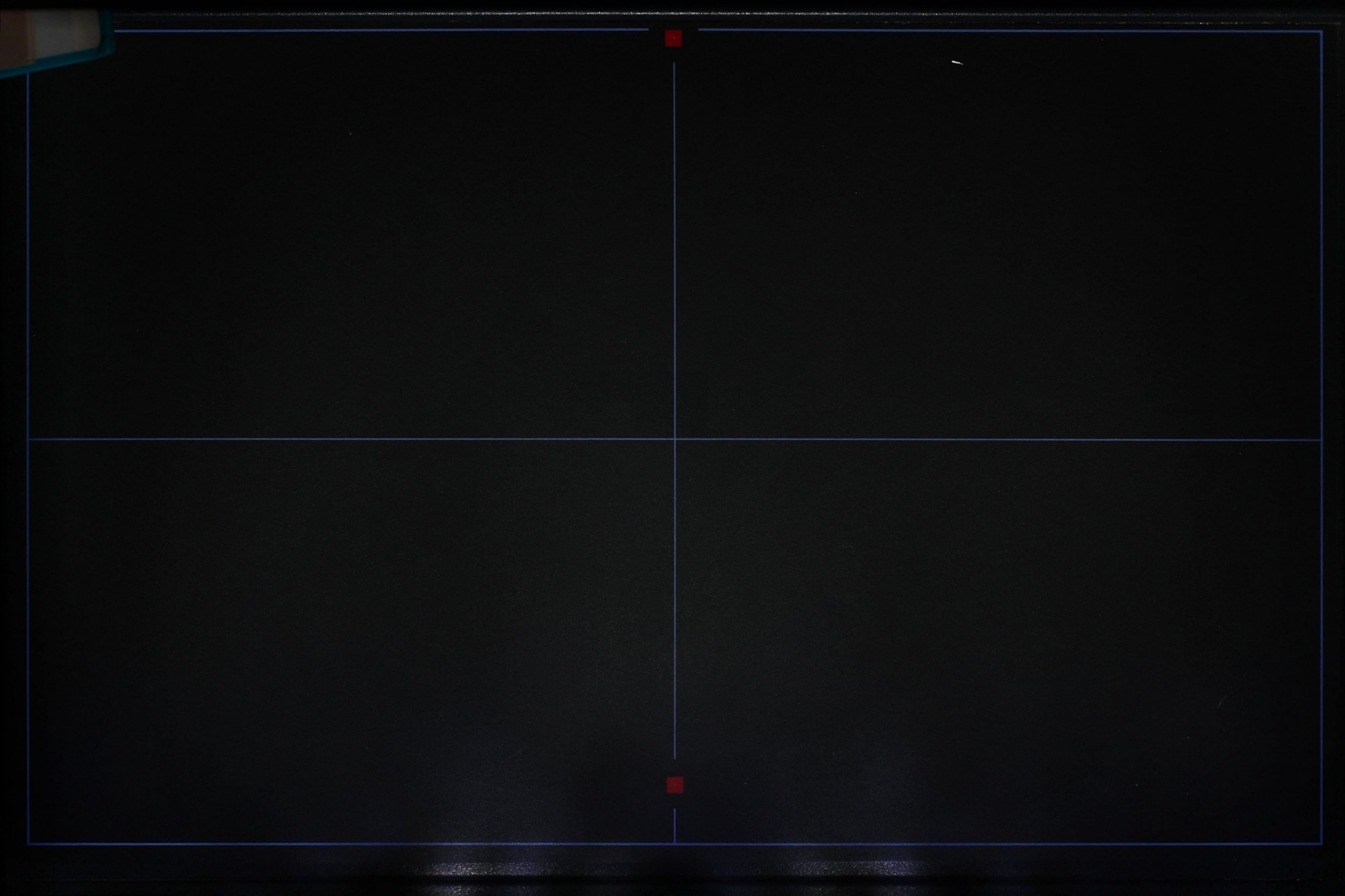

Preface

> A speech for the defence in a Paris murder trial, a road-safety slogan, Hobbes' political theory; each appeals to reason *of a kind*, but it remains an oblique and rhetorical kind. Each relies on comparisons rather than on direct statements, and none can override or supersede the conclusions of ethical reasoning proper. Nevertheless, just as slogans may do more for road safety than the mere recital of accident statistics, or of the evidence given at coroners' inquests, so the arguments of a Hobbes or a Bentham may be of greater practical effect than the assertion of genuinely ethical or political statements, however true and relevant these may be.
>
> Stephen Toulmin, *Reason in Ethics*, 1950.

The International Colloquium on Cognitive Science (ICCS), held in Donostia – San Sebastián every two years since 1989, has become one of the most important plazas for cognitive scientists in Europe to present the results of their research and to exchange ideas. The seventh edition, co-organized as usual by the Institute for Logic, Cognition, Language, and Information (ILCLI) and the Department of Logic and Philosophy of Science, both from the University of the Basque Country, took place from May 9 to 12, 2001, addressing the following main topics:

1. Truth: Epistemology and Logic.
2. Rationality in a Social Setting.
3. Music, Language, and Cognition.

4. The Order of Discourse: Logic, Pragmatics, and Rhetoric.

Around one hundred people from all over the world participated in ICCS-01. This volume contains the invited papers and tutorials presented there, as well as Martin Kusch's contributed paper, "Meaning Finistism and Truth," that merited the IBERDROLA Best Paper Award, given in 2001 for the third time. Let us briefly describe the papers one by one.

The first chapter, John Biro's "Cognitive Science and David Hume's Science of the Mind," is the only chapter in this volume devoted to the history of cognitive science, in commemoration of the 225th anniversary of the death of the great Scottish philosopher. Contemporary cognitive science is coupled with Hume's theory of mind. But, among all the figures of the history of philosophy that could be susceptible of study for the history of cognitive science, why does he choose Hume (besides the commemorative reason)? Was not he "academically" sceptic about any possibility of knowledge? So, why should he be relevant to cognitive scientists today? Biro argues that, far for being sceptical about the possibility of a science of the mind, Hume was maybe the first to consciously undertake the project of constructing one. Moreover, Hume anticipated some of the goals and some of the results of contemporary cognitive scientists. According to Biro, we should become aware of the fact that "it is an irony that the cognitive science of our day is sometimes explicitly contrasted with Hume's science of man, rather than being recognized as the latter's descendant," and that Hume is "the philosophical father of cognitive science".

The second chapter, Scott Soames' "Truth and Meaning: The Role of Truth in the Semantics of Propositional Attitude Ascriptions," focuses on the shortcomings that some semantic theories present when addressing the meaning of propositional attitude reports. The semantic theories taken under consideration are contemporary truth-conditional theories. According to Soames, these theories that try to specify the meanings of propositional attitude ascriptions, as they stand, are doomed to failure. However, he finds a way to extend these Davidsonian theories making them able to provide some information about the meanings of propositional attitude ascriptions. Nonetheless, he also finds that there are serious difficulties in extracting substantive testable consequences from even these extended theories.

The third chapter, Timothy Williamson's "Truth and Borderline Cases," adresses the relation between the derivability of the principle of bivalence, challenged by borderline cases for vague expressions, and some principles about truth and falsity. Starting with the latter, he replaces the usual disquotational characterization of truth by one that makes room for contextual variation in what is said by a sentence. He then explores the

implications this characterization has for bivalence. Bivalence requires a principle of uniformity in what a given sentence says in a given context, and even though borderline cases for vague expressions may appear to provide reasons to deny such a uniformity, Williamson argues that it is not so: "The apparent loophole in the classical case for bivalence is ... merely apparent", he concludes.

The fourth chapter, Martin Kusch's "Meaning Finitism and Truth," reconstructs and defends a theory of linguistic meaning developed since the late seventies by Barry Barnes and David Bloor that has been called "Meaning Finitism", exploring at the same time some of its consequences for our understanding of truth. Meaning finitism views meanings as continuously made and re-made by language users. Like Barnes and Bloor, Kusch combines meaning finitism with meaning collectivism, and, perhaps more explicitly than the formers, he defends one of the most important philosophical consequences of that "strong" position: relativism concerning truth. This paper, which closes the bunch of papers connected with truth, got the IBERDROLA Best Paper Award at ICCS-01.

Manuel Liz's "Subjective Experience and the External World: Gaps and Bridges" is the first of the three papers in the volume, chapter 5 to 7, devoted to the philosophy of cognitive science. He attempts to bridge what he calls the ontological and the epistemological gaps between our subjective experiences and the external world. The epistemological gap arises from the problem of knowing the external world given that "it always seems possible to conceive our subjective experience as remaining the same through many kinds of changes in the external world"; the ontological gap arises when trying to understand how the external world can produce subjective experiences given the fact that "it seems always possible to conceive the external world —especially our brains and bodies— as remaining the same through many kinds of changes in our subjective experience." According to Liz, taking seriously the ontological gap entails being able to bridge the epistemological one and the strategy he proposes to bridge the latter could also be successfully applied to the former.

Chapter 6, Marcelo Sabatés' "The Explanatory Relevance of Psychological Properties," studies epiphenomenalist approaches to mental properties and explores their consequences for the explanatoriness of such properties, and "perhaps more generally," tells Sabatés, "of other special science properties." According to many critics, epiphenomenalist approaches to mental properties end up concluding that they are explanatory irrelevant or even that they are not real properties. Sabatés attempts to show that the indirect and direct arguments going from epiphenomenalism to lack of explanatory relevance, though compelling, are not conclusive, and he certainly shows which should be the main concerns on the issue. As he

claims, this is a necessary first step "towards a theory allowing us to *realistically* explain with *real* epiphenomena."

Chapter 7, "Epistemology and Cognitive Theorizing" by Jesús Ezquerro and Fernando Martínez-Manrique, treats the mutual influence of studies in cognitive science, on the one hand, and theoretical projects in epistemology. The paper is divided into two parts. The first part traces a path through normative issues in the philosophy and history of science that grounds the view that the only plausible understanding of epistemology comes from a naturalist position. The second part of the paper studies the impact that a putative "folk epistemology" capacity (i.e., the natural ability to evaluate epistemic situations) could have on epistemology. This paper closes the series of papers on the philosophy of cognitive science.

Chapters 8 and 9 are concerned with the relation of music to cognition. The question made by Peter Kivy in the title of his paper, "Music, Language and Cognition: Which One Doesn't Belong," is clearly answered by the author: *language* doesn't belong. Music is not language. The first part of the paper explores the analogy between music and language. A wrong understanding of this analogy has caused pretty much confussion. Kivy insists that, though language-like in certain respects, music "is not language: it is not a language or part of a language." On the other hand, he claims that music and cognition belong together. More specifically, in the sencond part of the paper, he argues that the enjoyment and appreciation of music are concept-laden (but content-less) activities and, thus, deeply involved with cognitive processes.

While Kivy leaves it explicitly aside, Lairs Addis is primarily concerned with unconscious (or innate) knowledge in relation with music in his "Music and Knowledge." He adresses two questions that seem to involve some kind of unconscious knowledge. The first concerns knowledge "that we *bring to* music," what Addis calls "knowledge-*for*" listening to music; the second concerns knowledge that we, humans, get from music ("knowledge-*from*" listening to music). These two notions are also analysed in the light of a better-known philosophical distinction among knowledge-*that*, knowledge-*how*, and knowledge-*of*, as well as attending to music psychology with the aim of clarifying why music has the importance it has for human beings.

Chapter 10, "Event Coreference and Discourse Relations" by Laurence Danlos and Bertrand Gaiffe, is a paper on discourse semantics within the framework of the Segmented Discourse Representation Theory (SDRT). In particular, they focus on event coreference, which is shown to be the keystone for the correct analysis of four types of dicourses: particularizing, generalizing discourses, on the one hand, and explanation and result discourses, on the other. Their analysis allow them to present and explain some unusual linguistic phenomena (such as coreference between

existentially quantified elements, or asymmetrical behavior of explanation and result discourses), as well as to shed new light on the discourse relations involved.

The last two chapters address the issue of rationality in a social setting. Chapter 11, "Rationality in Context" by J. Francisco Morales, approaches the problem from the perspective of social psychology. To begin with, he discusses the very concept of rationality for undermining the alleged irrationality in some cases of social behavior. In particular, he is concerned with the possibility of reconciling rational choice theory with the influence of culture in human decision-making. With this purpose, he discusses a couple of cross-cultural studies on conflict management strategies and on the obedience to authorities, respectively. The author concludes that "paying attention to [the context] is a necessary step to move beyond the rational-irrational dichotomy to a more ... adequate conceptualization of human behavior."

Chapter 12, "Individual and Collective Rationality in a Social Framework" by Luis A. Pérez Miranda, presents and explores the very roots of rational choice theory and game theory. The first part of the paper is devoted to the discussion of the main problems of individual decision-making from the viewpoint of instrumental rationality. The second part examines the efforts to extend rational choice theory when there is more than one individual in the decision-making, that is, when we have a social or even a collective decision to be made. Is game theory suitable for explaining phenomena like coordination, cooperation, or the emergence of norms? Pérez Miranda argues that it needs a deep revision that, on the one hand, takes into account the role played by structural rationality and, on the other hand, gets "less idealised models sensitive to experimental data, or ... models that find their scientific support in accurate computer simulation."

As usual, this collection of papers selected from the ICCS-01 reflects, not only the diversity of the interdisciplinary origins and standpoints of the researchers who attended the Colloquium, but also serves as a proof of the richness, fruitfulness, and diversity of research in Cognitive Science today.

Kepa Korta and Jesus M. Larrazabal

Chapter 1

COGNITIVE SCIENCE AND DAVID HUME'S SCIENCE OF THE MIND

John Biro
University of Florida

What reason is there for coupling one of the newest branches of modern science with a long-dead philosopher, however august? And why with Hume, of all people? Was he not a sceptic, famous for questioning the possibility of any, and, thus, *a fortiori*, scientific, knowledge? Interesting as his reasons for such doubts may be to other philosophers, can they not be safely ignored by scientists, those actually engaged in pursuing and, surely, gaining, the knowledge he said was unattainable? But, then, what else has he to say to the latter? In what follows, I shall argue that far from being sceptical about the possibility of a science of the mind, Hume was perhaps the first to embark self-consciously on the project of constructing one. Furthermore, some of the goals and some of the results of his project anticipate in interesting and rarely noticed ways those of recent cognitive science and its philosophy.

1.

For almost two centuries after its appearance, Hume's philosophy was construed as essentially, perhaps entirely, negative.[1] Its goal was seen as

[1] With the possible exception of some of the practical philosophy. But even where these were read at all, the emphasis was typically on the sceptical or destructive passages (often taken

K. Korta and J. M. Larrazabal (eds.), Truth, Rationality, Cognition, and Music, 1-20.

showing that the unavoidable outcome of following through on some seemingly compelling empiricist principles was complete scepticism. That Hume should be read in this way over a stretch of time during which epistemological questions in general, and the sceptical threat in particular, dominated philosophy is hardly surprising. It is a common-place that with Descartes these questions had moved to center stage in a way that would have seemed, until then, quite unnatural to most philosophers since the ancient sceptics. Thus it is only to be expected that Hume should have been seen as engaged in the general post-Cartesian enterprize of assessing—in his case, apparently negatively—the prospects for human knowledge. It is also true that the first book of *A Treatise of Human Nature*, and *An Enquiry into Human Understanding*, the two parts of his *ôuvre* on which most interpreters have focussed until recently, contain a barrage of arguments that certainly seem to have been designed with a sceptical intent.

Thus Hume came to be seen as the arch-sceptic of the modern age. That such extreme scepticism is straightforwardly incompatible with the project Hume explicitly announced in the introduction to the very work which was interpreted as its expression was either not noticed or dismissed by the simple expedient of not taking him at his word.

In the Introduction to the *Treatise*, in what is one of the best-known passages in all his writings, Hume outlines a project it is hard to see a sceptic taking seriously: that of establishing a new science. He calls it "the science of man," and describes it as an investigation into "the nature of the ideas we employ, and of the operations we perform in our reasonings." Such an investigation is needed if any progress is to be made in any of the other sciences, since "all the sciences have a relation ... to human nature," and "lie under the cognizance of men, and are judged of by their powers and faculties." Thus, understanding the human mind is the key to understanding anything else:[2]

> There is no question of importance whose decision is not compriz'd in the science of man; and there is none, which can be decided with any certainty, before we become acquainted with that science. In pretending therefore to explain the principles of human nature, we in effect propose a compleat system of the sciences, built on a foundation almost entirely new, and the only one upon which they can stand with any security. [T xx]

---

out of context), such as the notorious one about the impossibility of deriving an 'ought' from and 'is'. (*A Treatise of Human Nature*, pp. 469-70. References to this work in what follows will take the form 'T pp.')

[2] For a more thorough discussion of this claim than is possible here, see Biro (1993). See also Monteiro.

Explaining the principles of human nature involves "examining the Mind... to discover its most secret Springs and Principles." Though these may "lie very deep and abstruse", the new "experimental method of Reasoning", modelled after that used with such spectacular success by Boyle, Newton and others in what might be called the new science of matter, holds out the hope of results no less far-reaching in the domain of "moral subjects".[3] The "total alteration" and "revolution" Hume says this new science brings to the intellectual scene consists in becoming an "anatomist" of human nature. The secret principles that "anatomizing human nature in a regular manner" enables us to discover need not, however, be *occult*; they are accessible to an investigator using the right method, a method that contrasts with that of the metaphysician. The latter's is mere arm-chair speculation, yielding "many chimerical systems" in which "hypotheses [are] embrac'd merely for being specious and agreeable," rather than because they can be shown to be true, or even probably true, by some objective method.

But what method? As an empiricist, Hume can allow no source of knowledge other than experience and recognizes no claim to knowledge not based on experience: "The only solid foundation we can give to this science... must be laid on experience and observation." (T xx). No science "can go beyond experience, or establish any principles which are not founded on that authority." Thus the right method calls for "exact experiments," aiming at "render[ing] all our principles as universal as possible". This, in turn, requires "tracing up our experiments to the utmost, and explaining all effects from the simplest and fewest causes." (ibid.)

The kind of speculative metaphysics with which Hume contrasts his new science poses questions to which no principles derived from experience can yield an answer. It pretends to offer deeper and more general principles based on pure reason and claims that only these can lead to genuine understanding. (An example of this attitude is the refusal of Leibniz and the Cartesians to admit that Newton had really *explained* anything.) Hume is adamant that in our new science we should not pretend to such "deeper" explanations. In particular, we should not give in to the temptation to try to explain *why* the principles of human nature our new method has revealed are the way they are: "we can give no reason for our most general and refined principles beside our experience of their reality." (T xxii) When he

[3] The relative importance of different influences on Hume from the direction of "natural philosophy" is a matter of controversy. The best-known interpretations (Noxon, Capaldi) assign the dominant role to Newton's influence; others (e.g., Barfoot) argue that Boyle's was greater or (Wright) that that of the mechanical philosophy, deriving from Descartes and Malebranche, perhaps greater still. Yet others (Jones) trace Hume's philosophy to quite other roots in classical thought. My discussion does not require taking sides on this question.

introduces his famous principles of association —the three "universal principles" which "guide" the operations of the imagination in uniting our ideas— Hume emphasizes that their reality requires no special proof beyond recognizing that their "effects are every where conspicuous." (T 13) He follows this claim immediately by reminding us that their "causes ... are mostly unknown, and must be resolv'd into *original* qualities of human nature, which I pretend not to explain" (*ibid.*, Hume's emphasis). It would be a mistake, however, to complain about this "impossibility of explaining ultimate principles" in the science of man, as it is "a defect common to it with all the other sciences, and all the arts, in which we can employ ourselves." (T xxii) Being sceptical about the possibility of answering the "deeper" questions of the metaphysician need not make one sceptical about the possibility of scientific knowledge.

Thus while there is certainly a sense in which Hume can be, as he so often is, said to be a sceptic, his scepticism should be understood as one about pretended supra-scientific metaphysical knowledge, rather than about scientific knowledge. It is this kind of scepticism that separates him from other philosophers before him, who conceived of philosophy as able to go beyond "mere" scientific knowledge and so able to provide a deeper and more certain knowledge of reality.[4] His extensive and devastating criticisms of the rationalists' attempts to deal with the threat of a scepticism that denies the possibility of scientific, indeed, any, knowledge, seemed to his contemporaries, and to many since, evidence that he shared their pre-occupation with that threat. Yet Hume is quite explicit in disclaiming such an interest and tells us clearly, in a variety of ways and contexts, that the main aim of his enquiries is something very different. An example is his admonition, in the opening paragraph of the section in the *Treatise* entitled "Of scepticism with regard to the senses," not to be concerned with the usual sceptical question about the existence of the external world:

> We may well ask, *What causes induce us to believe in the existence of body?* but 'tis in vain to ask, *Whether there be body or not?* That is a point, which we must take for granted in all our reasonings. (T 187)

That such an injunction should appear in this very section, nominally concerned with scepticism, is surely not an accident and should clinch the case that whatever Hume is doing, it is neither pressing, nor looking for—and failing to find—an answer, to the radical sceptical challenge. He tells us

[4] There are, in fact, other ways in which Hume can justly be said to be a sceptic, and recognizing them is important for a completely adequate interpretation of his overall philosophy, of which his science of the mind is only a part. But these are matters beyond my present topic. (For helpful discussions, see Fogelin (1985) and Fogelin (1993).)

explicitly what he *is* doing: "The subject, then, of our present enquiry is concerning the *causes* which induce us to believe in the existence of body... " (ibid., Hume's emphasis).

In sum, Hume *is* sceptical about various philosophical attempts at justifying our beliefs, especially when it comes to the most basic of these: in the external world, in the identity of our person, in causal connections, and the like, beliefs Hume thinks even a sceptic cannot seriously reject or live without. He often insists that it is just as well that nature has made sure that, in spite of all philosophy, we take these for granted, as without them "human nature would ... go to ruin" (T 225). But this recognition of our unreflective, instinctive and unavoidable acceptance of certain basic beliefs must not be confused with claiming to have a rational justification of them. It is philosophers' claims to the latter that are the targets of Hume's sceptical arguments, as are their pretensions to knowledge about the source of the principles a scientist of the mind can discover, and about the reasons why these principles are what they are.

I have noted that the method Hume urges for his new science consists of observation and experiment as much as does the method of natural philosophy. Yet it differs from the latter in some important respects. I shall discuss some of these differences later. But one must be noted right away: the impossibility in the new science of making experiments "purposely, with premeditation." Here Hume is thinking of the controlled experiments of the natural sciences, not of the thought experiments familiar in philosophy, which he himself often uses. In our new science we must "glean up our experiments... from a cautious observation of human life, and take them as they appear in the common course of the world, by men's behaviour in company, in affairs, and in their pleasures." (T xxiii) In spite of this, though, the science of man need "not be inferior in certainty... to any other of human comprehension." (ibid.)

2.

Hume, then, is quite explicit about the nature, status, and proper method, of the project he is undertaking. Yet his declarations have had remarkably little effect on the interpretation of his writings by champions and critics alike, from his day to ours. It is only very recently that some have begun to take him at his word and to see him as engaged in an inquiry at least continuous with what we today think of as the scientific study of the mind. But such recognition of continuity is still by no means universal, and it is an irony that the cognitive science of our day is sometimes explicitly contrasted with

Hume's science of man, rather than being recognized as the latter's descendant.[5]

One reason for this is the difference just noted in the two sciences' respective conceptions of observation and experiment. (I shall return to this.) Another, however, is the failure to appreciate the solutions he offers to the sceptical challenges he poses, solutions that contain, albeit in a form unfamiliar to many in recent cognitive science, at least the germs of some of the latter's central insights. I shall discuss examples of this failure later. First, however, let me offer a sketch of the main features of the picture of the mind that Hume, the anatomist, develops.

Hume's general answer to questions about how we come to have the various beliefs we have is that they are the product of a *non*-rational faculty. He labels this faculty variously as "instinct", "habit" or "custom", or "the imagination;" it is defined by a certain "propensity" to form ideas and beliefs. Some subtle differences behind this varying terminology notwithstanding, the main contrast is with reason, the faculty whose standards and operations some philosophers think can serve to provide an answer to the sceptics's challenge.[6]

The raw materials on which this non-rational faculty works and from which all mental life is constructed are impressions and their "faint copies," ideas, species both of the genus, perception: "All the perceptions of the human mind resolve themselves into two distinct kinds, which I shall call IMPRESSIONS and IDEAS." (T 1). Many of the most sceptical-sounding passages of the *Treatise* and of the first *Enquiry* are devoted to showing that our stock of these materials is more limited than philosophers have supposed. Hume shows us again and again that the impressions from which some putative idea posited by the metaphysician would have to derive are just not to be found in experience. What Hume calls his "first principle"—often labelled "the copy principle"—requires that all simple ideas be traceable back to simple impressions which they exactly resemble save for

[5] For an example of this, see Fodor (a). Contrast, however, Fodor (b), much more in sympathy with the view advocated here.

[6] Care has to be exercised with Hume's terminology here: while the chief contrast is between reason and the imagination, the latter is itself Janus-faced: "the general and more establish'd properties of the imagination," are to be sharply distinguished from "the trivial suggestion[s] of the fancy" (T 267). Which of these aspects of the imagination dominates in one's empirical reasoning ("reasonings concerning matters of fact") determines, for Hume, how reasonable one is. This makes sense only if we recognize that he uses terms such as 'reason' and 'reasoning' in two very different ways. In one sense, these terms refer to the kind of abstract reflective operation involved in the arbitrary comparison of ideas involved in what he calls "philosophical relations." In the other, they are labels for the kind of automatic, non-reflective, transitions that count as "natural relations," most important among them, causation. (For further discussion of these, see below.)

their lesser "forcefulness" or "vivacity." ("...all our simple ideas in their first appearance are deriv'd from simple impressions, which are correspondent to them, and which they exactly represent." (T 4)). But he does not deny the obvious, though remarkable, fact that from the rather limited stock of impressions that come my way, I am able to construct an edifice of beliefs that go far beyond those impressions and the ideas copied from them. First, my complex ideas are not confined to the complex impressions I have actually had: I can combine simple impressions in novel ways, into new complex ideas. (These, often called by Hume "fictions," may, but need not, give rise to belief, and they often do not, as with ideas of fiction in the usual sense. For Hume, though, not all "fictions" are fictions. See below.) Second, the course of my experience, the various patterns and regularities among the perceptions that make it up, is exploited by my mind in forming the beliefs I do. In both these ways, the mind must be conceived as essentially active: it is what it does with what it gets that matters, and it is this that Hume's science is designed to describe.

According to that science, the mind is led from one idea to another by three "principles of association": resemblance, contiguity, and cause and effect. These principles involve the mind's "taking notice" of certain properties of, and regularities among, its perceptions. Such taking notice need not be, and typically is not, conscious. What matters is that these properties and regularities be detected by the mind in a way that makes a difference to its subsequent operations and contents. Were it not for this active contribution on the mind's part, the mere presence of such properties and regularities would not be sufficient to explain the combinations and transitions among our ideas that actually occur, nor the genesis of the beliefs we actually form.

The remarkable regularities in the transitions we make from idea to idea and from (some) ideas to beliefs are the result of certain characteristics of the imagination, an ever-active (and sometimes over-active) non-rational faculty, the story of whose workings in large part constitutes Hume's scientific account of human nature. "Custom" or "habit" are Hume's usual short labels for these characteristics, among them a certain inertia that plays a role in Hume's explanations of some of the most remarkable, though often not even noticed, facts about the mind. These include the fact that in the absence of impressions from which the corresponding ideas could have been copied, we nonetheless come to believe that there are bodies and that we are the same person at one time as at another, even that we can "extend our identity beyond our memory" (T 262).

This inertia is one of the most important characteristics of the mind: "the imagination, when set into any train of thinking, is apt to continue, even when its object fails it, and like a galley put in motion by the oars, carries on

in its course without any new impulse." (T 198) In his treatment of our ideas of space and time and of mathematical and geometrical reasoning, Hume appeals to it to explain how we generate an "imaginary standard of equality," notions of "perfection beyond what [our] faculties can judge of," and of "correction[s] beyond what we have instruments and art to make." (ibid.) Thus the "useless and incomprehensible fictions" of the mathematicians, who give exact definitions and demonstrations that have no experiential application. In this case, Hume's purpose is to expose these fictions as "absurd" (T 51-2). His recommendation is to resist the tendency and thus avoid the absurdity.[7] But this inertial tendency, automatic and non-reflective, is not limited to mathematical reasoning. It is ubiquitous: nothing, says Hume, is "more usual, than for the mind to proceed after this manner with any action, even after the reason has ceas'd, which first determined it to begin" (T 48). In the discussion of our belief in the existence of external objects, the same tendency is invoked in the interest of quite a different goal: that of explaining how we *cannot* avoid forming our "opinion of the continu'd existence of body:"

> Objects have a certain coherence even as they appear to our senses; but this coherence is much greater and more uniform, if we suppose the object to have a continu'd existence; and as the mind is once in the train of observing an uniformity among objects, it naturally continues, till it renders the uniformity as compleat as possible. (T 198)

Hume distinguishes between "principles which are permanent, irresistable, and universal" and those "which are changeable, weak, and irregular" (T 225). This distinction is essential to the double use by Hume of the inertial, extrapolating, tendency of the mind. When the tendency is guided by principles of the first sort, as it is in the formation of our fundamental common-sense beliefs, a recognition that this is so is what constitutes what Hume calls a "sceptical solution" to the sceptic's doubts, whether about the existence of an external world, about genuine causal connexions, or about personal identity. While Hume sometimes uses the term "fiction' to label a

[7] Some find Part II of the first book of the *Treatise* one of the most perplexing of the whole work, and its purposes are much disputed (See, for example, Anderson, Fogelin). But it is worth noting that at least at times Hume seems to be interested there in making an *anti*-sceptical argument. Having argued that the idea of extension as infinitely divisible is incoherent, he says: "Now 'tis certain we have an idea of extension; for otherwise why do we talk and reason concerning it? ... Here then is an idea of extension, which consist of parts or inferior ideas, that are perfectly indivisible: consequently this idea implies no contradiction: consequently 'tis possible for extension really to exist conformable to it: and consequently all the arguments employ'd against the possibility of mathematical points are mere scholastick quibbles, and unworthy of our attention" (T 32)

fundamental natural belief produced by this property of the mind, we must be careful not to be mislead by this into thinking of such a belief as somehow fanciful and arbitrary. Fictions of this sort are not optional: they are forced on us by our nature. But we can distinguish such fictions from those resulting from philosophical speculation floating free of common sense, such as, for example, the "feining" of a spiritual substance to explain the identity of a self over time. (I shall discuss Hume's own explanation of that identity later.)

In these cases of what we may call natural fictions, the mind's extrapolating tendency operates "in such an insensible manner as never to be taken notice of," and "the imagination can draw inferences from past experience, without reflecting on it; much more without forming any principle concerning it, or reasoning upon that principle." (T 102)

Hume adds that this tendency "may even in some measure be unknown to us." (ibid.) It is important to see that by this he means only that we have no introspective access to the processes in question. In forming a belief that one event is the cause of another, for example, I obviously do not consciously recall the previous instances of constant conjunction upon which my inference is based, using them as evidence for the inference. Not only does "the custom operate[s] before we have time for reflection," but "I am never conscious of any such operation," and in "giv[ing] preference to one set of arguments above another I do nothing but decide from my *feeling* concerning the superiority of their influence" (T 103, my emphasis). Thus it is that "all probable [that is, causal] reasoning is nothing but a species of sensation." (ibid.)

This distinction between reason (a reflective faculty purporting to make inferences on the basis of evidence) and the imagination (a non-reflective faculty that naturally moves from experience to belief) is fundamental to Hume's anatomy of the mind. To quote again from his discussion of our belief in external objects:

> "...our reason neither does, nor is it possible it ever shou'd, upon any supposition, give us an assurance of the continu'd and distinct existence of body. That opinion must be entirely owing to the IMAGINATION: which must now be the subject of our enquiry." (T 193)

Nor is this particular kind of belief unique in this respect: quite generally, "belief is more properly an act of the sensitive, than of the cogitative part of our natures" (T 183). When it comes to our most general and most fundamental beliefs (such as those in the existence of an external world, in our own identity, in causal relations), these are, therefore, quite impervious to the influence of reason, which can neither ground nor destroy them. "Cogitative part," means here our faculty of theoretical reasoning, at work

when we construct demonstrations and philosophical arguments. There is, however, another sense of 'reasoning,', applicable to some of the natural and instinctive transitions we make from one perception one belief to another. Thus, for example, we are engaged in reasoning when we make a causal inference; indeed, that is what we primarily mean by 'reasoning' in ordinary, non-theoretical, contexts. ("... this inference is not only a true species of reasoning, but the strongest of all others" (T 97fn)). Hume calls this kind of reasoning "probable" or "experimental" and insists that we share it with infants, "nay, even brute beasts"—who presumably do not "cogitate" (*Enquiry,* 4.2). It is this latter kind of reasoning "on which the whole conduct of life depends, [and it] is nothing but a species of instinct or mechanical power, that acts in us unknown to ourselves" (ibid.).

3.

Hume's recommendation is to replace endless and fruitless "cogitating," in an attempt to give philosophical justifications of our beliefs, by the search for a scientific explanation of their origin. Doing so is what the "sceptical solution" of the sceptical challenge consists in (*Enquiry*, 5). It is to give up being a "metaphysician" and to become a scientist—an "anatomist"—of the mind. The resemblance between this recommendation and the so-called "naturalizing" programs common in recent philosophy of mind and epistemology is unmistakable. In these, too, the leading idea is to abandon a bankrupt a *priori* method in favour of an empirical, descriptive, one that holds out the promise of genuine progress. Many epistemologists have, in recent years, come to feel that arm-chair conceptual analysis is unlikely to tell us much about the real nature of human knowledge. (The uninspiring history of the so-called Gettier problem, involving increasingly arcane and artificial counter-examples to ever more byzantine definitions of knowledge is often taken as evidence of this.) Philosophers of mind, too, interested in understanding reasoning, perception, memory, language, and other mental phenomena, increasingly look to the new discipline (or constellation of disciplines) called "cognitive science" for illumination, rather than to inconclusive abstract philosophical analysis and argument.[8]

I have given examples elsewhere of how Hume's treatment of some of our fundamental concepts and natural beliefs anticipate, or are echoed in, cognitive science.[9] I want here to focus on the question I flagged earlier,

[8] For useful surveys of naturalization programmes in epistemology and in the philosophy of mind, see Goldman and Kornblith.

[9] Biro (d) and (e).

concerning the methods Hume uses in his science of the mind, both the one we have seen he recommends and a very different one we often find him actually using. I want to ask whether these different methods are compatible and, if so, how they are related to each other and to the methods of cognitive science. While the one he recommends is that of systematic every-day observation of others, the one that dominates Book I of the *Treatise* is one of first-person introspection. Neither is a method favoured by cognitive science, and it is worth asking three questions. First, are Hume's two methods compatible with one another? Second, in what ways do they differ from those of cognitive science? Third, do the methods Hume actually uses constitue a virtue or a limitation of his science?

We have already noted one striking difference between the method Hume recommends for his science and that of the "natural philosophy" on which, in other respects, that science is modelled: the difference between the use of controlled experiments and the observation of human behaviour as it occurs in the wild, as it were. (This difference is both reflected in, and obscured by, the differences between the ways Hume and we use the term 'experiment'.) In this regard, not only what we today call the "hard" sciences, but even much of modern social science (not all) fall on the non-Humean side of the divide. But in another way, a great deal of social science, by virtue of its being *social*, shares Hume's assumption that human nature cannot be studied in one human being, for it is found, indeed, in some of its most important aspects, is constituted, in the various relations people stand to each other. This aspect of Hume's science looms largest in the picture of the mind —the self, as he often calls it—found in the later parts of the *Treatise* and in other works, and it is in sharp contrast with the method prominent in Book I, with its more individualistic, atomistic picture. I shall return to it shortly. But first we need to note that the picture of the self drawn in Book I seems not at all to be based in the sort of "experiments" Hume recommends for the science of man. Rather, it involves a method common in empiricist philosophers before him and since, of looking to the meager materials provided by one's experience, from which one's ideas could be constructed or to which they could be reduced. The method this "looking" requires is one of observation, all right, but not of the things—"men's behaviour in company" and the like—recommended for the scientist of human nature, but of the sole, and curious, items directly accessible to consciousness: impressions and ideas. 'Observation' thus means, it seems, nothing more or less than *introspection*. No wonder, then, that most interpreters read the first book of the *Treatise* (and, since, as already noted, often not much else was read, the whole of Hume's philosophy) as part introspective psychology, part

phenomenalistic reduction and part phenomenology—not to mention the scepticism![10].

The details of Hume's criticism of the substantial theory of the self favoured by rationalists, which occupies much of the section on personal identity in Book I, are too well known to require recounting. It cannot fail to be obvious to even the most superficial reader that those criticisms are based primarily on the evidence of introspection.[11] When Hume talks of "enter[ing] most intimately into what I call *myself*" (T 252) and of looking and not finding a perception of an enduring self distinct from "its" ever-changing impressions and ideas—which, if the substantial theory were true, he *should* find—it is difficult to see how any other method could be involved. Less obvious, and usually less noticed, is the pivotal role played by something like introspection in the positive account of the origin of one's idea of oneself as the same ("imperfectly" identical) over time that at least some commentators see him going on to develop.[12] This positive role of introspection is due to the importance of memory as the foundation of the natural relation, causation, the relation that ties together the diverse perceptions that constitute the complex object whose idea the idea of the self is. Causation is "to be consider'd ... as the source of personal identity," and "[H]ad we no memory, we never shou'd have any notion of causation, nor consequently of that chain of causes and effects which constitute our self or person."(T 262) Thus Hume's somewhat surprising conclusion: "...the memory not only discovers [the self's] identity, but also contributes to its production" (T 261)

But, of course, 'memory' in this context can mean only a special kind of perception, distinguishable from others by un-mediated internal observation, that is, phenomenologically. Perceptions of this sort refer to a time (putatively) earlier and to an experience (apparently) had at that earlier time, but do not imply anything about the actual existence of that time and

[10] The first of these readings is often a critical one, contrasting such psychology with real philosophy. (For a recent example, see Bennett). A combination of the second and the last was orthodoxy among analytic philosophers until recently; the third, again combined with elements of the sceptical interpretation, has been common in Europe (see, for example, Davie and Connon; for an attenuated phenomenologist reading, see Livingston); the last is too ubiquitous and too well-known to need documentation historically: it tempts just about everybody from Read until Kemp Smith (and even some today – see Stove, Fogelin).

[11] Primarily, though not exclusively. They also involve examining the way we use the language of identity and thereby providing an analysis of the concept of identity, resulting in the all-important distinction between perfect and imperfect identity. For detailed discussion, see Biro (a) and (b).

[12] Some do not see him as doing that at all (e.g., Penelhum); and it goes without saying that even those who do, hardly agree on what that account is. (Geach, Bricke, Robison, Traiger, McIntyre, Biro (a) and (b).)

that experience. So, the essence of the work memory does in tying together the perceptions constituting a self cannot lie in anything else than its phenomenologically, introspectively, available properties. The properties in question are obviously available in this way only to one mind or self, the one whose identity they account for.[13] Thus, even when Hume speculates on what one might find, were one able to "look into the breast of another," his very question underlines the essentially first-person character of both the information sought and the only possible method of seeking it.

The contrast with the overtly third-person, external, "observations" in which pursuing the new experimental method is said by Hume to consist, could hardly be greater. And this divergence re-appears at many places besides these discussions of the self, where it is easy to see that the nature of the problem Hume is struggling with—explaining the origin of one's idea of one's *own* identity[14]—demands a first-person approach. But in many other contexts, too, in which memory plays a central explanatory role—in grounding our ideas of time, of causation, of experience itself, as well as our belief in the existence of bodies—the first-person, introspective, method dominates Hume's discussion. I shall not give evidence for this general claim here, having done so elsewhere.[15] But to give just one example: to have an adequate (Hume calls it "just") idea of time, not only must we be presented with a succession of impressions such as Hume's "five notes play'd on a flute," but we must appreciate that succession (represent it to ourselves, as Kant, or some late-twentieth-century cognitivists might put it) *as* a succession. To quote William James on a closely related point: "A succession of feelings, in and of itself, is not a feeling of succession. And since, to our successive feelings, a feeling of their succession is added, that must be treated as an additional fact requiring its own elucidation."

The sort of fact involved here is, for both James and Hume, a phenomenological fact, and an appeal to facts of this sort seems to play a crucial role in many of the latter's explanations of how we come to have our fundamental ideas and beliefs. Such facts are of a very different sort than, and our access to them, and with it their epistemological status, is also very different from, those with which the science of man is supposed to be concerned. They are different also from the putative explanatory facts that science posits: we explain what we experience by postulating theoretical entities and processes that are not themselves part of that experience — explanatory entities to which we have no un-mediated introspective access. Chief among these are those propensities of the mind to which Hume

[13] For how to dispell the air of circularity about this, see Biro (b).

[14] For a defence of seeing Hume's treatment of personal identity in this light, see Biro (a).

[15] For details, see Biro (a), (e).

appeals again and again in his explanations of all our beliefs, both natural and philosophical. Familiar under the titles of "custom" and "habit," these natural, irresistible and unnoticed associating and extrapolating tendencies of the imagination *are* the scientific explanations Hume is seeking. They involve "experiments" and "observations" that consist not in peering *into* one's own mind in but looking *out* at the world, the method recommended by Hume for the anatomist of the mind.

Similar points could be made concerning almost any topic Hume treats in the *Treatise*.[16] He seems almost always to be engaged in two searches at once, one for the internal bed-rock, one for the properly objective, hence external, explanation. (And, as I just suggested, matters are further complicated by Hume's unabashed appeal to unobservables.) The two pursuits require him to use different methods; focusing on one to the exclusion of the other is one of the things that lead to the various one-sided and incompatible interpretations in which Hume scholarship abounds.

I shall come back shortly to the question of whether these various aims and methods are compatible. First, however, I want to re-emphasize the striking similarity between Hume's change of course from the speculative philosophical approaches to human nature in his time and that recommended by the various "naturalizing" programmes in epistemology and the philosophy of mind I mentioned earlier. Just as the former were meant by Hume to be a new departure promising real results and progress in the place of the idle —and worse— speculations of the metaphysicians, so the latter are, in large part, a reaction to what is seen as the triviality and inconclusiveness of arm-chair conceptual analysis and so-called "ordinary-language philosophy". As suggested earlier, the more and more *recherche* the examples and counter-examples of the Gettier industry, the less promise, it seemed, of genuine illumination about knowledge. In the philosophy of mind, the endless and seemingly pointless distinction- and idiom-mongering of some of the lesser lights of the ordinary-language school of the fifties and the sixties, along with disenchantment with various brands of behaviourism, led many to look, in something like the way Hume looked to Newton, to the emerging cognitive sciences as a new paradigm for how questions about the mind should be approached.

Many of the results of Hume's first efforts to begin a new science of the mind also have a clear echo in those being delivered by modern empirical studies of our cognitive processes. According to these, one of the most

[16] Compare his account of the difference between having an idea of a colour in terms of having a mastery of a colour concept and having a "just" idea of that colour in the sense of having phenomenal acquaintance with it (T 5); or his analysis of the term 'cause' with his account of the origin of the idea of necessary connection.

striking features of our cognitive capacities and performance, whether in perception, in linguistic processing, or in reasoning generally, is that the states, mechanisms and operations our best theories of them posit are often thought of as —to use current idiom— sub-doxastic, modular, and automatic. Sub-doxastic, because since their subject —the entity to which they are attributed— is not the cognizer himself, but some component sub-system we regard as the locus of the operation or process postulated to explain the cognitive function in question, we think of the states and processes involved as obtaining or taking place below the threshold of the cognizer's consciousness. For this reason, the subject is not a reliable source of information about these processes. Hence the preference in these studies for a third-person approach, rather than a first-person one, for laboratory experiments, such as reaction-time studies and the like, instead of introspection. (Compare Hume's advocacy, noted earlier, of "careful and exact experiments ... judiciously collected and compared," through "a cautious observation of human life.") Second, the processes of interest are modular, in that they are, in the overwhelming majority of cases, found to be task-specific, doing their work largely in isolation from each other and from the cognitive states we would attribute to the person taken as a whole. Thus the processes underlying one particular kind of cognitive capacity or performance rarely, if ever, interact with those of another, and their output is similarly independent of the output of other modules. (Think of the very common case of the different senses delivering different verdicts on the same object or event, with the need for a considerable amount of central processing to reconcile these). Finally, they are, for both these reasons, insensitive to the cognizer's beliefs —even if these are reflective and conscious, rather than merely tacit, and even if he makes an effort to bring them to bear on their workings. (Think of the robustness of perceptual illusions known to be illusions).[17] Hence the common characterization of many, if not most, processes posited to explain our cognitive capacities and performance as "cognitively impenetrable" or "informationally encapsulated."[18] We are not a long way from Hume's "instinct, or mechanical power, that acts in us unknown to ourselves."

[17] I have already mentioned Hume's recognition of our tendency to over-generalize. The same sort of inductive over-generalization has been found to be ubiquitous in our cognitive life, from linguistic processing (certainly in phonological and morphological processing, but even in syntax, e.g., with so-called "garden-path" sentences, and in prosody), to perception (of, for example, edges, or of motion), and general problem-solving and reasoning (as in making clearly fallacious probabilistic inferences). For details, see Ulman, Kahneman, Goldman.

[18] See Pylyshyn.

With all these similarities between Hume's new science and the even newer one of today, we should not overlook the important differences between them. We have already noted one: the difference in attitude to introspection as a method. Equally important is Hume's refusal to abandon those elements of the traditional framework that derive from common sense and our everyday practices, rather than from the rarefied and esoteric activities of philosophers *or* scientists. That is why, to return once again to his clear and explicit explanation of his method at the outset of his project, the experiments of the new scientist must consist of a "cautious observation of human life;" their objects must be taken "as they appear in the common course of the world." Hume therefore has a much more complex task than the modern cognitive scientist, or even the modern naturalizing philosopher. He must try to fit together into a coherent whole a number of elements that do not easily go together: introspection, scientific theorizing, conceptual analysis —and an ultimate allegiance to common sense and common language as the touchstone. This last desideratum complicates his task but, as I now want to suggest, it also yields the key to understanding the surprising extent to which he succeeds in it.

One of the striking features of the common-sense view of the world, and of human nature in particular, is that it is not solipsistic in the way that Hume's treatment of the self in Book I of the *Treatise* has appeared to many readers to be.[19] We naturally see ourselves as part of an external world, both physical and social. Not only is it idle to ask whether bodies exist (though it is not at all idle to ask how we come to believe that they do), it is equally idle to ask whether other people do. Surprising as this may sound to ears attuned to some of the traditional interpretations of Hume's account of the self, it will seem natural to those who remember that that account is not limited to, and is not completed in, the notorious discussion of Book I. The self of the later parts of the *Treatise*, social, engaged in action, seeing itself as so engaged, is not a different self from the lone introspective brooder of Book I, nor is it somehow—and questionably—derived from the latter by some inference. It is, rather, the very same self considered in a different light. Indeed, this later self is the real self, so to say, the one we all take ourselves to be. And that self depends for its very being and identity on its relations to others like it. Hume's emphasis on the mechanism of sympathy so prominent in the later books would not make any sense were this not so.

[19] I have myself emphasized the first-person and foundational character of that treatment (Biro (a)). But it is important to recognize that Hume's aim in Book I is descriptive and explanatory, rather than justificatory: it is to give a psychological account of how one's beliefs (including one's belief in one's own identity) arise, rather than an epistemological grounding of them. Part of Hume's message is that first-person materials are inadequate for the latter. (See Biro (b) and below.)

I have put the point in the strong, even extreme, way I just have —the self depends for its very being and identity on others— to draw attention to another striking anticipation of twentieth-century philosophical developments by Hume, this time in a very different neck of the philosophical wood. That the self is, in a strong sense, a social product is a common-place in much of the still vital philosophical tradition that derives from German and French thinkers of the nineteenth and twentieth centuries. Social-construction theories of identity, of personhood, of ethnicity, of gender, of —you name it, are all over the place, some to be taken seriously, some... well... Whatever their quality, these theories all rely essentially on the insight elegantly expressed by Hume in saying that "the minds of men are as mirrors to one another" (T 365), and systematically followed up by him in the later books of the *Treatise* (and elsewhere) to yield what could be defended as the most subtle, most comprehensive, most coherent, and most realistic, moral psychology and moral philosophy since Aristotle.[20] A full justification of this claim is, of course, well beyond the scope of the present paper. For now it must serve merely to take us back to the question I postponed a little way back: Can these different aspects of Hume's inquiries, and the different methods they require him to use, be fitted into one coherent whole, can they be seen as complementary parts of a coherent scientific-*cum*-philosophical project? I suggest that the answer to at least the latter question may be "Yes!," even if we have to concede that Hume himself perhaps does not quite succeed in fully knitting the different strands together.

To make out the correctness of an affirmative answer would take much evidence and argument. I have enough of neither, nor the space to give what I do have. So, I shall close by giving what is no more than the barest outline of an account that, if made out, would yield such a positive answer.

The key to the account I have already gestured at: it lies in placing Hume's *theories* of the nature and workings of the mind on the right side of the distinction between the justificatory project of traditional epistemology and the descriptive one of cognitive (and moral) psychology. It is important to emphasize that this extends even to the first of the two aims I attributed to Hume, namely, to that part of his explanation of the genesis of our beliefs in the service of which he employs the first-person, introspective, phenomenological method. As we have seen, the explanation is not complete without its complementary third-person, experimental, scientific element. But in both components Hume's goal is to discover by one or another kind of observation how the mind actually works, in contrast to others who either

[20] Perhaps one can see so-called social externalism of the kind advocated by Burge as an expression of one aspect of the same insight.

speculate on, or stipulate, how it *must* work. (Neither, as I have argued, is it his goal to show that we cannot know how it works.) And, while I cannot go into the matter here, I want to note at least that I believe it can be shown conclusively that this descriptive thrust of Hume's project does not preclude his giving an account of immanent normativity.[21]

From this perspective, we can, I think, say that while there is work for both methods to do, they are perhaps, in the end, neither as different nor as independent of each other as they may at first sight seem. The link between them lies in the role of language: in the non-solipsistic contents of the terms we must rely on even when we adopt a first-person method. This is, of course, a long story, one to which I can no more than point.[22] But the basic idea is that even when turning inward, what one finds are items whose very identity depends on their representational content, which is, in turn, inseparable from the meaning of the words one uses to express it. Hume's theory of meaning, so far from being one on which, as it is often alleged, the meaningfulness of words derives from mental items with no representational or semantic properties, is just the opposite: it is the public meanings of the words that individuate the mental items to which they (purport to) point. Thus it is that he can allow that even the blind can be said to possess ideas —concepts, as we would say—of colours in virtue of their ability to use colour terms. Hume's phenomenology may be seen as what Dennett has called "hetero-phenomenology," in contra-distinction to "auto-phenomenology," as turning, before the reader's eyes, into something like what Austin must have had in mind in speaking as he did of "linguistic phenomenology." And should not be surprised if before long cognitive scientists themselves come to question the assumption, so far dominant in their field, that a science of cognition requires an internalist, individualist, approach of the sort *one* strand in Hume's method exemplifies.

As I have said, all this is sketchy, perhaps hopelessly so. Still, it may give a feel for how Hume's new —at his time— insights about the mind quite uncannily anticipate those of recent cognitive science. While his idiom is, of course, not ours, while there are real differences between the method he recommends and the one he actually uses, as well as between these and those we favour, (and while his philosophical concerns go beyond the scientific study of the mind, something I have had to put aside in this

[21] Among those who *have* shown this, in various ways, are Ardal, Norton, Penelhum, and Swain.

[22] Some of the elements of the story have been provided by writers like Ardal, in his account of Hume's conventionalist account of language, Livingston, in his account of Hume's historical —and therefore social— theory of meaning, and Baier and others who have emphasized the essentially social picture Hume ultimately gives of the self.)

discussion), he may nonetheless be said with more justice than any philosopher before him to be the philosophical father of cognitive science.

REFERENCES

Anderson, R.F., *Hume's First Principles* (Lincoln: University of Nebraska Press. 1966).

Ardal, P., "Convention and Value," in *David Hume: Bicentenary Papers,* G.P. Morice ed., (Edinburgh: University of Edinburgh Press, 1977).

Austin, J.L., "A Plea for Excuses," in *Philosophical Papers* (Oxford: Oxford University Press, 1961).

Baier, A., *A Progress of Sentiments: Reflections on Hume's Treatise* (Cambridge: Harvard University Press, 1991).

Barfoot, M. "Hume and the Culture of Science in Early Eighteenth-Century Britain."

Bennett, J., *Locke, Berkeley, Hume: Central Themes* (Oxford: Oxford University Press, 1971).

Biro, J.I. (a),"Hume on Self-Identity and Memory," *The Review of Metaphysics* 30, 1976.

- (b),"Hume's difficulties with the self," *Hume Studies* V, 1979.

- (c),"Description and Explanation in Hume's Science of Man," *Transactions of the Fifth International Congress on the Enlightenment,* (New York: The Voltaire Foundation, 1979).

- (d), "Hume and cognitive science," *History of Philosophy Quarterly* 2, 1985.

- (e), "Hume's New Science of the Mind," in Norton (b).

- (f), "Memory, Mind, and Society," unpublished mss.

Bricke, J. (a), "Hume on Self-Identity, Memory and Causality," in Morice.

- (b) *Hume's Philosophy of Mind* (Princeton: Princeton University Press).

Burge, T., "Individualism and the Mental," *Midwest Studies in Philosophy,* 1979.

Broughton, J., "What does the scientist of man observe?" *Hume Studies,* vol. XVIII, no.2, 1992.

Capaldi, N., *David Hume: The Newtonian Philosopher* (Boston: Twayne Publishers, 1975).

Connon, R.W., "The Naturalism of Hume Revisited," in Norton (a).

Dennett, D., *Brainstorms* (Cambridge: MIT Press, 1978).

Davie, G., "Edmund Husserl and 'the as yet, in its most important respect, unrecognized greatness of Hume'," in Morice.

Fodor, J., (a) "Mental Representation: An Introduction," in Rescher.

- (b) *Humean Variations* (mss.).

Fogelin, R.J., (a) *Hume's Scepticism in the Treatise of Human Nature* (London: Routledge and Kegan Paul, 1985).

- (b) "Hume's scepticism," in Norton (b).

Garrett, D., *Cognition and Commitment in Hume's Philosophy* (New York: Oxford University Press, 1997).

Goldman, A., *Epistemology and Cognition* (Cambridge: Harvard University Press, 1987).

Hume, D., (a) *A Treatise of Human Nature,* L.A. Selby-Bigge, ed., rev. by P.H. Niddich (Oxford: Clarendon Press, 1987).

- (b) *An Enquiry Concerning Human Understanding,* L.A.Selby-Bigge, ed., (Oxford: Clarendon Press, 1970).

James, W., *The Principles of Psychology* (New York: Dover, 1950).

Kahnemann, D et al. (eds.), *Judgments under Uncertainty: Heuristics and Biases* (Cambridge: Harvard University Press, 1982).

Kornblith, H. (ed.), *Naturalized Epistemology* (Cambridge: Harvard University Press, 1985).
Livingston, D. (a), "Hume's Historical Theory of Meaning," in Livingston and King.
- (b), *Hume's Philosophy of Common Life* (Chicago: University of Chicago Press, 1984).
Livingston, D. and King, J. (eds.), *Hume: A Re-evaluation* (New York: Fordham University Press, 1976).
Monteiro, J-P., "Hume's Conception of Science," *Journal of the History of Philosophy* 19, 1981.
Morice, G.P. (ed.), *David Hume: Bicentenary Papers* (Edinburgh: University of Edinburgh Press, 1977).
Norton, D.F. (ed.), *The Cambridge Companion to Hume* (Cambridge: Cambridge University Press, 1993).
Norton, D.F. et al. (eds.), *McGill Hume Studies* (San Diego: Austin Hill Press, 1979).
Noxon, J., *Hume's Philosophical Development* (Oxford: Clarendon Press, 1973).
Owen, D., *Hume's Reason* (Oxford: Oxford University Press, 1999).
Penelhum, T., (a) "Hume on Personal Identity," *The Philosophical Review* 54, 1955.
- (b) "Hume's Theory of the Self Revisited," *Dialogue* 14, 1975.
- (c) *Hume* (London: Macmillan), 1975.
Pylyshyn, Z., *Computation and Cognition: Toward a Foundation for Cognitive Science* (Cambridge: Bradford/MIT Press, 1984).
Rescher, N. (ed.), *Scientific Inquiry in Philosophical Perspective* (Washington: University Press of America, 1987).
Robison, W.L., (a) "Hume on Personal Indentity," *Journal of the History of Philosophy* 12, 1974.
- (b) "In Defense of Hume's Appendix," in Norton (a).
Rosenberg, A., "Hume and the Philosophy of Science," in Norton (b).
Shope, R.K., *The Analysis of Knowledge: a Decade of Research* (Princeton: Princeton University Press, 1983).
Smith, N.K., *The Philosophy of David Hume* (London: Macmillan, 1949).
Stove, D.C., "The Nature of Hume's Skepticism," in Norton (a).
Stroud, B., (a) *Hume* (London: Routledge and Kegan Paul, 1977).
- (b) "Hume's Scepticism: Natural Instincts and Philosophical Reflection," *Philosophical Topics* 19.
Swain, C. "Being Sure of One's Self: Hume on Personal Identity," *Hume Studies,* 1991.
Ullman, S., *The Interpretation of Visual Motion* (Cambridge: Harvard University Press, 1979).
Winkler, K., "The New Hume," *Philosophical Review* 50, 1991.
Wright, J., *The Skeptical Realism of David Hume*, (Manchester: Manchester University Press, 1983).

Chapter 2

TRUTH AND MEANING

The Role of Truth in the Semantics of Propositional Attitude Ascriptions

Scott Soames
Princeton University

In contemporary philosophy, theories of truth take two main forms. One form aims to provide an analysis of truth that illuminates the content of claims that x is true or not true, that explains the uses to which the notion of truth is legitimately put, and that can be employed to diagnose and resolve paradoxes involving truth, such as the liar. The other main form taken by contemporary theories of truth is found in semantics, where theories of truth are used to interpret, or give the meanings of, sentences. When we use theories of truth to play this role, we are not attempting to provide an analysis of truth, or to specify the content of the truth predicate. Rather, we take truth for granted, and use it to illuminate the meanings of sentences by giving their truth conditions. Here it is usually taken for granted either that the meaning of a sentence consists of nothing more than its truth conditions, or that truth conditions constitute a central component of that meaning. As a result, theories of meaning are sometimes taken to be nothing more than theories of truth, while at other times theories of truth are viewed as central subcomponents of theories of meaning.

Today I will talk about the role of truth in theories of meaning. My focus will be on propositional attitude ascriptions, and the problems they pose for truth-conditional theories. I will argue that contemporary theories that aspire to interpret, or specify the meanings of, such ascriptions, simply by deriving their truth conditions, are doomed to failure. Next, I will argue that although it remains possible for certain truth-conditional theories to provide some information about the meanings of propositional attitude ascriptions, there

K. Korta and J. M. Larrazabal (eds.), Truth, Rationality, Cognition, and Music, 21-44.

Kornblith, H. (ed.), *Naturalized Epistemology* (Cambridge: Harvard University Press, 1985).
Livingston, D. (a), "Hume's Historical Theory of Meaning," in Livingston and King.
- (b), *Hume's Philosophy of Common Life* (Chicago: University of Chicago Press, 1984).
Livingston, D. and King, J. (eds.), *Hume: A Re-evaluation* (New York: Fordham University Press, 1976).
Monteiro, J-P., "Hume's Conception of Science," *Journal of the History of Philosophy* 19, 1981.
Morice, G.P. (ed.), *David Hume: Bicentenary Papers* (Edinburgh: University of Edinburgh Press, 1977).
Norton, D.F. (ed.), *The Cambridge Companion to Hume* (Cambridge: Cambridge University Press, 1993).
Norton, D.F. et al. (eds.), *McGill Hume Studies* (San Diego: Austin Hill Press, 1979).
Noxon, J., *Hume's Philosophical Development* (Oxford: Clarendon Press, 1973).
Owen, D., *Hume's Reason* (Oxford: Oxford University Press, 1999).
Penelhum, T., (a) "Hume on Personal Identity," *The Philosophical Review* 54, 1955.
- (b) "Hume's Theory of the Self Revisited," *Dialogue* 14, 1975.
- (c) *Hume* (London: Macmillan), 1975.
Pylyshyn, Z., *Computation and Cognition: Toward a Foundation for Cognitive Science* (Cambridge: Bradford/MIT Press, 1984).
Rescher, N. (ed.), *Scientific Inquiry in Philosophical Perspective* (Washington: University Press of America, 1987).
Robison, W.L., (a) "Hume on Personal Indentity," *Journal of the History of Philosophy* 12, 1974.
- (b) "In Defense of Hume's Appendix," in Norton (a).
Rosenberg, A., "Hume and the Philosophy of Science," in Norton (b).
Shope, R.K., *The Analysis of Knowledge: a Decade of Research* (Princeton: Princeton University Press, 1983).
Smith, N.K., *The Philosophy of David Hume* (London: Macmillan, 1949).
Stove, D.C., "The Nature of Hume's Skepticism," in Norton (a).
Stroud, B., (a) *Hume* (London: Routledge and Kegan Paul, 1977).
- (b) "Hume's Scepticism: Natural Instincts and Philosophical Reflection," *Philosophical Topics* 19.
Swain, C. "Being Sure of One's Self: Hume on Personal Identity," *Hume Studies*, 1991.
Ullman, S., *The Interpretation of Visual Motion* (Cambridge: Harvard University Press, 1979).
Winkler, K., "The New Hume," *Philosophical Review* 50, 1991.
Wright, J., *The Skeptical Realism of David Hume*, (Manchester: Manchester University Press, 1983).

attitude ascription changes truth value.[18] Although solving these problems is no easy task, I believe it can be accomplished. However the argument for this is something that must be left for another time.[19]

[18] For an attempt to deal with apparent instances of substitution failure in propositional attitude ascriptions see my treatment of the issue in *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*, (New York: Oxford University Press, 2002).

[19] The material in this paper extends and develops a point sketched in the first few pages of chapter 7 of *Beyond Rigidity*. I would like to thank the commentators on my paper and the participants at the ICCS-01 conference for useful comments. Timothy Williamson, in particular, made a helpful observation concerning examples (11) and (12).

Chapter 3

TRUTH AND BORDERLINE CASES

Timothy Williamson
University of Oxford

1. INTRODUCTION

According to the principle of bivalence, truth and falsity are jointly exhaustive and mutually exclusive options for a statement. It is either true or false, and not both, even in a borderline case. That highly controversial claim is central to the epistemic theory of vagueness, which holds that borderline cases are distinguished by a special kind of obstacle to knowing the truth-value of the statement. But this paper is not a defence of the epistemic theory. If bivalence holds, it presumably does so as a consequence of what truth and falsity separately are. One may therefore expect bivalence to be derivable from a combination of some principles characterizing truth and other principles characterizing falsity. Indeed, such derivations are easily found. Their form will of course depend on the initial characterizations of truth and falsity, and not all such characterizations will permit bivalence to be derived. This paper focusses on the relation between its derivability and some principles about truth and falsity. Borderline cases for vague expressions are primary examples of an urgent challenge to bivalence.

A key variable in the relation is obviously the choice of a logic. The background logic here is classical. That choice does not automatically prejudge the issue in favour of bivalence, for the latter can fail even in a classical context on some non-standard accounts of truth and falsity; traditional supervaluationist semantics provides the best-known instance.

K. Korta and J. M. Larrazabal (eds.), Truth, Rationality, Cognition, and Music, 45-66.

My strategy is to start not with semantic theories designed to have some particular result for vague languages but with principles which seem natural from the standpoint of the theory of truth and falsity. I then explore their implications for bivalence. More specifically, simple considerations about the semantics of indexicals motivate replacing the usual disquotational characterization of truth by one which makes explicit allowance for contextual variation in what is said by a sentence. The apparatus I use for that purpose *appears* to make room for failures of bivalence in borderline cases. For the argument for bivalence requires a principle of uniformity in what a given sentence says in a given context, and borderline cases may appear to motivate a denial of such uniformity. I will argue that the appearance is illusory.

2. SAYING AND DISQUOTATION

A formal framework is needed in which to conduct the investigation. Since the concern is with vagueness in language, I will treat sentences as the bearers of truth and falsity. That is consistent with the idea that the truth-value of a sentence is determined by the truth-value of a proposition which it expresses. Of course, a sentence may express different propositions with respect to different contexts of utterance, if it contains indexicals or demonstratives. Thus the truth-value of a sentence is also relative to a context of utterance. The expression **Say(s,c,P)** will mean: the sentence s as uttered in the context c says (or is used to say) that P (where 'P' may be replaced by a declarative sentence); in terms of propositions, s in c expresses the proposition that P. **True(s,c)** (respectively, **False(s,c)**) means: s is true (respectively, false) in c. One could easily extend the discussion from vagueness in language to vagueness in non-linguistic thought by letting **Say** mean the more general notion of expressing and s range over thought types as well as sentences.

I will take for granted the principles of classical logic, appealing without comment to substitution instances of theorems of the classical propositional calculus and its extensions for the relevant types of quantifier, and use the rule of *modus ponens* similarly. Although these logical assumptions are of course not uncontroversial in the case of vagueness, justifying them is not my current concern. They are in any case not exclusive to a particular theory of vagueness; they are common to epistemicism, supervaluationism and some other views. Moreover, as it happens, many — but not all— of the arguments will use only uncontroversial fragments of classical logic.

Within this framework, the natural principle about truth is that if a sentence says that something is so, then it is true if and only if that thing is so. More precisely:

(T) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset [\mathbf{True}(s,c) \equiv P]]$

(T) is not itself a disquotational principle about truth, for it contains no occurrence of **P** in quotation marks (explicit or implicit). To recover a disquotational biconditional of the form $\mathbf{True}(\langle\langle P \rangle\rangle,c) \equiv P$ from (T), one would need a corresponding auxiliary premise of the form $\mathbf{Say}(\langle\langle P \rangle\rangle,c,P)$. Such a premise holds if the sentence which replaces **P** contains no context-dependent elements. For example, if '$E = mc^2$' is context-independent, then in any context it says that $E = mc^2$, and so is true if and only if $E = mc^2$. Even if **P** is replaced by a context-dependent sentence, the auxiliary premise $\mathbf{Say}(\langle\langle P \rangle\rangle,c,P)$ still holds if **c** refers to the context in which that premise is being asserted. For example, as uttered by me 'I am Timothy Williamson' says that I am Timothy Williamson, and so is true if and only if I am Timothy Williamson. But 'I am Timothy Williamson' as uttered by you does not say that I am Timothy Williamson, for 'I' as uttered by you does not refer to me. The sentence as uttered by you is not true. The disquotational biconditional holds only under restricted circumstances. But (T) does better; it handles context-dependence with ease. Since 'I am Timothy Williamson' as uttered by you does not say that I am Timothy Williamson, the corresponding instance of (T) is vacuously true. That sentence as uttered by you does say that you are Timothy Williamson, and it is true as so uttered if and only if you are Timothy Williamson. Thus (T) is more basic than the disquotational biconditional; it explains both the successes and the failures of the latter.

A further advantage of (T) over the disquotational biconditional is that the latter but not the former must be revised to meet semantic paradoxes such as the Liar. For example, in some context c one can construct a self-referential sentence s to be $\sim\mathbf{True}(s,c)$. The corresponding disquotational biconditional is $\mathbf{True}(\langle\langle \sim\mathbf{True}(s,c) \rangle\rangle,c) \equiv \sim\mathbf{True}(s,c)$; since $s = \langle\langle \sim\mathbf{True}(s,c) \rangle\rangle$ holds, one can deduce $\mathbf{True}(s,c) \equiv \sim\mathbf{True}(s,c)$, which is logically false. Thus not every instance of $\mathbf{True}(\langle\langle P \rangle\rangle,c) \equiv P$ holds. But this argument does not show that (T) is invalid; it merely falsifies the antecedent $\mathbf{Say}(\langle\langle \sim\mathbf{True}(s,c) \rangle\rangle,c,\sim\mathbf{True}(s,c))$. Not even the sentence 'This sentence is not true' succeeds in saying of itself that it is not true (for further discussion see Williamson 1994: 197 and 1998).

The variable **P** takes sentence position in (T). In particular, it flanks the biconditional ≡. Correspondingly, the third argument place in $\mathbf{Say}(s,c,P)$ is for a sentence, not a name (not even a name of a sentence); with respect to

that argument place, **Say** works like an operator rather than a (first-level) predicate. Thus (T) involves quantification into sentence position. Such quantification might be interpreted substitutionally. The sentences substituted for **P** would be declarative sentences of the language in which the theorist is working (English, say). This would not restrict the instances of (T) which hold non-vacuously to those in which the sentence to which **s** refers is itself a sentence of English, for we can express in English what a sentence of some other language says. For example, the Serbian sentence 'Jedan sto je u sobi' as uttered in a suitable context says that one table is in the room, and so is true if and only if one table is in the room. However, if in some context a sentence s says something which cannot be expressed in English, at least in the context in which the theorist is working, then no instance of (T) holds non-vacuously for that case. (T) would still be true, but it would not be as informative about such cases as we should like it to be. We might therefore consider a non-substitutional reading of the quantifier $\forall$**P**, on which $\forall$**P** φ(**P**) might be false even if φ(**P**) was true whenever **P** is replaced by a sentence of the language in which the theorist is working. For example, given fixed assignments to the variables **s** and **c**, $\forall$**P** ~**Say**(**s**,**c**,**P**) might be false even though every corresponding substitution instance of the form ~**Say**(**s**,**c**,**P**) was true. Such a non-substitutional interpretation would *not* automatically be objectual, for example over all propositions conceived as objects of a special kind. Objectual quantification is quantification whose semantics is given by non-substitutional quantification *into name position* in the metalanguage; both 'proposition' and 'object' are nouns, not sentences. One should not dismiss without argument the possibility that non-substitutional quantification into sentence position is irreducible because its semantics can be given faithfully only by non-substitutional quantification into sentence position in the metalanguage too. It is not obviously wrong to suppose that one can understand such non-objectual non-substitutional quantification even if it cannot be expressed unequivocally in English; that might be a defect of English. We shall not attempt to decide between the substitutional and non-substitutional options here. Both are consistent with the arguments below. Even if the substitutional interpretation is not the intended one, it still provides a consistency proof of a standard theory of quantification into sentence position. The use made below of such quantification is quite undemanding, for sentence variables are not replaced by sentences that themselves contain either quantification into sentence position or one of the expressions **True**, **False** or **Say**.

Corresponding to (T), the natural principle about falsity is that if a sentence says that something is so, then it is false if and only if that thing is not so. More precisely:

(F) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset [\mathbf{False}(s,c) \equiv \sim P]]$

For example, in a context in which the Serbian sentence 'Jedan sto je u sobi' says that one table is in the room, it is false if and only if it is not the case that one table is in the room. One needs the same auxiliary premise **Say(«P»,c,P)** to infer the disquotational biconditional **False(«P»,c)** ≡ **~P** from (F) as one does to infer the corresponding disquotational biconditional **True(«P»,c)** ≡ **P** from (T). Much more will be said about both (T) and (F) below.

3. BIVALENCE

How should the principle of bivalence be formulated within this framework? The principle should not imply that non-declarative sentences are true or false, for presumably they are not intended to say that something is the case. For the same reason, the principle does not imply that a declarative sentence is true or false if it does not say that something is the case. For example, if a language teacher writes the sentence 'That belongs to her' on the board to illustrate a point of grammar, without attempting to supply a reference for the demonstratives, the principle of bivalence does not require the sentence to be false. Thus one can reasonably build into the antecedent condition that the sentence says that something is the case, just as in (T) and (F). The usual formulation then says, on that condition, that the sentence is either true or false, where the disjunction is understood as inclusive. Since classical logic is being assumed, that is equivalent to the claim that the sentence is false if it is not true:

(WB) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset [\sim \mathbf{True}(s,c) \supset \mathbf{False}(s,c)]]$

Call (WB) the principle of *Weak Bivalence*. It is weak because, by itself, it is consistent with the supposition that (for each fixed context) **True** and **False** stand for exactly the same property, one possessed by every sentence saying that something is so, even if there are many such sentences. In that case **True** and **False** would stand for one and the same truth-value, not two, and the term 'bivalence' would be a misnomer. One can strengthen weak bivalence by adding its converse, the principle that truth and falsity are mutually exclusive for a sentence that says that something is so:

(ME) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset [\mathbf{False}(s,c) \supset \sim \mathbf{True}(s,c)]]$

Suppose that a sentence is true only if it says that something is so; then one can drop the antecedent **Say(s,c,P)** and (ME) implies the simpler principle that truth and falsity are unconditionally mutually exclusive. (WB) and (ME) can be combined into a single principle of *Strong Bivalence*:

(SB) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset [\sim \mathbf{True}(s,c) \equiv \mathbf{False}(s,c)]]$

In classical logic, the combination of the biconditional and negation can be reformulated as an exclusive disjunction: a sentence that says that something is so is either true or false and not both. (SB) stands to (WB) as exclusive stands to inclusive disjunction. For any context in which at least one sentence says that something is so, (SB) requires **True** and **False** to stand for distinct properties, and the term 'bivalence' is appropriate. Contrary to the usual practice, I will concentrate on the strong rather than weak form of bivalence. Of course, (SB) is equivalent to (WB) given the almost universally accepted principle (ME). Nevertheless, (SB) does better than (WB) in making some logical and philosophical connections salient (Williamson (1994, 1995 and elsewhere) uses 'bivalence' for weak bivalence; the idea of working with strong bivalence was proposed in an earlier version of Andjelković (1999)). Dialetheists think that paradoxical sentences such as the Liar are both true and false; on some readings they accept (WB) but reject (SB) and (ME). I discuss the status of (SB), (WB) and (ME) below.

The logical relations between (T), (F) and (SB) are rather simple: any two of them entail the third. To check that, note that their main biconditionals are (classically equivalent to) $\sim\mathbf{True}(s,c) \equiv \sim P$, $\mathbf{False}(s,c) \equiv \sim P$ and $\sim\mathbf{True}(s,c) \equiv \mathbf{False}(s,c)$ respectively and that the biconditional behaves symmetrically and transitively. Thus, given (T), (F) is equivalent to (SB); given (F), (T) is equivalent to (SB); given (SB), (T) is equivalent to (F).

For future reference, it is convenient to label the two directions of (T) and (F) separately:

(T→) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset [\mathbf{True}(s,c) \supset P]]$

(T←) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset [P \supset \mathbf{True}(s,c)]]$

(F→) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset [\mathbf{False}(s,c) \supset \sim P]]$

(F←) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset [\sim P \supset \mathbf{False}(s,c)]]$

One can easily check that these logical relations hold amongst the unidirectional principles:

(T←) and (F←) entail (WB);

(T→) and (F→) entail (ME);

(F←) and (ME) entail (T→);

(F→) and (WB) entail (T←);

(T←) and (ME) entail (F→);

(T→) and (WB) entail (F←).

By considering unintended interpretations of **True** and **False**, one can check that no further entailments obtain amongst the six halves of (T), (F) and (SB) beyond those implicit in the foregoing list, if **True** and **False** are treated as primitive predicates, not as logical constants (Andjelkovic´ and Williamson 2001).

Epistemicists about vagueness typically argue from (T) and (F) to (SB), or more specifically from (T←) and (F←) to (WB) (Williamson 1994: 188-189 and 1995). Someone could consistently accept both (T) and classical logic while still rejecting (F), and therefore (SB), and therefore epistemicism (Andjelković (1999) points out that such a combination is possible; Williamson (1999) responds). Thus the notion of falsity should play a significant role in the discussion of vagueness; principles such as (F) need explicit attention. Unfortunately, the literature has neglected (F) and cognate principles about falsity.

4. FALSITY AND NEGATION

Can the notion of falsity be defined in terms of the notion of truth? If so, one might be able to use the definition to reduce (F) to (T), and therefore to derive (SB) from (T). In that case, the neglect of the notion of falsity would be more apparent than real, since the notion would have been treated implicitly in the treatment of the notion of truth. Presumably such a definition should not use auxiliary conceptual resources so strong that they would suffice by themselves for a direct definition of falsity without the detour through truth.

The simplest attempted definition of falsity is as non-truth: **False(s,c)** $=_{def}$ **~True(s,c)**. Under this definition, (SB) becomes a trivial logical truth, and (F) reduces to (T). However, the definition has very implausible consequences. A sentence which does not say that something is so counts as false simply because it is not true. Indeed, mountains and lakes count as false for the same reason. One could revise the definition by defining falsity as the conjunction of non-truth with saying that something is so: **False(s,c)** $=_{def}$ **~True(s,c) & ∃P Say(s,c,P)**. According to epistemicism, such a definition would give at least extensionally correct results. But it would not be acceptable to those who reject (SB); a less controversial definition would be preferable. Moreover, it employs quantification into sentence position; given such quantification, one might as well define falsity directly, rather than making an unnecessary detour through the notion of truth (see below).

Falsity is often defined as the truth of the negation (see below for what counts as a negation). Such a definition of falsity may be understood as implying that an item is false only if it has a true negation, and therefore only if it has a negation, so that it is not false if it has no negation at all. Since lakes and mountains have no negations, they are not false. If a sentence s fails to say that something is so, then either s lacks a negation or its negation also fails to say that something is so, for what the negation of s says would be the contradictory of what s says; either way, s does not count as false.

Let the singular term **Ns** refer to a negation of the sentence to which the term **s** refers. If the latter has several verbally different negations, it does not matter which of them the former refers to. Thus the definition of falsity as the truth of the negation yields this principle:

(FN) $$\forall s \forall c[\mathbf{False(s,c)} \equiv \mathbf{True(Ns,c)}]$$

The biconditional in (FN) is not conditional on the antecedent **Say(s,c,P)**, for if s fails to say that something is so then, as noted above, Ns also fails to

say that something is so; consequently, neither **False(s,c)** nor **True(Ns,c)** holds, and therefore the biconditional holds vacuously.

Obviously, (FN) does not help unless we can independently characterize negation (**N**). The easiest way to do so is by using negation in the metalanguage to say what Ns says:

(N) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \equiv \mathbf{Say}(Ns,c,\sim P)]$

Note that, given (N), one cannot automatically assume that a sentence is the negation of its negation. For two applications of (N) yield **Say(s,c,P)** ≡ **Say(NNs,c,~~P)**; on a fine-grained notion of saying, **Say(s,c,~~P)** is not equivalent to **Say(s,c,P)** (we might still allow that a sentence is a *contradictory* of its negation). On a very fine-grained view, not endorsed here, the negation of a sentence consists of a negation operator and that sentence, the former applied to the latter.

Given (FN) and (N), we can reduce (F) to (T) and thereby derive (SB). For as a special case of (T) we have **Say(Ns,c,~P)** ⊃ **[True(Ns,c)** ≡ **~P]**, which by (N) yields **Say(s,c,P)** ⊃ **[True(Ns,c)** ≡ **~P]**, which by (FN) yields **Say(s,c,P)** ⊃ **[False(s,c)** ≡ **~P]** and therefore (F). Once we have (T) and (F), we can derive (SB).

Even without (FN), one can use (T) and (N) to derive a principle related to (SB) (although without the implication that the sentences takes one of two values) with **True(Ns,c)** in place of **False(s,c)**: in a context in which a sentence says that something is so, either it or its negation is true. Thus the rejection of (F) by itself would not avoid all the controversial consequences of an epistemic theory of vagueness, since one would still be committed to the claim that when a vague sentence in a borderline case says that something is so, either it or its negation is true, and not both, even though we have no idea how to find out which of the two contradictory sentences is true.

Let us return to the original principle (SB). To assess it, one must consider the notion of falsity itself, and therefore (FN). What happens to (FN) if s has no negation? If s is not a candidate for truth or falsity —for example, if s is a lake or mountain— then (FN) holds vacuously. But might not s be a candidate for truth or falsity and still have no negation, because it is in an expressively impoverished language? Creatures might in principle communicate information about their environment to each other in a limited system of signals without negation. Perhaps some animals actually do so. Such signals could be true or false, depending on the relation between the signalled state of the environment and its real state. Suppose, for example, that the animals can signal that there is food over there. Thus, when there is food over there, they can communicate truly that there is food over there.

When there is no food over there, they can communicate falsely that there is food over there. That does not require them to have the capacity to negate a signal; they may be unable to signal that there is no food over there. Of course, questions can be raised about the legitimacy of attributing propositional content to signals in so simple a system. Nevertheless, I see no decisive obstacle to such attributions. Creatures are fallible; the capacity to get things right requires the capacity to get them wrong. If the capacity to negate signals is not strictly necessary for communicating true information, it is also not strictly necessary for communicating false information.

A friend of (FN) might reply that a negation operator N can always in principle be added to the system, so that one can define the falsity of s as the truth of its hypothetical negation Ns. This proposal must be formulated with some care, for the falsity of s in a given situation is equivalent to the truth of its negation in that very situation, not in a counterfactual system in which negation has been introduced into the system. For the presence of negation in the signal system might indirectly affect the aspect of the environment about which information is being communicated, by affecting the behaviour of the animals. Intuitively, the complications of working out what might happen in counterfactual circumstances in which a signal had a negation are quite irrelevant to the truth or falsity of its actual tokens. Nor is the definition of the falsity of s as the *actual* truth of a merely *possible* or abstract negation of s particularly attractive. There is a simpler way.

I propose to treat falsity as a primary notion on a par with truth. If negation has a role to play in the characterization of falsity, it is negation as used in the metalanguage by the theorist, not negation as used in the object-language by the speakers under study. For example, ~ occurs in (F) as the metalanguage negation, whereas N is mentioned in (FN) as the object-language negation. Given (T) and (N), (F) is equivalent to (FN) whenever **Ns** is well-defined, so (F) and (FN) do not compete for correctness, although they may compete for primacy.

Even in a language without negation, vagueness can arise in virtue of borderline cases. In some contexts, a negationless signal may be neither clearly true nor clearly false. A sorites series may be possible for it: for instance, a series of contexts, each indiscriminable from the next, in the first of which it is true and in the last of which it is false. Although one needs logical constants to formulate the sorites reasoning explicitly, one does not need them merely to have difficulty in using the signal correctly in such a series of contexts. For example, the case may be classified differently depending on which end of a sorites series it is approached from. Without the logical constants, one cannot properly reflect on the significance of the phenomena of vagueness, but at least some of the phenomena themselves can occur.

The possibility of vague but negationless signals also suggests that borderline cases should not primarily be conceived as those in which speakers are poised symmetrically between a positive and a negative assertion, since the latter is not always an option. Certainly the pragmatic significance of the absence of a positive assertion does not give it the semantic significance of a negative assertion.

The concepts of truth and falsity are parallel in most respects, but not in all, at least as they are explained by (T) and (F). For (F) uses negation in the metalanguage at a point at which (T) does not use any operator at all. I have not defended the use of negation rather than some other operator to define falsity, although it seems very natural; if (F) is coherent, what does it characterize other than falsity? By substituting other operators in the metalanguage for ~, one can construct a variety of other concepts. For example, if ~ is replaced by $\Box$, read 'necessarily', the result characterizes the concept of necessary truth. If ~ is replaced by Δ~, where Δ is read 'clearly', a concept of clear falsity is characterized, and one should not expect to derive the analogues of (SB) or (WB). Within classical logic, the principle of bivalence stands or falls with the connection between falsity and standard negation.

5. DEFINITIONS OF TRUTH AND FALSITY

One might be tempted to regard (T) and (F) as implicit definitions of **True** and **False** respectively. That would be a mistake. Even granted the correctness of (T) and (F), in one respect they are too weak to constitute definitions; in another respect they are too strong to do so. They guarantee neither uniqueness nor existence.

In what follows we will treat **True** and **False** as so far uninterpreted symbols, but the other primitive terms of the formal language as already interpreted. One can therefore ask how various principles constrain the interpretation of **True** and **False**. (T) and (F) are too weak to define **True** and **False** because they are quite neutral about the application of those terms to a sentence when it says nothing. Presumably a sentence is neither true nor false in a context in which it says nothing, but (T) and (F) impose no such constraint. They are equally consistent with the unappealing supposition that a sentence is both true and false in a context in which it says nothing. Consequently, they do not characterize **True** and **False** uniquely. To be more precise, let (T_1) and (T_2) be the results of subscripting **True** in (T) by '1' and '2' respectively. Then from (T_1) and (T_2) on the assumption **Say(s,c,P)** one can deduce $\mathbf{True_1(s,c) \equiv P}$ and $\mathbf{True_2(s,c) \equiv P}$ and therefore $\mathbf{True_1(s,c) \equiv True_2(s,c)}$, giving:

(T!-) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset [\mathbf{True}_1(s,c) \equiv \mathbf{True}_2(s,c)]]$.

But one cannot drop the assumption and assert unconditionally:

(T!) $\forall s \forall c[\mathbf{True}_1(s,c) \equiv \mathbf{True}_2(s,c)]$.

Thus truth is not *the* value of **True** which satisfies (T), because there is no such unique value. Similarly, if (F_1) and (F_2) are the results of subscripting **False** in (F) accordingly, they entail:

(F!-) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset [\mathbf{False}_1(s,c) \equiv \mathbf{False}_2(s,c)]]$.

But one cannot drop the assumption and assert unconditionally:

(F!) $\forall s \forall c[\mathbf{False}_1(s,c) \equiv \mathbf{False}_2(s,c)]$.

Thus falsity is not *the* value of **False** which satisfies (F), because there is no such unique value. We can fill these gaps by explicitly stipulating that a sentence is true or false only if it says something:

(T+) $\forall s \forall c[\mathbf{True}(s,c) \supset \exists P\ \mathbf{Say}(s,c,P)]$

(F+) $\forall s \forall c[\mathbf{False}(s,c) \supset \exists P\ \mathbf{Say}(s,c,P)]$

With these additions, truth and falsity are uniquely characterized, in the sense that (T!) follows from (T_1), (T_2), (T_1+) and (T_2+) (the last two being the results of subscripting **True** in (T+) accordingly), and (F!) follows from (F_1), (F_2), (F_1+) and (F_2+). It is worth noting that (T+) and (F+) rely more heavily on quantification into sentence position than do (T) and (F). For one could drop the universal quantifiers from (T) and (F) and treat the remainders simply as schemata, whereas no such treatment is possible for the existential quantifiers in (T+) and (F+).

A deeper problem is that (T) and (F) are too strong to be mere definitions of **True** and **False**, because they are *creative*. That is, (T) has a new consequence not involving **True** and (F) has a new consequence not involving **False**. More specifically, each of them entails a principle of *uniformity* to the effect that everything said by a given sentence in a given context has the same truth-value:

(U) $\forall s \forall c \forall P \forall Q[[\mathbf{Say}(s,c,P)\ \&\ \mathbf{Say}(s,c,Q)] \supset [P \equiv Q]]$

For the antecedent of (U) yields **True(s,c)** ≡ **P** and **True(s,c)** ≡ **Q** by instances of (T), from which **P** ≡ **Q** follows. Similarly, the antecedent of (U) yields **False(s,c)** ≡ ~**P** and **False(s,c)** ≡ ~**Q** by instances of (F), from which ~**P** ≡ ~**Q** follows, and therefore **P** ≡ **Q** too.

Let L be a language without the predicates **True** and **False** but in which (U) can be formulated, and consider a theory Z in L of which (U) is not a theorem. If one adds **True** to L and (T) to Z, the result is not a conservative extension of Z, because (U) is a theorem of the new theory in the old language without being a theorem of the old theory. For the same reason, if one adds **False** to L and (F) to Z, the result is again not a conservative extension of Z. If (T) and (F) were genuine definitions, their addition would yield conservative extensions of Z. They behave more like theories than definitions.

A question now arises for the epistemicist derivation of the bivalence principles (SB) and (WB). It relies on (T) and (F), which are logically stronger than definitions of **True** and **False**. Thus the opponent of epistemicism might accept classical logic but block the derivation of (SB) and (WB) by rejecting (T) or (F) or both, while insisting that this does not amount to redefining **True** and **False**. In particular, the anti-epistemicist might reject (U), and therefore both (T) and (F), on the grounds that (U) fails in borderline cases for a vague sentence s. Perhaps they will suggest that a vague sentence says many things, corresponding to its different possible sharpenings; in a borderline case, some of these things differ from others in truth-value, contrary to (U).

I will postpone asking whether (U) really does fail in borderline cases, and suppose for the sake of argument that it does. If so, what happens to the bivalence principles (SB) and (WB)? In order to answer that question, one needs characterizations of truth and falsity. (T) and (F) will not do for present purposes because they entail (U), which is being supposed to fail. One must therefore characterize truth and falsity in some alternative way. Within the present framework, the natural idea is to define truth and falsity explicitly by quantifying into sentence position.

A standard proposal is that a sentence is true if something that it says to be so is so. The biconditional corresponding to such a definition is this:

(TDEF1) $\forall s \forall c[\mathbf{True}(s,c) \equiv \exists P[\mathbf{Say}(s,c,P) \ \&\ P]]$

The obvious analogue for falsity of such a definition is that a sentence is false if something that it says to be so is not so. The corresponding biconditional stands to (F) as (TDEF1) stands to (T):

(FDEF1) $\forall s \forall c[\mathbf{False}(s,c) \equiv \exists P[\mathbf{Say}(s,c,P) \ \&\ \sim P]]$

Like (T+) and (F+), both (TDEF1) and (FDEF1) make much more essential use of quantification into sentence position than do (T) and (F). Schemata without such quantification are no substitute for (TDEF1) and (FDEF1).

Before continuing with the main line of argument, I pause to consider an objection. By explicitly defining truth, for instance by (TDEF1), are we not liable to incur inconsistency, by Tarski's theorem on the undefinability of truth? Fortunately, we can avoid this danger by not permitting the unrestricted application of the expression **Say** to sentences of the metalanguage as values of the term in its first argument-place (**s**). This still permits **Say** to be applied to sentences of many diverse languages, provided that their content is not of a special metalinguistic kind. Such a restriction requires extensive discussion, but our present concern is not with the semantic paradoxes. Quantification into sentence position does not itself introduce any inconsistency. As noted above, one can give it a consistency proof by using a substitutional interpretation. Indeed, one could give an alternative consistency proof by using a different unintended interpetation in which the semantic value of the sentential variable **P** is simply its truth-value, although that would not permit **Say** to behave in the intended way, as an intensional operator with respect to **P**. One could accommodate some of that intensionality by treating the semantic value of **P** as the set of possible worlds in which it is true, although even that would not capture its full intended meaning, since we take **Say** to behave *hyper*intensionally: a sentence may say that P without saying that Q even though in all possible worlds P if and only if Q. For example, '2+2=4' says that 2+2=4 without saying that 7+5=12, even though all mathematical truths are true in all possible worlds and therefore in the same possible worlds. Given the limited nature of the present use of quantification into sentence position, as noted above, we can consistently treat the semantic values of variables in sentence position as structured or otherwise finely grained propositions, thereby allowing for hyperintensionality. We will not pursue these issues here, but simply repeat that none of the logical framework within which we are operating falls to a semantic paradox.

Since (TDEF1) and (FDEF1) are in effect definitions, they do not entail (U). Therefore, they do not entail (T) or (F). More specifically, one can easily check that (TDEF1) entails (T$\leftarrow$), and therefore does not by itself entail (T$\rightarrow$) (for if it did, it would entail (T)), and that (FDEF1) entails (F$\leftarrow$), and therefore does not by itself entail (F$\rightarrow$) (for if it did, it would entail (F)).

We can check that (TDEF1) and (U) together do entail (T$\rightarrow$), and therefore (T), by an obvious argument. Similarly, (FDEF1) and (U) together entail (F$\rightarrow$), and therefore (F). This shows that (U) is the *strongest*

consequence of the conjunction of (T) and (F) to contain neither **True** nor **False**, in the sense that if A contains neither **True** nor **False** and (T) and (F) together entail A then (U) entails A. For (U), (TDEF1) and (FDEF1) together entail (T) and (F); thus if (T) and (F) entail A, then by transitivity (U), (TDEF1) and (FDEF1) entail A; but if one replaces **True** and **False** throughout the deduction by the right-hand sides of (TDEF1) and (FDEF1) respectively, the result trivializes (TDEF1) and (FDEF1) and is still a valid deduction of A from (U), since the premises and conclusion do not contain **True** or **False**. Thus (U) exhausts the non-conservative aspect of (T) and (F).

What are the implications of (TDEF1) and (FDEF1) for bivalence? As already noted, they entail (T←) and (F←) respectively, which together entail (WB), so (TDEF1) and (FDEF1) jointly yield weak bivalence. But they do not yield the strong bivalence principle (SB), because they do not yield (ME). In fact, one can easily show that if any sentence is a counter-example to (U), then (TDEF1) counts it as true and (FDEF1) counts it as false. Thus if (U) fails for vague sentences in borderline cases, (TDEF1) and (FDEF1) make those sentences both true and false. Such a consequence would generally be regarded as unacceptable.

One or two theorists of vagueness do suggest that a vague sentence is true if and only if at least one of its admissible sharpenings is true, and is false if and only if at least one of its admissible sharpenings is false (Hyde 1997). Such a view is know as *subvaluationism*, by analogy with supervaluationism (discussed below). Subvaluationism resembles the envisaged combination of (TDEF1) and (FDEF1) with the negation of (U). The various things said by a sentence correspond to its various sharpenings. However, subvaluationism has highly counter-intuitive consequences. Consider, for example, 'small' as a vague predicate of natural numbers in a context in which one admissible sharpening puts its cut-off point between n-1 and n while another equally admissible sharpening puts the cut-off point between n and n+1. Then the conjunctions 'n-1 is small and n is not small' and 'n is small and n+1 is not small' both count as true according to subvaluationism, because each is true on at least one sharpening. Moreover, since we might know that both sharpenings were admissible, we might know that both conjunctions were true. Presumably, therefore, we should assert each of them if the question arises. But we are not entitled to assert (rather than stipulate) of any place in particular that the cut-off point for a vague predicate comes there. Still less are we entitled to assert both conjunctions, for they are mutually inconsistent; not even subvaluationism makes their conjunction 'n-1 is small and n is not small and n is small and n+1 is not small' true. Consequently, one should reject subvaluationism. It faces all the main problems of supervaluationism and more besides. In what follows I assume the principle (ME) that truth and falsity are mutually exclusive.

Thus, in the absence of (U), I reject the combination of (TDEF1) and (FDEF1).

One could preserve both (ME) and either one of (TDEF1) and (FDEF1) while rejecting (U) by treating truth and falsity as contradictories. For example, suppose that one accepts (TDEF1). Then to treat falsity as non-truth is to accept the dual of (FDEF1):

(FDEF2) $\forall s \forall c[\mathbf{False}(s,c) \equiv \forall P[\mathbf{Say}(s,c,P) \supset \sim P]]$

A sentence is false if nothing that it says to be so is so. (TDEF1) and (FDEF2) jointly entail the full bivalence principle (SB) itself, and therefore both (ME) and (WB). But since they merely define **True** and **False**, they do not entail (U). Thus they do not jointly entail (T) or (F). In particular, (FDEF2) does not entail (F): although it entails (F$\rightarrow$), it does not entail (F$\leftarrow$). In any context in which (U) fails, the relevant sentence counts as true by (TDEF1) and as not false by (FDEF2). Therefore, if (U) fails for all vague sentences in borderline cases, all such sentences count as true and not false —and one can know that. Thus the combination of (TDEF1), (FDEF2) and the negation of (U) represents a route to strong bivalence incompatible with that taken by the epistemicist.

A similar route to strong bivalence equally inconsistent with epistemicism combines (FDEF1) with the negation of (U) and a definition of truth as non-falsity. It yields the dual of (TDEF1):

(TDEF2) $\forall s \forall c[\mathbf{True}(s,c) \equiv \forall P[\mathbf{Say}(s,c,P) \supset P]]$

A sentence is true if everything that it says to be so is so. (TDEF2) and (FDEF1) jointly entail (SB), and therefore both (ME) and (WB). Since they merely define **True** and **False**, they do not entail (U) or (T) or (F). In particular, (TDEF2) does not entail (T): although it entails (T$\rightarrow$), it does not entail (T$\leftarrow$). In any context in which (U) fails, the relevant sentence counts as not true by (TDEF2) and as false by (FDEF2). Therefore, if (U) fails for vague sentences in borderline cases, such sentences count as false and not true —and one can know that.

Although the two combinations just considered secure strong bivalence, they do so at a high cost to classical intuitions. They undermine the natural conception of truth and falsity as somehow parallel notions: one of the defining conditions is quantified universally, the other existentially. Moreover, the presence of just one of (TDEF1) and (FDEF1) suffices to yield some of the counter-intuitive consequences of subvaluationism. For example, given the persuasive principle (N) linking what a sentence says to what its negation says, one can show that in any context in which a sentence

is a counter-example to (U), so is its negation: from **Say(s,c,P) & Say(s,c,Q)** one can derive **Say(Ns,c,~P) & Say(Ns,c,~Q)**. Now suppose that (U) fails in a borderline case *n* for the vague predicate 'small'. Then (TDEF1) counts both '*n* is small' and '*n* is not small' as true. That (FDEF2) does not count '*n* is small' as false too is not much consolation, for that is achieved only by not counting a sentence with a true negation as false; (FN) is violated from right to left. Similarly, (FDEF1) counts both '*n* is small' and '*n* is not small' as false. That (TDEF2) does not count '*n* is not small' as true too is equally little consolation, for that is achieved only by not counting the negation of a false sentence as true; (FN) is violated from left to right. Both (TDEF1) and (FDEF1) are to be rejected. What remains is the combination of (TDEF2) and (FDEF2).

Given (TDEF2) and (FDEF2), failures of (U) produce truth-value gaps rather than truth-value gluts. If vague sentences in borderline cases are counter-examples to (U), then such sentences are neither true nor false. This view has an obvious analogy with supervaluationism, on which a sentence is true if and only if all its admissible sharpenings are true, and is false if and only if all its admissible sharpenings are false. As with subvaluationism, the various things said by a sentence would correspond to its various sharpenings.

An immediate problem is that (TDEF2) and (FDEF2) together do not make truth and falsity mutually exclusive, even in the presence of (U). If a sentence says nothing, they count it as vacuously both true and false. Indeed, they count mountains and lakes as both true and false. To overcome this problem, one can add clauses requiring that a sentence is true or false only if it says something:

(TDEF2*) $\forall s \forall c[\mathbf{True(s,c)} \equiv [\exists P\ \mathbf{Say(s,c,P)} \ \&\ \forall P[\mathbf{Say(s,c,P)} \supset P]]]$

(FDEF2*) $\forall s \forall c[\mathbf{False(s,c)} \equiv [\exists P\ \mathbf{Say(s,c,P)} \ \&\ \forall P[\mathbf{Say(s,c,P)} \supset \sim P]]]$

This modification preserves the analogy with supervaluationism; the latter does not permit sentences to be vacuously both true and false, for admissible sharpenings are defined at the level of a language as a whole and supervaluationists assume that there is always at least one such sharpening. (TDEF2*) and (FDEF2*) jointly entail (ME), (T→) and (F→); they do not entail (U), (T), (T←), (F), (F←), (WB) or (SB). Since counter-examples to (U) must satisfy the extra clause $\exists P$ **Say(s,c,P)**, they still count as neither true nor false by (TDEF2*) and (FDEF2*). If borderline cases falsify (U), they are classified as involving truth-value gaps.

6. THE CASE FOR THE UNIFORMITY PRINCIPLE

In the preceding discussion I assumed, for the sake of argument, that (U) can fail: a sentence may in the same context say something that is so and something that is not so. If so, the supposed failure appears to give succour to supervaluationism. It is time to evaluate the assumption that (U) can fail.

There is a non-technical notion of saying on which to say something can also be to say some of its immediate logical consequences. In that sense, saying that it is cold and wet might involve both saying that it is cold and saying that it is wet; saying that Nenad is absent might involve saying that someone is absent. That notion of saying is clearly irrelevant to the present problem, for it would yield counter-examples to (U) even in unproblematic non-borderline cases. For example, perhaps it is cold but not wet; perhaps Nenad is present but someone else is absent. Since the relevant instances of (T) and (F) should hold in such unproblematic cases, so should their consequences, the relevant instances of (U). Thus we should interpret **Say(s,c,P)** in (T), (F) and (U) to mean something like: s says in c *just* that P. But then how can (U) fail? If, in a given context, a sentence says *just* that P and says *just* that Q, how could the proposition that P be anything other than the proposition that Q? Similarly, if we read **Say(s,c,P)** as something like 'The propositional content of s in c is that P', then the uniqueness implied by the definite article leaves no obvious room for (U) to fail.

Someone might intentionally use physically the same words to say each of two things simultaneously. But it is not *ad hoc* to treat the two intentions as creating two simultaneous contexts of utterance for that sentence, so (U) is still not falsified (see Andjelković and Williamson 2001 for detailed discussion).

Suppose that Harry is a borderline case for baldness. It is natural to hold that the sentence 'Harry is bald' says in the present context that Harry is bald. But, if so, why think that it says something else too? *What* else does it say? If the sentence does not say that Harry is bald, the question arises again: what does it say? 'Harry is bald' seems to say just that Harry is bald. Even the semantic paradoxes do not seem to be counter-examples to the qualified disquotational principle that *if* the sentence «P» says anything at all in this context, it says just that P. Thus it is hard to understand how (U) could fail in the present context. But if it cannot fail in the present context, then it cannot fail in any ordinary context, for there is nothing special about the present context. Since borderline cases such as Harry occur in the present context, vagueness would not falsify (U).

Someone might use the supervaluationist or subvaluationist notion of an admissible sharpening in an attempt to explain a sense of **Say** on which (U) could fail: a sentence would say its admissible sharpenings. Such an

explanation is only as clear as the notion of an admissible sharpening. Supervaluationists have great difficulty in giving an adequate account of that notion in such a way that it does not reduce to something epistemic (Williamson 1994: 164, 1995 and 1997: 216-217, the last of which responds to McGee and McLaughlin 1995).

Fortunately, the uniformity principle (U) can be supported by arguments more rigorous than the foregoing remarks. The strategy is as follows. We start with (TDEF2*), a definition of truth of the sort that someone who rejects (U) might accept, and show that under plausible assumptions about compositionality it still leads to (U). Since (TDEF2*) makes **True** behave in a distinctively supervaluationist fashion only if (U) fails, the appearance of support lent by (TDEF2*) to supervaluationism is illusory. I will also show that under plausible assumptions about compositionality (TDEF1) too leads to (U). Thus the appearance of support lent by (TDEF1) to subvaluationism is equally illusory.

The argument assumes that sentences s and t can be connected by a material biconditional into a sentence Est (≡ is the material biconditional in the metalanguage). Although the argument does not apply directly to sentences in a language without such a connective, that is not a serious limitation. For a language with a material biconditional may contain arbitrarily vague sentences, or sentences with any other features that might be thought to undermine (U). Thus if (U) holds for languages with a material biconditional, the relevant theories that predict the failure of (U) are false, and there is no longer any reason to reject (U).

I will now argue from (TDEF2*) to (U). I first note a plausible principle of compositional semantics. Roughly, since E expresses ≡, Est says the biconditional of what s says and what t says:

(E1) $\forall s \forall t \forall c \forall P \forall Q[[\mathbf{Say}(s,c,P) \,\&\, \mathbf{Say}(t,c,Q)] \supset \mathbf{Say}(Est,c,P \equiv Q)]]$

Next, note that if a sentence says something then its biconditional with itself is true:

(E2) $\forall s \forall c \forall P[\mathbf{Say}(s,c,P) \supset \mathbf{True}(Ess,c)]$

Supervaluationists would certainly accept (E2), since every sharpening of Ess is true; it is a classical tautology. A special case of (E1) is:

(1) $\forall s \forall c \forall P \forall Q[[\mathbf{Say}(s,c,P) \,\&\, \mathbf{Say}(s,c,Q)] \supset \mathbf{Say}(Ess,c,P \equiv Q)]]$

Next, recall that (TDEF2*) yields (T→), the half of (T) acceptable to supervaluationists. A special case of (T→) is:

(2) $\forall s \forall c \forall P \forall Q[\mathbf{Say}(Ess,c,P \equiv Q) \supset [\mathbf{True}(Ess,c) \supset [P \equiv Q]]]$

From (1) and (2) one has:

(3) $\forall s \forall c \forall P \forall Q[[\mathbf{Say}(s,c,P) \ \& \ \mathbf{Say}(s,c,Q)] \supset [\mathbf{True}(Ess,c) \supset [P \equiv Q]]]$

But (E2) allows one to discharge the condition **True(Ess,c)** from (3) and thereby derive (U). Once one has (U), one can easily recover (T) from (TDEF2*) and (F) from (FDEF2*). The bivalence principles (SB) and (WB) then follow as before from (T) and (F).

Conversely, one can explain (E2) on general grounds by deriving it from (T) and (E1). For an instance of (T) is $\mathbf{Say}(Ess,c,P \equiv P) \supset [\mathbf{True}(Ess,c) \equiv [P \equiv P]]$, which yields $\mathbf{Say}(Ess,c,P \equiv P) \supset \mathbf{True}(Ess,c)$, while an instance of (E1) is $\mathbf{Say}(s,c,P) \supset \mathbf{Say}(Ess,c,P \equiv P)$; together, these give (E2).

Although (TDEF2*) and (FDEF2*) appear to invite a supervaluationist treatment of vagueness, they do not really do so. Such a treatment would involve the denial of (U). The foregoing argument shows that that in turn would require the supervaluationist to deny (E1). But that is too high a price to pay, for it destroys our conception of what biconditionals say.

There is a similar argument for (U) from (TDEF1), the truth definition with an apparently subvaluationist flavour. The argument also assumes that each relevant sentence s has a negation Ns; this assumption is harmless in the present dialectical situation for a reason just like that already given in relation to the assumption that the relevant sentences are closed under the biconditional function E. In place of (E2), this argument needs:

(E2*) $\forall s \forall c \sim \mathbf{True}(NEss,c)$

(E2*) does not need the condition that NEss says something, for if it says nothing it is certainly not true. Subvaluationists should accept (E2*), since no admissible sharpening of NEss is true; it is a classical contradiction. As before, one derives (1) from (E1). One then applies (N) to (1) to reach:

(1*) $\forall s \forall c \forall P \forall Q[[\mathbf{Say}(s,c,P) \ \& \ \mathbf{Say}(s,c,Q)] \supset \mathbf{Say}(NEss,c,\sim[P \equiv Q])]$

Next, recall that (TDEF1) yields (T←), the half of (T) acceptable to subvaluationists. A special case of (T←) is:

(2*) $\forall s \forall c \forall P \forall Q[\mathbf{Say}(NEss,c,\sim[P \equiv Q]) \supset [\sim[P \equiv Q] \supset \mathbf{True}(NEss,c)]]$

From (1*) and (2*) one has:

(3*)$\forall s \forall c \forall P \forall Q[[\mathbf{Say(s,c,P)} \ \& \ \mathbf{Say(s,c,Q)}] \supset [\sim[P \equiv Q] \supset \mathbf{True(NEss,c)}]]$

But (E2*) allows one to apply *modus tollens* to the final conditional in (3*) and thereby derive (U). Once one has (U), one can easily recover (T) from (TDEF1) and (F) from (FDEF1). The bivalence principles (SB) and (WB) then follow as usual from (T) and (F). Although (TDEF1) and (FDEF1) appear to invite a subvaluationist treatment of vagueness, they do not really do so. Such a treatment would involve the denial of (U). The argument shows that that would require the subvaluationist to deny (E1). As before, that is too high a price to pay.

The crucial premise in both arguments for (U) is (E1). It articulates the natural way in which a biconditional sentence says something when its constituent sentences say something. One might come to doubt (E1) by equating saying something with not clearly not saying it. This can be formalized in terms of the 'clearly' operator Δ. Given that the antecedent of (E1) is clearly sufficient for its consequent, one has:

(E1Δ) $\forall s \forall t \forall c \forall P \forall Q \ \Delta[[\mathbf{Say(s,c,P)} \ \& \ \mathbf{Say(t,c,Q)}] \supset \mathbf{Say(Est,c,P \equiv Q)}]]$

But from (E1Δ) one cannot infer:

(E1$\Delta\Delta$) $\forall s \forall t \forall c \forall P \forall Q[[\sim\Delta\sim\mathbf{Say(s,c,P)} \ \& \ \sim\Delta\sim\mathbf{Say(t,c,Q)}] \supset \sim\Delta\sim\mathbf{Say(Est,c,P \equiv Q)}]]$

Similarly, when $\Box$ and $\Diamond$ mean necessity and possibility respectively, $\Box[[\mathbf{P} \ \& \ \mathbf{Q}] \supset \mathbf{R}]$ does not generally entail $[\Diamond\mathbf{P} \ \& \ \Diamond\mathbf{Q}] \supset \Diamond\mathbf{R}$. For example, if **P** expresses a contingency then $\Box[[\mathbf{P} \ \& \sim\mathbf{P}] \supset [\mathbf{P} \ \& \sim\mathbf{P}]]$ is trivially true while $[\Diamond\mathbf{P} \ \& \ \Diamond\sim\mathbf{P}] \supset \Diamond[\mathbf{P} \ \& \sim\mathbf{P}]$ is false. For analogous reasons, if it is not clear that s does not say that P and not clear that s does not say that Q, it does not follow that it is not clear that Ess does not say that P if and only if Q. It may be clear that Ess does not say that P if and only if Q just because it is clear that P and not Q. But if it is clear that Ess does not say that P if and only if Q, then that is so because it is clear that s does not *both* say that P *and* say that Q. Thus (E1) and (E1Δ) themselves are unthreatened. Saying something should not be confused with the more complex matter of not clearly not saying it. There is a natural and theoretically central notion of saying which satisfies (E1), (E1Δ) and (U). That is the notion needed in an account of truth and falsity. If one so wishes, one can then introduce a secondary notion of quasi-saying by the formula $\sim\Delta\sim\mathbf{Say(s,c,P)}$, but quasi-saying something does not amount to saying it. One might for a moment be tempted to suppose that quasi-saying is a more precise notion than saying, because it takes vagueness explicitly into account with its use of the clarity operator. However, the phenomenon of higher-order vagueness implies that

quasi-saying is a vague notion too, just like saying. Unlike saying, quasi-saying has no special theoretical significance.

We have examined attempts to reject the uniformity principle (U) and therefore the principles (T) and (F) about truth and falsity in favour of definitions of truth and falsity which, in the spirit of supervaluationism or subvaluationism, yield only one or other half of (T) and (F). Those attempts violate independently plausible constraints of semantic compositionality. Once the uniformity principle is restored to its rightful place, those definitions yield (T) and (F) after all. Thus, given classical logic, they also yield the principle (SB) of strong bivalence. The apparent loophole in the classical case for bivalence is therefore merely apparent.

ACKNOWLEDGEMENT

This paper draws extensively on my contribution to Andjelković and Williamson 2001. I am very grateful to Miroslava Andjelković for her part in that joint project. That paper contains amplifications of some of the arguments in this one.

REFERENCES

Andjelković, Miroslava. (1999). 'Williamson on bivalence'. *Acta Analytica* 14 (issue 23): 27-33.

Andjelković, Miroslava, and Williamson, Timothy (2000). 'Truth, falsity and borderline cases'. *Philosophical Topics* 28: 211-244.

Fine, Kit. (1975). 'Vagueness, truth and logic'. *Synthese* 30: 265-300. Reprinted in Keefe and Smith (1996).

Hyde, Dominic. (1997). 'From heaps and gaps to heaps of gluts'. *Mind* 106: 440-460.

Keefe, Rosanna, and Smith, Peter, eds. (1996). *Vagueness: A Reader*. Cambridge, Mass. and London: MIT Press.

McGee, Vann, and McLaughlin, Brian. (1995). 'Distinctions without a difference'. *Southern Journal of Philosophy* 33 (supplement): 203-251.

Williamson, Timothy. (1994). *Vagueness*. London and New York: Routledge.

Williamson, Timothy. (1995). 'Definiteness and knowability'. *Southern Journal of Philosophy* 33 (supplement): 171-191.

Williamson, Timothy. (1997). 'Imagination, stipulation and vagueness'. In E. Villanueva, ed., *Philosophical Issues 8: Truth*. Atascadero CA: Ridgeview.

Williamson, Timothy. (1998). 'Indefinite extensibility'. *Grazer Philosophische Studien* 55: 1-24.

Williamson, Timothy. (1999). 'Andjelković on bivalence: a reply'. *Acta Analytica* 14 (issue 23): 35-8.

Chapter 4

MEANING FINITISM AND TRUTH[1]

Martin Kusch
Cambridge University

1.

'Meaning finitism' is a theory of linguistic meaning that has been developed since the late seventies by Barry Barnes and David Bloor.[2] Barnes and Bloor are the two founding fathers of the 'Edinburgh School' in the Sociology of Scientific Knowledge. In this paper I seek to reconstruct and defend meaning finitism by exploring its consequences for our understanding of truth.[3] Given the constraints of time, I will here only be able to scratch the surface of the

[1] A much earlier version of this paper was discussed at the Moral Sciences Club in Cambridge in 1998. I am grateful to Anjan Chakravartty, Michael Esfeld, Anandi Hattiangadi, Jane Heal, Susan James, Tim Lewens, Peter Lipton and Hugh Mellor for many objections and suggestions on that occasion. I have also profited from comments by David Bloor, David Chart, Jeremy Gray, Matthias Hild, Matthew Ratcliffe and Paul Teller.

[2] The best general accounts are Barry Barnes, David Bloor, John Henry, *Scientific Knowledge: A Sociological Analysis* (London: Athlone Press, 1996), Chapter Three; and D. Bloor, *Wittgenstein, Rules and Institutions* (London: Routledge, 1997). See also B. Barnes, 'On the Extensions of Concepts and the Growth of Knowledge,' *Sociological Review* 30 (1982), 23-44.

[3] I have used and developed Barnes' and Bloor's views on meaning in my earlier work. See M. Kusch, *Psychological Knowledge: A Social History and Philosophy* (London: Routledge, 1999); Harry Collins and M. Kusch, *The Shape of Actions: What Humans and Machines Can Do* (Boston: M.I.T. Press, 1998).

K. Korta and J. M. Larrazabal (eds.), Truth, Rationality, Cognition, and Music, 67-83.

quasi-saying is a vague notion too, just like saying. Unlike saying, quasi-saying has no special theoretical significance.

We have examined attempts to reject the uniformity principle (U) and therefore the principles (T) and (F) about truth and falsity in favour of definitions of truth and falsity which, in the spirit of supervaluationism or subvaluationism, yield only one or other half of (T) and (F). Those attempts violate independently plausible constraints of semantic compositionality. Once the uniformity principle is restored to its rightful place, those definitions yield (T) and (F) after all. Thus, given classical logic, they also yield the principle (SB) of strong bivalence. The apparent loophole in the classical case for bivalence is therefore merely apparent.

ACKNOWLEDGEMENT

This paper draws extensively on my contribution to Andjelković and Williamson 2001. I am very grateful to Miroslava Andjelković for her part in that joint project. That paper contains amplifications of some of the arguments in this one.

REFERENCES

Andjelković, Miroslava. (1999). 'Williamson on bivalence'. *Acta Analytica* 14 (issue 23): 27-33.

Andjelković, Miroslava, and Williamson, Timothy (2000). 'Truth, falsity and borderline cases'. *Philosophical Topics* 28: 211-244.

Fine, Kit. (1975). 'Vagueness, truth and logic'. *Synthese* 30: 265-300. Reprinted in Keefe and Smith (1996).

Hyde, Dominic. (1997). 'From heaps and gaps to heaps of gluts'. *Mind* 106: 440-460.

Keefe, Rosanna, and Smith, Peter, eds. (1996). *Vagueness: A Reader*. Cambridge, Mass. and London: MIT Press.

McGee, Vann, and McLaughlin, Brian. (1995). 'Distinctions without a difference'. *Southern Journal of Philosophy* 33 (supplement): 203-251.

Williamson, Timothy. (1994). *Vagueness*. London and New York: Routledge.

Williamson, Timothy. (1995). 'Definiteness and knowability'. *Southern Journal of Philosophy* 33 (supplement): 171-191.

Williamson, Timothy. (1997). 'Imagination, stipulation and vagueness'. In E. Villanueva, ed., *Philosophical Issues 8: Truth*. Atascadero CA: Ridgeview.

Williamson, Timothy. (1998). 'Indefinite extensibility'. *Grazer Philosophische Studien* 55: 1-24.

Williamson, Timothy. (1999). 'Andjelković on bivalence: a reply'. *Acta Analytica* 14 (issue 23): 35-8.