨Ch.3⟩

Routley, R. [1979] 'Repairing Proofs of Arrow's General Impossibility Theorem and Enlarging the Scope of the Theorem', *Notre Dame Journal of Formal Logic* **20**, 879–890. ⟨Ch.5⟩

Routley, R. and V. Routley [1975] 'The Role of Inconsistent and Incomplete Theories in the Logic of Belief', *Communication and Cognition* **8**, 185–235. ⟨Ch.5⟩

Ryle, G. [1954] 'Pleasure', in *Dilemmas*, Cambridge University Press, Cambridge. ⟨Ch.3⟩

Saarinen, E. [1977] 'Backwards-Looking Operators in Intensional Logic and in Philosophical Analysis', Reports from the Department of Philosophy, University of Helsinki. ⟨Ch.1⟩

Saarinen, E. [1978] 'Intentional Identity Interpreted', *Linguistics and Philosophy* **2**, 151–223. ⟨Ch.2⟩

Salmon, N., and S. Soames (eds.) [1988] *Propositions and Attitudes*, Oxford University Press, Oxford. ⟨Ch.0⟩

Sanford, D. H. [1970] 'What is a Truth-Functional Component?', *Logique et Analyse* **13**, 482–486. ⟨Ch.7⟩

Searle, J. R. [1979] *Expression and Meaning*, Cambridge University Press, Cambridge. ⟨Ch.6⟩

Searle, J. R. [1983] *Intentionality: An Essay in the Philosophy of Mind*, Cambridge University Press, Cambridge. ⟨Ch.6⟩

Segerberg, K. [1973] 'Two-Dimensional Modal Logic', *Journal of Philosophical Logic* **2**, 77–96. ⟨Ch.1⟩

Segerberg, K. [1982] *Classical Propositional Operators*, Clarendon Press, Oxford. ⟨Ch.7⟩

Seuren, P. A. M. [1977] 'Forme logique et forme semantique: Un argument contre M. Geach', *Logique et Analyse* **79**, 338–47. ⟨Ch.2⟩

Shoemaker, S. [1970] 'Persons and their Pasts', *American Philosophical Quarterly* **7**, 269–285. ⟨Ch.7⟩

Shoemaker, S. [1979] 'Identity, Properties, and Causality', pp. 321–342 in P. A. French *et al.* (eds.), *Midwest Studies in Philosophy IV. Studies in Metaphysics*, University of Minnesota Press, Minneapolis. ⟨Ch.7⟩

Smiley, T. [1996] 'Rejection', *Analysis* **56**, 1–9. ⟨Ch.1^{+}⟩

Smith, M. A. [1986] 'Peacocke on Red and Red′', *Synthese* **68**, 559–576. ⟨Ch.7⟩

Smith, M. A. [1987] 'The Humean Theory of Motivation', *Mind* **96**, 36–91. ⟨Ch.6⟩

Smith, M. A. [1988] 'Reason and Desire', *Proceedings of the Aristotelian Society* **88**, 243–58. ⟨Ch.6⟩

Smith, M. A. [1994] *The Moral Problem*, Blackwell, Oxford. ⟨Chs.2^{+},7⟩

Sober, E. [1990] 'Putting the Function Back Into Functionalism', pp. 97–106 in W. G. Lycan (ed.), *Mind and Cognition: A Reader*, Oxford: Basil Blackwell, (Excerpted from Sober, E., 'Panglossian Functionalism and the Philosophy of Mind', *Synthese* **64** (1985), 165–93.) ⟨Ch.6⟩

Sorensen, R. A. [1988] *Blindspots*, Clarendon Press, Oxford. ⟨Ch.6⟩

Sorensen, R. [1991] '"*P*, Therefore, *P*" Without Circularity', *Journal of Philosophy* **88**, 245–266. ⟨Ch.7⟩

Stalnaker, R. C. [1984] *Inquiry*, MIT Press, Cambridge, Mass. ⟨Ch.6⟩

Stenius, E. [1969] 'Mood and Language Game', pp. 251–271 in J. W. Davis, D. J. Hockney and W. K. Wilson (eds.), *Philosophical Logic*, Reidel, Dordrecht. ⟨Ch.6^{+}⟩

Stich, S. P. [1990] *The Fragmentation of Reason*, MIT Press, Cambridge, Mass. ⟨Ch.6⟩

Stocker, M. [1979] 'Desiring the Bad: An Essay in Moral Psychology', *Journal of Philosophy* **76**, 738–53. ⟨Ch.2⟩

Stocker, M. [1983] 'Psychic Feelings: Their Importance and Irreducibility', *Australasian Journal of Philosophy.* **61**, 5–26. ⟨Ch.3⟩

Stocker, M. [1987] 'Emotional Thoughts', *American Philosophical Quarterly* **24**, 59–69. ⟨Ch.3⟩

Teichmann, R. [1990] 'Actually', *Analysis* **50**, 16–19. ⟨Ch.1^{+}⟩

Thalberg, I. [1964] 'Emotion and Thought', *American Philosophical Quarterly* **1**, 45–55. ⟨Ch.3⟩

Unger, P. [1968] 'An Analysis of Factual Knowledge', *Journal of Philosophy* **65**, 157–70. ⟨Ch.6⟩

Urmson, J. 0. [1967] 'Memory and Imagination', *Mind* **76**, 83–91. ⟨Ch.6⟩

Velleman, J. D. [1992] 'The Guise of the Good', *Noûs* **26**, 1–26. ⟨Ch.6⟩

Wehmeier, K. F. [2000*a*] 'The Importance of Being in the Mood', unpublished paper, Faculteit der Wijsbegeerte, Rijksuniversiteit Leiden. ⟨Ch.1^{+}⟩

Wehmeier, K. F. [2000*b*] 'Modality, Mood, and Descriptions', later version of Wehmeier [2000*a*], Faculteit der Wijsbegeerte, Rijksuniversiteit Leiden. ⟨Ch.1^{+}⟩

White, A. R. [1975] *Modal Thinking*, Clarendon Press, Oxford. ⟨Ch.3⟩

Williams, B. A. O. [1966] 'Consistency and Realism', pp. 187–206 in Williams [1973]. ⟨Ch.6⟩

Williams, B. A. O. [1970] 'Deciding to Believe', pp. 136–51 in Williams [1973]. ⟨Ch.6⟩

Williams, B. A. O. [1973] *Problems of the Self*, Cambridge University Press. ⟨Ch.6⟩

Williams, B. A. O. [1978] *Descartes: The Project of Pure Enquiry,* Harvester. ⟨Ch.4⟩

Williams, J. N. [1979] 'Moore's Paradox: One or Two?', *Analysis* **39**, 141–2. ⟨Ch.6⟩

Williamson, T. [1992] 'Inexact Knowledge', *Mind* **101**, 217–242. ⟨Ch.4^{+}⟩

Williamson, T. [1995] 'Is Knowing a State of Mind?', *Mind* **104**, 533–565. ⟨Ch.0⟩

Wolniewicz, B. [1970] 'Four Notions of Independence', *Theoria* **36**, 161–164. ⟨Ch.7⟩

Woolhouse, R. S. [1975] 'Leibniz's Principle of Pre-Determinate History', *Studia Leibnitiana* **7**, 207–228. ⟨Ch.1⟩

Wright, C. [1992] *Truth and Objectivity*, Harvard University Press, Cambridge, Mass. ⟨Ch.7⟩

Yablo, S. [1993] 'Definitions — Consistent and Inconsistent', *Philosophical Studies* **72**, 147–175. ⟨Ch.7⟩

Zangwill, N. [1998] 'Direction of Fit and Normative Functionalism', *Philosophical Studies* **91**, 173–203. ⟨Ch.6^{+}⟩

Zolin, E. E. [2000] 'Embeddings of Propositional Monomodal Logics', *Logic Journal of the IGPL* **8**, 861–882. ⟨Ch.7^{+}⟩