1

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file - Votre référence

Our file - Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

MEMES, PARFIT, AND THE CONSTRUCTION OF THE SELF

by

TIMOTHY KENYON, B.A.

**A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of**

Master of Arts

Department of Philosophy

**Carleton University
Ottawa, Ontario
December 6, 1993
c copyright
1993, Timothy Kenyon**

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-89838-0

Canada

Name

Timothy A. Kienzer

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

Philosophy

SUBJECT TERM

0422

SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION

General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523

Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language 0679
General 0289
Ancient 0290
Linguistics 0291
Modern 0291
Literature 0401
General 0294
Classical 0295
Comparative 0297
Medieval 0298
Modern 0316
African 0591
American 0305
Asian 0352
Canadian (English) 0355
Canadian (French) 0593
English 0311
Germanic 0312
Latin American 0315
Middle Eastern 0313
Romance 0314
Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy 0422
Religion 0318
General 0321
Biblical Studies 0319
Clergy 0320
History of 0322
Philosophy of 0469
Theology 0469

SOCIAL SCIENCES

American Studies 0323
Anthropology 0324
Archaeology 0326
Cultural 0327
Physical 0310
Business Administration 0272
General 0770
Accounting 0454
Banking 0338
Management 0385
Marketing 0501
Canadian Studies 0503
Economics 0505
General 0508
Agricultural 0509
Commerce-Business 0510
Finance 0511
History 0358
Labor 0366
Theory 0351
Folklore 0578
Geography 0578
Gerontology 0578
History 0578
General 0578

Ancient 0579
Medieval 0581
Modern 0582
Black 0328
African 0331
Asia, Australia and Oceania 0332
Canadian 0334
European 0335
Latin American 0336
Middle Eastern 0333
United States 0337
History of Science 0585
Law 0398
Political Science 0615
General 0616
International Law and Relations 0617
Public Administration 0814
Recreation 0452
Social Work 0626
Sociology 0627
General 0938
Criminology and Penology 0531
Demography 0628
Ethnic and Racial Studies 0629
Individual and Family Studies 0630
Industrial and Labor Relations 0700
Public and Social Welfare 0344
Social Structure and Development 0709
Theory and Methods 0999
Transportation 0453
Urban and Regional Planning 0453
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture 0473
General 0285
Agronomy 0475
Animal Culture and Nutrition 0476
Animal Pathology 0359
Food Science and Technology 0478
Forestry and Wildlife 0479
Plant Culture 0480
Plant Pathology 0817
Plant Physiology 0777
Range Management 0746
Wood Technology 0306
Biology 0287
General 0308
Anatomy 0309
Biostatistics 0379
Botany 0329
Cell 0353
Ecology 0369
Entomology 0793
Genetics 0410
Limnology 0307
Microbiology 0317
Molecular 0416
Neuroscience 0433
Oceanography 0821
Physiology 0778
Radiation 0472
Veterinary Science 0786
Zoology 0760
Biophysics 0786
General 0760
Medical 0760

EARTH SCIENCES

Biogeochemistry 0425
Geochemistry 0996

Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Palaeobotany 0345
Palaeocology 0426
Palaeontology 0418
Palaeozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences 0768
Health Sciences 0566
General 0300
Audiology 0992
Chemotherapy 0567
Dentistry 0350
Education 0769
Hospital Management 0758
Human Development 0982
Immunology 0564
Medicine and Surgery 0347
Mental Health 0569
Nursing 0570
Nutrition 0380
Obstetrics and Gynecology 0354
Occupational Health and Therapy 0381
Ophthalmology 0571
Pathology 0419
Pharmacology 0572
Pharmacy 0382
Physical Therapy 0573
Public Health 0574
Radiology 0575
Recreation 0575

Speech Pathology 0460
Toxicology 0383
Home Economics 0386

PHYSICAL SCIENCES

Pure Sciences 0485
Chemistry 0749
General 0486
Agricultural 0487
Analytical 0488
Biochemistry 0738
Inorganic 0490
Nuclear 0491
Organic 0494
Pharmaceutical 0495
Physical 0754
Polymer 0405
Radiation 0605
Mathematics 0986
Physics 0606
General 0608
Acoustics 0748
Astronomy and Astrophysics 0607
Atmospheric Science 0798
Atomic 0759
Electronics and Electricity 0609
Elementary Particles and High Energy 0610
Fluid and Plasma 0752
Molecular 0756
Nuclear 0611
Optics 0463
Radiation 0346
Solid State 0984
Statistics 0984

Applied Sciences

Applied Mechanics 0346
Computer Science 0984

Engineering 0537
General 0538
Aerospace 0539
Agricultural 0540
Automotive 0541
Biomedical 0542
Chemical 0543
Civil 0544
Electronics and Electrical 0348
Heat and Thermodynamics 0545
Hydraulic 0546
Industrial 0547
Marine 0774
Materials Science 0548
Mechanical 0743
Metallurgy 0551
Mining 0552
Nuclear 0549
Packaging 0765
Petroleum 0554
Sanitary and Municipal 0790
System Science 0428
Geotechnology 0796
Operations Research 0795
Plastics Technology 0994
Textile Technology 0994

PSYCHOLOGY

General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

The undersigned recommend to the Faculty of Graduate Studies
and Research acceptance of the thesis

"MEMES, PARFIT, AND THE CONSTRUCTION OF THE SELF"

submitted by Timothy Kenyon
in partial fulfillment of the requirements for
the degree of Master of Arts

Thesis Supervisor

Chair, Department of Philosophy

CARLETON UNIVERSITY

December 20, 1993

ABSTRACT

Richard Dawkins introduces the concept of a *meme* as a unit of behavioral imitation in order to explain patterns of replicating behavior. I take memes to be mental states, and adapt Daniel Dennett's "intentional stance" argument that propositional attitudes are abstractions to apply to the broader class of memes. After examining the constraints that govern the ascription of memes, I summarize Dennett's theory of the self, which presents the self as a narrative comprising a series of memes. I then contend that the meme theory of the self is particularly relevant to two specific issues: namely, Derek Parfit's theory of personal identity, and the questions surrounding Multiple Personality Disorder. The meme theory underwrites Parfit's claims about the roles of memory and continuity in engendering judgements of identity, and moreover renders unproblematic the appearance of multiple selves in one body, such as is found in Multiple Personality Disorder.

Acknowledgments

I owe my deepest thanks to Colleen. As meme constellations go, surely she is without equal.

Both of our families have steadfastly supported my efforts even when they didn't get a word I said. A round of thanks.

Robert Stainton and Christopher Viger provided invaluable comments and criticisms of earlier drafts. Their efforts are greatly appreciated.

Without the good sense of J.A. Brook this project would never have begun, and without his good humour it would never have been completed. He had a hand in anything excellent about this thesis, and strove mightily against the inclusion of all else.

Tim Kenyon
12 22 93

Table of Contents

ABSTRACT	ii
Acknowledgments	iii
Table of Contents	iv
Chapter One: Telling a Meme from a Hawk and a Handsaw	1
I. Introduction	1
i. The need for a meme	3
ii. Locating memes in the mental schema	5
II. What Memes Are Not	8
i. Type identity	8
ii. Token identity	10
a. Neural facts and representation	10
b. Frames of reference, function, and generalization	14
c. Inside the argument from misrepresentation	17
iii. Behavior	20
III. What Memes Are	22
i. The Ontological Answer	22
ii. The Ascriptive Answer	26
iii. Conclusions	29
Chapter Two: Dennett and Consciousness	31
I. The Cartesian Theater	32

II. Multiple Drafts	35
i. Introduction to the model	36
ii. A digression for the purpose of ungarbling	37
III. Evolution, Serial Consciousness, and Parallel Brains	41
IV. User Illusions and Virtual Captains	48
i. You're only fooling yourself	48
ii. Meme, meme, meme and me, me, me	50
 Chapter Three: Parfit and Identity	 52
I. Deeper Facts, Revisited	52
i. The problem	52
ii. Parfit's solution	54
II. Memory, Continuity, and Similarity	55
III. The Answer is: No Answer	59
 Chapter Four: The Meme Theory of the Self	 61
I. Description of the Theory	62
i. Memory, memery and virtual captains	62
ii. The self as a constellation	67
II. Continuity, Causality, and Narrative	69
III. Ironing Out Some Wrinkles	73
i. From whence the self?	73
ii. Narrative and agency	74
IV. MPD as a Test	75
i. Braude and apperceptive unity	77
ii. Hacking and evidence	80
iii. Meme theory and MPD	84
 Conclusions	 88

i. What has been done?	88
ii. The construction of the self	89
Bibliography	91

Chapter One: Telling a Meme from a Hawk and a Handsaw

I. Introduction

What follows may be thematically divided into two parts, the plausibility of the second more or less dependent upon the success of the first. Had I been inclined to partition this essay along such thematic lines, the two parts might have been colloquially labelled: "Here's What a Meme Is;" and "Here's What We Can Do with Memes." The agenda for the former theme is more or less clear; I plan to explore and sharpen the definition of a class - an overlooked class, I think - of mental entities. This endeavor comprises the first chapter. The second theme revolves around the concepts of selfhood and identity, and particularly diachronic identity. In the remaining three chapters, I first outline a *meme theory of the self*, and then show how this theory situates and enriches Derek Parfit's views on personal identity. This aspect of my project requires that I exegetize the relevant contributions of Parfit and Daniel Dennett. Finally, I discuss the implications of a particular psychological disorder for the meme theory of the self.

As regards the first theme of my project, the task of providing necessary and sufficient conditions for a meme has the potential to engender as substantial a literature as currently exists on propositional attitudes. Offering suggestions on how to approach this task is, I think, as much as an introductory treatment need accomplish. My contention is that an approach of simply listing criteria is hopeless for this sort of inquiry, so I employ a variety of thought experiments designed to rule out certain *prima facie* possibilities for what a meme might be, and thus draw a rough line around the appropriate class of mental entities.

As regards the second theme, notions of "the disintegration of the subject" and "the self as narrative" have recently found increasingly frequent expression in philosophy and literary criticism, especially in Continental varieties of those fields. I am satisfied to *outline an approach* under which such ideas, having been popularized by rhetoric in some circles, may be supported by argumentation.

Much of what follows concerns what Dennett says - what he says about propositional attitudes, verificationism, evolution, consciousness, language, and Multiple Personality Disorder.¹ While I believe the central components of Dennett's approach to be correct, I also wish to draw out the errors, philosophical and tactical, that Dennett commits on the periphery of his philosophy. Almost without exception, I submit, what Dennett says wrongly, or expresses poorly, can be made right in a fashion that preserves the *spirit* of his project. Righting several wrongs, such as Dennett's conflation of low-level homuncular explanation and neurological explanation, is one aim of this paper. Moreover, the boundaries around Dennett's arguments may often be redrawn such that potentially damaging criticism no longer applies, and yet the new argument proves the required point. An example of this is found in the first chapter, in which I suggest an Einsteinian-physics version of a Dennettian argument that seems, at first glance, to rest on the truth of naive physical Realism. Effecting such redrawings is another aim of this paper. And most important is the project of showing how the methods and conclusions of a Dennettian interpretationist programme may be developed into a comprehensive view of the self, one with significant connections to specific issues in the philosophy of mind. The two issues I tie into Dennett's meme theory of the self are the Parfitian view of personal identity, and the metaphysical implications of Multiple Personality Disorder.

Parfit's theory of personal identity is, or was, revolutionary in its own right; both it and the meme theory of the self benefit from an investigation of their mutually supportive conclusions. Both approaches deny generally accepted "deeper" facts about the mental - Parfit denying a deep fact of personal identity over time, and Dennett denying any intrinsic fact about representational states. Multiple Personality Disorder, in turn, presents a bizarre possibility: a multiplicity of selves sharing a single body.

¹ I would like to suggest that, even were exegesis all I undertook in this essay, it would be a worthy project to clarify and adequately develop the important - even revolutionary - ideas that Dennett has presented over the past twenty-five years. He has written only two genuine books; his other major publications have been collections of papers. And given the vast number of threads that he pulls together to make his position seem plausible, the challenge of assimilating the arguments of those papers into a single coherent point of view is a demanding one indeed.

This is a metaphysical consideration in need of examination by a theory of the self, and yet the opinion flourishes in some quarters that philosophy of mind cannot legitimately appeal to this disorder. These specific issues invite treatment, and I provide it in the work that follows.

i. The need for a meme

The ascription of propositional attitudes is practically indispensable when giving explanations of complex human behaviors. The postulation of propositional attitudes also plays an important role in accounting for the *spread* of these behaviors between individuals. We often speak of the spread of certain beliefs, or ways of thinking, or concepts, or ideologies. And whatever one's initial position on the composition or ontological status of things like beliefs, concepts, and so forth, we will agree that our talk of the spread of such things is meant as an explanation for spreading behavior of a recognizably similar sort. For example, one might notice that more and more people are commenting on the genius of Wittgenstein, and are buying his books, and writing papers extolling his insights. One might explain this to some small extent by speaking of the spreading belief that Wittgenstein was brilliant. Notice, however, that there are many cases of spreading behavior that do not seem particularly amenable to explanation in terms of propositional attitudes. For example, if I hear someone humming Beethoven's Ninth, and I then begin to hum it myself, I would be forcing the idea of a belief if I were to explain my humming in terms of any *belief* about the musical structure or aesthetic worth of the tune. One would say, rather, that the tune itself - and not necessarily any belief about it - is "in my head." Similarly, in cases of spreading behavior in more simple organisms (like novel songs spreading through populations of reed warblers), one may wish to be cautious about attributing to the animal the appropriate set of beliefs and desires to account for the behavior. Here, too, we desire an account of the propagation of these behaviors that grants the same intelligibility and predictivity granted by

explanations in terms of propositional attitudes.

A broader concept than the propositional attitude, and one that subsumes such cases quite nicely, is Richard Dawkins' *memes*. Near the close of his 1976 work *The Selfish Gene*, Dawkins introduces the meme as a unit of imitative behavior. Having developed and explicated the predictive methodology of treating genes as if they were solely and selfishly concerned with copying themselves, Dawkins speculates that certain aspects of behavior could be given a similar treatment. Patterns of behavior, he suggests, might spread and replicate for no good reason other than that *they can*; like genes in a gene pool, these memes would "struggle" to expand in the available space of a "meme pool". The relatively new medium in which the meme pool resides, Dawkins argues, comprises the set of human brains.

Examples of memes are tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches. Just as genes propagate themselves in the gene pool by leaping from body to body via sperms or eggs, so memes propagate themselves in the meme pool by leaping from brain to brain via a process which, in the broad sense, can be called imitation (1976:192).

Taking Dawkins' suggestion to be both plausible and promising, one must recognize that the same ontological worries associated with propositional attitudes are associated with memes. Indeed, these concerns will be compounded by the current unfamiliarity and fuzziness of the concept of a meme. Given the usefulness of the meme in explaining behavior that resists explanation via propositional attitudes, it is worthwhile to develop an account of the nature of memes. Fundamental to this account, and the object of this paper, is an outline of what memes are and are not. Are all these things, as Dawkins claims, really *in my head*? What is the relation between the contents of Dawkins' list and the mental generally? While I have already suggested that memes are invoked to perform the same task as propositional attitudes (viz., explaining behavior), it is clear

that some memes are not propositional attitudes. This implies some classificational division within the class of mental entities. I will begin with the classification issue, since it illuminates and foreshadows my approach to the ontological issue.

ii. Locating memes in the mental schema

Since the class of memes includes beliefs, desires, and a great many other mental entities, one might well wonder whether "meme" is just a new term for "mental object." Many would claim that this cannot be the case, though, since Dawkins' definition presents memes as susceptible to transmission through imitation, and the utility of the meme as I have presented it depends upon this feature. However, there are mental states that do not appear to be transmissible by imitation; an example is pain. To be sure, someone in pain may engage in behavior that others could imitate; indeed, the appropriate behavioral patterns might catch on as some sort of fashionable dance. Most would agree, though, that what matters about pain has resisted transmission in such cases. I am not at all certain that I agree with this analysis, but for my current purposes it does no harm to grant that pain is not a meme, and the class of memes therefore does not contain or equal the class of the mental.

Figure 1:
Divisions within the mental

This analysis enables one to sketch a broad picture of the categorical relations between propositional attitudes, memes, and the mental. The conception of the mental set out in Figure 1 makes two important assumptions, neither of which is a source of difficulties for this way of conceiving the mental. First, this diagram assumes that there are no non-meme propositional attitudes. This is a reasonable claim, I think, but I am not inclined to defend it beyond saying that, upon reflection, no examples come to mind of attitudes that cannot conceivably be spread via imitative behavior. The second assumption implicit in Figure 1 is that there are no non-mental memes. This assumption is warranted simply because that is how I am defining memes. Certainly there could be examples of non-mental *replicators* that are closely associated with certain memes; sheet music is a good example, for it seems to replicate itself. That is, copies of a certain type of sheet music play an especially salient causal role in the explanation of how more copies of that type of sheet music come into existence. And this is exactly the sense in which genes replicate as well. But I am not discussing things like sheet music, or genes, here. Since I am explicating an idea developed only cursorily by Dawkins, and since I am doing so with specific purposes in mind, I can be openly Procrustean in my delineation of the concept of the meme.

It follows from the relations indicated by Figure 1 that the work already done on the issue of the ontological status of propositional attitudes may serve as a guideline - but not much more - for an investigation of the ontological status of memes. Because some memes are non-attitudinal, one cannot simply apply the conclusions of an inquiry into propositional attitudes directly to the entire class of memes. If those conclusions *are* so applicable, this must be shown, not assumed.

In measuring the applicability to memes of existing views of propositional attitudes, I will take as a clear case of a non-belief meme the example of a tune. I have mentioned above that the aspect of belief seems insufficient as a mental explanation of one's humming a tune. Furthermore, the aspect of desire is similarly separable from tune-humming; we are all familiar with the maddening phenomenon of one's humming for (what seems like) hours on end a tune one utterly despises. In these cases, the most

salient desire that figures is a desire to *stop* humming the tune. Tunes, furthermore, are *nonrepresentational*. A tune, suitably employed, may indeed represent something, but then so may a bottle cap. It is not essential to the concept of a tune that it be about anything at all. A belief or a desire, on the other hand, is necessarily about something. In whatever sense the tune itself is "in my head," then, it is neither a belief nor a desire.

So, what must the notion of a tune *qua* meme capture? One condition is that it must admit of replication independent of transcription. While one might make a case for regarding sheet music as a replicator, our tune-meme must be such that it could spread through a population totally ignorant of the methods of transcription. Otherwise, the ascription of the meme cannot explain how a tune could become popular in a group of musical illiterates. But this still leaves a few candidates for what the meme might actually be, including: a sequence of air-mediated compression waves of certain frequencies; the behavior that produces such waves; the neural patterns that cause that behavior; or, as I will argue, an *abstractum*. In the course of evaluating the tenability of these familiar conceptions of beliefs and desires, then, let us examine how those conceptions apply to the broader class of memes. In what sense can both beliefs and tunes be *in my head*?

Ideas certainly are commonly envisioned as cranial residents, but clothes fashions are memes as well, and ideas and clothes fashions, on many construals, occupy different ontological categories. Consider: "This year's fashions are higher hemlines and lower heels." This is a perfectly untendentious reference to clothes fashions which appears to equate them with physical objects. Must ideas, tunes and the rest then be physical objects if they are all to be memes? I shall argue that memes are in fact abstract entities and not physical objects, an argument that will proceed in the main from Dennett's more specific case for beliefs being abstractions and not physical objects. In the course of presenting my case, I will outline what I see as a far more comprehensive account than Dennett's of how mental objects *qua* abstractions may be squared with our intuitions about the mental - and especially with the seeming power of mental events to cause behavior. In order to deal with these two issues, I propose to begin by examining what

a belief, idea or meme is *not*. That is, I am not attempting to define memes in the standard philosophical sense of giving necessary and sufficient conditions; rather, I am considering candidates for the explanatory task given above, and eliminating those that do not fit the job description.

It is, I think, a natural first reaction to suppose that beliefs are in some sense neurons. It follows from the definitions given above that a belief is an entity posited as the cause of a certain pattern of behavior, enabling one to explain the spread of the behavior in terms of the spread of the causal entity. We accept, furthermore, that behavior is caused by neurons. Therefore, since the belief is said to cause the behavior, and since neurons in fact cause the behavior, one may conclude that beliefs are neurons.

To believe this is to be a Physicalist. There are, however, two avenues open to Physicalism. One is the type identity theory, and the other the token identity theory, or functionalism.² The former holds that types of mental states are identical to types of neurophysiological states, while the latter makes the weaker claim that all tokens of mental states are tokens of brain states. I believe that neither of these theories are satisfactory, and will suggest a two-stage argument for rejecting both of them and abandoning altogether a Realist conception of belief.

II. What Memes Are Not

i. Type identity

Under type identity, all beliefs that *p* are identical to, say, neuron firings of type

² Clearly I am using the term "physicalism" in a different sense than its original, under which it was synonymous with type identity. I use this term, rather than "materialism," because the conception of belief that I will summarize and refine is materialist without being an identity thesis. That is, it holds that there are no immaterial substances - every *thing* is material - but it denies that beliefs and desires are physical.

x-y-z. This theory predicts that occurrences of the belief that *p* and *x-y-z* firings will someday be discovered to correspond in all organisms of which believing that *p* may be reasonably predicated. Token identity theory makes a weaker claim than this since it merely insists that any mental event be some physical event or other. One type of mental state, on this account, need not involve the same type of brain structures on different occasions or tokenings. The first stage of the argument against Physicalism is directed against the type identity theory, and is generally known as the argument from multiple instantiations.

One way in which to frame the argument from multiple instantiations is to note the spread of memes across species. I can credibly whistle a meadowlark's tune, and though this might be possible, it seems unlikely that there could be the right sort of shared brain structure between the meadowlark and me for common neural structure to account for this. Thus, it seems there is no reason to believe that similar behavior in different organisms *need* be caused by anatomically similar events. Now, both the meadowlark's behavior and my own are caused by some neural event or other, but not the same anatomically differentiated *kind* of event. But the meme, by definition, is the same in us both. So the meme cannot be a type of brain event.

Multiple instantiations of this variety do not constitute an uncontested argument against the identity theory, though. One famous response to this argument, mentioned by Paul Churchland, is that the relevant identities could be species-specific (42). Now, anyone sympathetic to the species-specific response faces a somewhat awkward question: In what sense are "joy-in-a-human" and "joy-in-a-Martian" both *joy* if there is no interspecies criterion for identifying joy? On the assumption that human joy and Martian joy are identical to their respective, and non-identical, brain state types, we reach the conclusion that human joy and Martian joy are not identical to each other. Still, they are plainly both *joy* in some sense. Their differentiation, not surprisingly, springs from the fact that one is humanlike and one is Martianlike. But this leaves the species-specific thesis in trouble, for now we see that human joy is human in virtue of its being exemplified by humans, and Martian joy is Martian in virtue of its being exemplified by

Martians, while both must be joy in virtue of some further criterion shared across all species. Something crucial has been left out of the species-specific account - namely, that which engenders the ascription of common mental states to different species. It seems, though, that whatever could be added to the account to make sense of such ascriptions, whatever matters in ascribing mental states, would be an interspecies factor. And thus, whatever matters is not species-specific.

Furthermore, not all cases of multiple instantiations are between species. Individuals can re-learn certain abilities and come to regain certain beliefs after undergoing neurosurgery that severs neural connections or even removes brain matter, which seems a powerful indication that the same behavior within a species can have different neural causes.

The argument from multiple instantiations, whatever its force, is plainly applicable to the tune-meme case as well. Indeed, one of the examples given above is of the spread of a meadowlark tune from the bird to me. While that argument - that it just seems unlikely that the meadowlark and I are of sufficiently similar neural structure for the meme to be a neuron - is hardly knockdown, the second argument given against type identity above is more suggestive. Since it is (mercifully) possible for patients to undergo surgery that initially removes their ability, say, to whistle a tune, and yet be able to relearn that ability, it must be that the tune-meme is not a type-differentiated neuron or neural complex. However, this leaves open the possibility that the tune-meme is a *functionally* differentiated brain state.

ii. Token identity

a. Neural facts and representation

Clearly the arguments just presented are unproblematic for, and even congenial to, token identity theories. Functionalists could cheerfully accept the argument from multiple instantiations and yet argue that, whatever physical events caused the whistling in the meadowlark and myself, these events were identical to those particular occurrences of the meme. A far more complex argument is needed to undermine the token identity thesis; such an argument is offered by Dennett (1987), and might be termed the argument from the problem of misrepresentation, although, as I will discuss below, its applications are broader than this name suggests.

This argument is itself a multi-stage affair which ultimately attempts to show that propositional attitudes must be different sorts of things than physical objects - including brains, their constitutive parts, and their token states and processes. This conclusion is based on the claim that there is no objective fact of the matter about what mental state an organism is in, while there is clearly a fact of the matter concerning what physical state it is in. Thus, runs the argument, the two cannot be the same. Dennett's argument for this claim is based on the phenomenon of misrepresentation.

To understand the motivation for the argument from misrepresentation, one must understand that Dennett more or less views explanation and prediction as two sides of the same coin. Thus, he proposes a tripartite division of predictive strategies to suggest three types of explanation one may give for events. Each strategy he calls a *stance*, since in adopting the particular strategy we are taking an attitude toward the world, under which particular sorts of explanations and putatively causal entities are mandated.

Consider the physical strategy, or physical stance; if you want to predict the behavior of a system, determine its physical constitution... and the physical nature of the impingements upon it, and use your knowledge of the laws of physics to predict the outcome for any input...

Sometimes, [however,] it is more effective to switch from the physical stance to what I call the design stance, where one ignores the actual... details of the physical constitution of an object, and, on the

assumption that it has a certain design, predicts that it will behave *as it is designed to behave* under various circumstances...

Sometimes even the design stance is practically inaccessible, and then there is yet another stance or strategy one can adopt: the intentional stance (Dennett 1987:16-17).

From the intentional stance, one ascribes beliefs and desires to a system based upon its history (both morphological and phylogenic) and its niche in the world. In the context of the further assumption that the system is fully rational - and Dennett follows Davidson (1984) in claiming that this assumption is a necessary feature of belief-desire ascription - one may make reasonably accurate predictions about the object's behavior under the circumstances. Any system of which one can in fact make such predictions (like a frog, or a human) is an *intentional system*. But on this account, the ascription of beliefs and desires to an intentional system is never determinate, but rather only pragmatically constrained. Cases of misrepresentation, Dennett argues, are cases in which the practical constraints upon ascription fail to prefer one ascription over another, and such cases therefore reveal the thoroughgoing indeterminacy of belief-desire ascription.

For example, take a frog out of its natural environment, place it in a laboratory, and project a series of small dark shadows moving across a light background. The frog will shoot its sticky tongue out at the spots. Or, more simply, sit by the swamp and toss lead pellets at the frog; it will catch and ingest them. Now, does the frog believe that it is ambushing flies in these cases?³ If so, it is mistaken. But here is another construal which accounts for the frog's behavior: it believes it is shooting its tongue out at dark specks. If this is the case, the frog is absolutely correct. The element of error, or misrepresentation, has evaporated with the new interpretation. Dennett argues that misrepresentation can always be dissolved in this manner,

³ Those for whom talk of frog beliefs is unacceptable may substitute the informational content of the afferent signal from the frog's eyes; the problem remains the same.

by adding disjunctions (the signal means something less demanding: fly *or* pellet *or* dark moving spot *or* slug of kind *K or ...*) until we arrive back at the brute meaning_n of the signal type ...(1987:302; 84-87; 290-295).

If there were a fact of the matter about what belief state the frog is really in, then there would necessarily be a chance that the frog holds an erroneous belief. But, since there appear to be at least several different ascriptions of beliefs that make equally good sense of the frog's behavior, and no principled way to select one as canonical, there seems to be no principled way to distinguish an error from a veridical perception. Thus there seems to be no such fact of the matter. On the other hand, the physical state of the frog is perfectly determinable; we might not be able to give an atom-by-atom description of it, but we accept that such an unambiguous description is possible in principle.

At this point Dennett presents his argument: If no conceivable test could decide between the hypothesis under which a system accurately represents *x* and the hypothesis under which that system misrepresents *y*, then there is no fact of matter concerning which representational state that system is *really* in (1987:300). Since there clearly is a fact of the matter concerning the system's physical state, the mental, intentional, representational state of the organism cannot be identical to the physical state of the organism.

This move in argumentation is one that has been labelled "verificationist" - or even "instrumentalist," a pejorative term associated with the view that only observation reports based on the data of the unaided senses are acceptable premises in scientific hypothesis. But this implausible view surely has little or nothing to do with Dennett's point here, which is a far more modest one. Dennett insists only that one must not impute a distinction to a case in which no evidence could bear out a real difference. Since I find *at least* this sort of instrumentalism - if it must be called that - plausible, it falls to me to defend it against the most common attempts to undermine it.

b. Frames of reference, function, and generalization

One way for the functionalist to meet this argument is to challenge Dennett's claim about the determinateness of physical states. Einsteinian and quantum physics suggest that observers in different frames of reference can disagree on the shape, mass, and particulate composition of an object. So, it might be asked, how can the determinate nature of the physical serve as a criterion of difference from the representational? This potential line of thought, though, offers no help to the functionalist. One need not simply claim that the determinate/indeterminate distinction is what sets the physical apart from the representational; the more profound point revealed by Dennett's examples is that a system's representational state can remain opaque even if its physical state is transparent. Far from mitigating the distinction, a lack of physical determinacy would actually *exacerbate* the problem. The indeterminacy of the physical would seem to make type or token identities still more unlikely; now both the representational *and* the physical would be in an observer-dependent flux. This gives us even greater reason to deny that the representational maps onto the physical, and Dennett's argument thus succeeds even more clearly within the physics of frames of reference.

Now suppose that two observers occupy a common frame of reference, or sufficiently similar frames that they agree on all their physical descriptions and observations. (The physics of frames of reference tells us that this is more or less what we all do). For all practical purposes, these observers could assume the truth of the most naive sort of Realism about the physical, as Dennett seems to assume in his argument from misrepresentation. This yields up the same situation proposed by Dennett in his original arguments, in which the representational status of behavioral systems can be underdetermined regardless of our agreement on their physical status. Thus, *neither agreement nor disagreement about the description of a system's physical state mitigates the underdetermination of the system's representational state*. This, I believe, precisely captures what is expressed by the claim that the intentional and the physical are different *stances*. The distinction need not be that the representational is determined

pragmatically, and thus is subject to indeterminacy, while the physical is determined absolutely, and thus is never indeterminate. One may rather make the distinction on the grounds that our way of carving things up physically never strictly implies the way we carve things up into representations.⁴

Another response to the argument from misrepresentation might be to reintroduce the possibility of representational error by restricting the set of alternative interpretations to those that make sense evolutionarily. Why, after all, would a frog evolve a nervous system capable of accurately representing flying lead pellets? But this response is unsuccessful. Given a frog's environment, stingy natural selection would do as well to make a frog's brain states *really* mean "dark moving spots across field of vision" as it would to make the same state *really* mean "food." After all, in the history of frog evolution, dark flying spots usually *were* food. There is no test that could distinguish in principle between these two intentional characterizations of a single brain state.

Finally, one might object to the argument from misrepresentation on the grounds that it shows nothing about the vast majority of our beliefs and desires, the determinacy of which seems beyond any reasonable doubt. Let us call this the *generalization objection*. This objection essentially claims that, even granting the indeterminacy of limiting cases of belief-desire ascription, we may advance an identity theory regarding the bulk of such ascriptions. To argue from cases of frog misrepresentation, or poor visibility conditions, or Twin Earth idiosyncrasies (Is your belief really about a horse, or really about a shmorse?) to the nature of all beliefs and desires is to mount an argument suspiciously of the form: Some *x* are *F*; therefore, all *x* are *F*. The identity theorist might argue that some cases are indeterminate precisely because no identities obtain, but that in most cases such identities do obtain, and thus most propositional attitudes are determinate. On this objection, the indeterminacy of frog beliefs (to paraphrase Fodor's devastating anti-Gibsonian witticism) adds only to the mounting evidence that frogs are not persons (Fodor 1990:208).

⁴ These points emerged from a useful discussion with Chris Viger.

There are a few lines of response to this objection, the most basic being that science generally is the business of arguing from some to all. This is, after all, the basis of inductive inference. While the quantificational calculus tells us that this is strictly a fallacy, we might nevertheless recognize that when the only evidence at our disposal indicates that some x are F , the conclusion that all x are F is hardly outlandish.

A better response to the generalization objection is one roughly based on simplicity. We have as phenomena to be explained: the seeming determinacy of most propositional attitudes and, as I have argued (and the generalization objection grants), an indeterminacy of some cases of ascription. Now, even in the indeterminate cases, while there is no choosing one ascription as superior to another, we may arbitrarily ascribe some plausible (though not canonical) set of beliefs and desires to the system in question. Both Physicalism and Dennett's holistic account must agree that such ascriptions can be purely a matter of pragmatics - with the identity theories contending that our being able to (falsely, according to the Physicalist) ascribe propositional attitudes in marginal cases involves a sort of extension by courtesy from cases of *genuine* determinateness. But Dennett's account proposes that ascription is pragmatically governed through and through, and that the unproblematic cases of belief-desire attribution are only problem-free in virtue of the relevant system's being situated in the right environment for the normative logic of the ascriber to select one set of propositional attitudes as superior. The seemingly determinate cases are those for which a "unique semantic interpretation is practically (but never in principle) dictated" (1987:31). But any identity theory must propose two different goings-on to account for the determinate/indeterminate distinction: a pragmatic account of indeterminate cases and an identity account of the determinate ones. In the absence of any untendentious examples of such identities, the only motivation for asserting that the bulk of ascriptions are determinate is the fact that they *seem* determinate. But this seeming determinacy surely cannot count as support for an identity account, since the Dennettian account explains and predicts it as well.

This leaves us with cases of indeterminacy, and no reason to regard these other than as exemplary of the entire class of propositional attitudes. Moreover, we have in

the pragmatic account of ascription a mechanism by which the important features of the indeterminate cases may be extrapolated to the broader class of beliefs and desires. It appears, therefore, that granting the undecidability in principle - as opposed to the mere underdetermination, or undecidability in fact - of the limiting cases of belief-desire ascription legitimizes Dennett's extrapolation to the general nature of propositional attitudes. This is not to say that the argument from misrepresentation is incontestable, but rather that the crux of the matter is indeed the move from underdetermination to indeterminacy. The generalization objection is *not* a crucial objection.

c. Inside the argument from misrepresentation

Recall at this point my claim that the argument from misrepresentation has a broader relevance than that merely of representational states. My reason for making this claim is simply that the argument applies equally well in assessing the nature of wholly non-representational memes, such as tunes. Strictly speaking, the heart of the argument is not misrepresentation, but rather the undecidability of ascription that misrepresentation can engender. According to Dennett's argument, since an organism's behavior severely underdetermines any ascription of beliefs and desires to it, there can be no fact of the matter about its beliefs and desires. In the case of propositional attitudes, this undecidability appears to spring from the way in which beliefs and desires are ascribed - that is, through a principle of charity regarding the rationality of the predicted system. To use Davidson's term, the *constitutive principles* of intentional attribution are that a system is rational enough to act on its beliefs in pursuing its desires, and that it has mostly the beliefs it ought to have, given its history and niche (Dennett 1987:17-18). Indeterminacy of the representational content of propositional attitudes arises when the assumption of rationality and the postulation of reasonable ancillary beliefs cannot support one ascription over another. Rationality, however, hardly seems a necessary constitutive principle for broader meme attribution. Need one be rational to have one's humming caused by a meme? It falls to me to show that the argument strategy enabled

by the possibility of misrepresentation in belief-desire psychology is also viable in the discussion of non-representational memes.

I believe that a strong argument *can* be raised against any view of non-representational memes under which they are determinate states. Specifically, it seems that even the ascription of memes like the tune-meme depends on conditions of ascriptive plausibility, albeit less defined and less familiar than those by which propositional attitudes are ascribed. It is not clear that the behavior "caused" by memes *is* ever such that it supports a principled, canonical attribution of one meme. Consider "Ave Maria": it is written in a certain key; it consists of several distinct and recurring phrasings; and it is several minutes long when played at standard speed. If I am heard to hum two recognizable bars from "Ave Maria," do I have the same meme as a person who hums two different bars of the same piece? I think not; tune-memes ought to be small enough to capture the fact that, for the most part, when a pattern of humming behavior spreads it is quite a short pattern of behavior that spreads.⁵ In the songwriting industry these are known as "hooks"; musical phrasings short enough to be easily remembered and repeated. So in this case, we are probably confronted with two distinct memes and not an underdetermined occurrence of one. Does this mean that there is a deeper fact about what meme someone *really* has? Here is why I think it does not.

When composers sue other composers for plagiarism, the judge and jury on the case often are confronted by a serious difficulty in adjudicating the situation. The reason is that the guidelines for what constitutes a hook are so ill-defined that it is next to impossible to say beyond any doubt when two hooks are identical. Even if the song of the accused is in a totally different key, at a completely different speed, with an altered mood and different instrumentation from the song of the accuser, plagiarism can be ruled to have occurred. That is, it may be ruled that one songwriter spreads with a song a tune-meme that another songwriter has already invented and spread. But if all these elements may be different and the meme the same, wherein lies the meme? Surely not

⁵ That is, people do not often hum an entire song to themselves; they tend to repeat a few short phrases of the song.

even in the pattern, describable mathematically, of the series of notes in a hook, since the stolen hook could be one note less, one note more or one note different and still be judged the same tune-meme. The jury must decide whether the accused has presented an established meme with "smokescreen" modifications or invented a new meme altogether.

Although lawyers from both sides tend to bring in experts to show sound graphs and computer printouts, at the end of the day the issue is decided by judge and jury listening to both songs and *legislating*, not discovering, whether the hook, the meme, from one is incorporated in the other. Similarly, if I hum a phrase from "Ave Maria," it might be in the wrong key, at the wrong speed, and I may get two or three notes wrong, and yet everyone could agree that my humming was the result of my having the meme for (some portion of) "Ave Maria." In so agreeing, my listeners must decide whether I am inaccurately reproducing "Ave Maria" or accurately producing some other meme. They must do so, furthermore, via a combination of a principle of *musical charity*, which will mostly consist in assuming that I am at least capable of accurately reproducing some meme or other, and various ancillary considerations. Such ancillary considerations will be whether I am known to own a recorded version of "Ave Maria," or whether I listen to classical radio stations, whether I have just said, "I am about to hum a segment of 'Ave Maria,'" and so forth. We may think of these as *conditions of ascriptive plausibility*. Clearly, though, there may be cases in which such considerations are ambiguous. The result for the debate over accuracy versus inaccuracy is that it may be simply impossible to choose one explanation over the other. And thus, even if a tune-meme is wholly nonrepresentational, one may make the same verificationist move made in the argument from misrepresentation: If the state of my brain as I hum is x , and yet there is no way in principle to discover whether my "meme state" is p or q , my meme state cannot be identical to my brain state. The argument from misrepresentation applies equally well to memes.

iii. Behavior

A rejection of the view that memes are neural states does not entail the conclusion that a meme is in no way a physical aspect of the organism. It may seem reasonable to fall back to the position that what is common to the meadowlark and myself, for example, is the behavior which leads one to propose that a meme has spread from the bird to me. Since behavior is in some sense a physical phenomenon, it is possible that the meme may yet be physical - that is, if beliefs are behavior.

One argument against a belief *qua* meme being behavior is fundamentally the argument against logical behaviorism. For fairly well-known reasons, I reject the possibility that a meme for, say, liking the opera could be expressed in terms of a long (or even infinite) conjunct or disjunct of behavioral dispositions.

The meme for liking the opera could not mean the same thing as "will go to the opera house." I might like the opera, yet not go - for instance, if I am not informed that the opera is currently in town. Can the behavioral description be modified to allow for this circumstance? Consider "Tim will go to the opera house, if he sees the poster advertising that the opera is in town." But I see all sorts of poster every day, and many of them I do not even stop to read. Append "and reads such a poster" to this behavior description, and the case is no better; I often read entire passages when in a hurry, without them noticeably impinging on my understanding. The point is that this behavioral description of the meme for liking the opera cannot get off the ground until it incorporates something along the lines of "*knows* the opera is in town," "*reads and believes* the poster," or some such. But each of these descriptions smuggles in a belief/meme predication itself. It is important to note that nothing is solved by supposing that I do in fact go to the opera house. In order for my going to the opera house even to be relevant to my having the meme for liking the opera - in order for it not to be an accident or coincidence that I go to the opera house - it must be posited that I believe that going to the opera house will yield an opportunity to take in an opera. The point is that any attempt to reduce a belief or a desire to behavior will need to incorporate a mental

predicate to explain how that behavior relates to the belief.

In fact, this entire objection may be seen as an extension of the argument from misrepresentation: just as a system's representational states are underdetermined by its neurological state, they are also underdetermined by its behavioral patterns.⁶ This is a result of the impossibility of separating the real, intrinsic patterns within behavior (if there are such patterns) from the patterns that our minds impose upon that which we observe. The phenomenon of the Tuneless Whistler exemplifies this difficulty.

Imagine yourself and a friend in an elevator, accompanied by one of those disturbingly enthusiastic whistlers who possess no talent for whistling. The whistler's efforts might seem at first devoid of any hook or rhythm, but no one who has ever been in such a situation would be surprised to find themselves "picking out" the scattered memes within the whistling. However, upon emerging from the elevator, neither should you be surprised to discover your friend whistling a different tune than your own, while both of you insist that the Tuneless Whistler served as your common inspiration. The mind, I submit, is a *pattern-extractor* - so much so, in fact, that it often "discovers" patterns in the environment where we might generally agree there are none. Since we have seen that there may be cases in which a meme is clearly transmitted, and yet the behavioral pattern-in-itself is indeterminate, the meme can be cashed out neither as behavior nor as patterns of behavior.

A final thought experiment should suffice to support this conclusion. Imagine a behaviorally complex and medically hypercapable robot that one day hears a meadowlark singing. This robot has no vocal outputs nor any apparatus suited for whistling, but it is capable of quickly (and painlessly, what luck) inserting tiny intrusive probes into the brains of passers-by. Through these probes, the robot stimulates the appropriate neural tracts with the appropriately encoded spiking frequencies, and initiates in the subject an

⁶ I am indebted to J.A. Brook, for his remarks on how this point relates to underdetermination.

experience qualitatively indistinguishable from hearing a meadowlark's whistled tune.⁷

If a meme is anything like what I take it to be, one will have to say that both the robot and the passers-by (who walk away whistling the tune they "heard") are infected with the same meme. But the robot's spreading of the meme involves neither waves of compressed air nor any sort of behavior typically associated with the production of such waves. Moreover, there is no reason to believe that the electrical stimulation is even remotely similar in encoding from one passer-by to another; our robot is just sophisticated enough to know what pattern to emit at whatever stage of signal transduction its probe happens to encounter. Thus the robot's behavior in spreading the meme to one person could be totally different from its behavior in spreading the same meme to the next person. Here we have, then, a tune-meme that is neither a sequence of sound waves nor the behavior implicated in causing such sound waves.

I have so far attempted to investigate memes by examining what propositional attitudes could not be. I chose this tactic both because Dawkins gives examples of beliefs as prime examples of memes and because it just makes sense that memes ought to subsume beliefs as explanatory entities of the sort apt to explain spreading types of behavior. For these same reasons, I will commence my proposal for how memes ought to be construed by dealing with the issue of how beliefs and desires ought to be construed.

III. What Memes Are

i. The Ontological Answer

My position is closely akin to Dennett's, since I concur with his conclusion that beliefs and desires are *abstracta* - Reichenbach's term for what Dennett calls

⁷ Although, to paraphrase J.L. Austin on the argument from illusion, they could hardly fail to notice the robot sticking probes into their heads; that might make the experience in some way distinguishable from the standard encounter with whistling.

"calculation-bound entities or logical constructs" (1987:53). This means that beliefs have the same ontological status as centers of gravity, a similarity that Dennett draws out for explicatory purposes.

An object's center of gravity is something postulated as a means to quick predictions of the object's movement - a means that requires far less information than is necessary to make similar predictions from an approach that mentions explicitly only the combined gravitational properties of the component part of the object. Although in principle the same predictions could be made from either approach, a center of gravity is not reducible to any physical component. Even if it happened that a component atom of the object coincided spatially with the object's center of gravity, one would not say that the atom *was* the center of gravity. The reason for this is bound up with what a center of gravity is. When I say that a cube's center of gravity is located at a point designated by the coordinates (x,y,z), I am saying that it is *as if* the object's mass inheres in that point. That is what it means to be a center of gravity, and an atom, however fortuitously located, could not share such an improbable but predictively rewarding property.⁸

Centers of gravity are clearly not physical things, yet their predictive power arises from their congeniality to being treated as causes. One would say - falsely, but often usefully - that the cube falls off the table *because* its center of gravity moved past the vertical plane of the table's edge, meaning to suggest a strong causal relation between the two events. This ability of beliefs and desires to be invoked as causes is the most profound consideration in the decision to treat mental states as neurons, a point we explored at the outset of this chapter. How does that square with what I am now arguing; namely, that neither beliefs nor centres of gravity are physical things? Easily;

⁸ Now, are centers of gravity real? Dennett insists that they are, along with beliefs, desires and Equators (1987:53). My own purpose in raising the question is that it allows me to say that I am not concerned to answer it here. The clarification I offer concerning the ontological status of beliefs and memes and centers of gravity may help resolve the question as to whether they are also real things, for all that they are not physical. But that is not my aim.

both centers of gravity and beliefs are *abstracta*, and therefore neither causes physical events. Neither *could*. But it is sufficiently *as if* they did play a causal role that one may derive an appreciable predictive power from so invoking them. The example of centers of gravity illustrates how expedient and ultimately unproblematic it can be deliberately to confuse one logical class of thing with another, as when we confuse *abstracta* like memes and centers of gravity with physical states.

Dennett is sometimes characterized as treating beliefs like useful fictions. As a result of his wanting to call mental states real though non-physical, he prefers the *abstracta* terminology. But my approach leaves the reality question alone and characterizes the intentional stance as embodying a *useful conflation*. Whether one prefers to call beliefs fictions or logical constructs, they are not physical. They are in some sense abstractions. Yet their predictive power depends on their being considered causally efficacious. Taking memes to be causally efficacious strikes me as a straightforward Rylean category mistake, although my contention is that it need not be a mistake, in the conventional sense, if one commits it fully aware of what it does and does not entail.

Since beliefs and desires are, as Dennett argues, abstract objects, they cannot be causes. Nevertheless, predictive power and psychological intelligibility may derive from the misconstrual of a mental state as a physical state. If one possesses the talent, as humans typically possess, of ascribing the right state to the right behavioral system at the right time, the supposition that a mental state is a physical state can be a profoundly useful one.⁹ Thus it is true both that mental states are causally inefficacious and that their causal power is their crucial property, however abstract, since it is in virtue of this fictional causal property that mental states are worth ascribing.

It is not clear that Dennett's critics have adequately understood the way that intentional stance ascriptions conveniently step aside when explanations from the physical

⁹ The right kind of behavioral system is simply one for which this tactic works; an intentional system. We use the tactic, again, because it works; the ability to recognize behavioral patterns predictable in this fashion was a survival advantage to our ancestors, who passed the ability down to us. Even today, as Dennett comments, "try catching frogs without it."

or design stances are required. As recently as their book (1992), Fodor and Lepore argue that Dennett cannot account for, or appeal to, the presumed evolutionary advantage of a system's having true beliefs. Their reason for levelling this charge is that, according to Dennett, beliefs are not real things, and surely one cannot impose "a selectional history on *things that don't exist*" (1992:148). Beliefs that a system does not actually possess, in other words, cannot be an aid to survival. But as Dennett rightly responds, this is precisely analogous to the claim that an anti-Realist conception of centers of gravity is hopeless, since the low center of gravity that a sailboat does not actually possess cannot be an aid to remaining upright (1993:216). If we are embarrassed by the second conclusion, we ought to be equally embarrassed by the first.

Viewing beliefs as abstractions and folk psychology as a useful conflation is, in fact, perfectly consistent with Quine's claim that

If we are limning the true and ultimate structure of reality, the canonical scheme for us is the austere scheme that knows no quotation but direct quotation and no propositional attitudes but only the physical constitution and behavior of the organism (1960:221).

Strictly speaking, one must know when to admit that references to the causal power of beliefs (and, for that matter, of centers of gravity) are conflationary. But this admission does not place in jeopardy the erstwhile claim that references to beliefs are explanatorily valuable. Such references offer expedient and reasonably accurate predictions about the behavior of certain systems, and when I am trying to catch fish or buy a reliable used car from an experienced seller, I am not particularly trying to limn the true and ultimate structure of reality. I am rather trying to outwit a mackerel or a smooth talker. Given my cognitive limitations, neither is even remotely possible for me without ascribing propositional attitudes as causally efficacious entities. And not only does the intentional stance *qua* useful conflation assist with predicting behavior, but, if properly construed as a conflation, it does not interfere with predicting physical structure physically (Quine's

"limning"). Thus, the ascription of memes is of as much value as its varying predictivity confers.

ii. The Ascriptive Answer

I have argued that memes are abstract entities posited for the purpose of predicting and rendering intelligible behavior, without having to engage in the near-impossible task of making such predictions from a physical stance. But I have also claimed that the familiar background condition of practical rationality by which propositional attitudes are ascribed does not necessarily provide a context for the ascription of all memes. At least some memes - tunes, for example - require neither ancillary mental states nor assumptions of rationality in order to be invoked as causes of behavior. Against what background assumptions are such memes ascribed? Some account is owing to explain what generally constrains the predication of those memes not subordinate to the demands of rationality.

There are, I submit, two fundamental constraints upon the ascription of all memes. The first and more narrow level is that of psychological intelligibility. While there are many, many conjunctions of mental states capable of making sense of a given instance of behavior, there will be at least some ascriptions that clearly do not. For example, Jane's sitting down could hardly be explained, *ceteris paribus*, by saying of Jane that she is tired of sitting. Thus, even if mental states and psychological narratives are not determined, they are at least constrained by a combination of constitutive principles and *the way things turn out*.

It is important to note that even if the psychological intelligibility conferred by intentional ascriptions is intimately related to the predictive power of such ascriptions, the two are separable effects. Imagine, for example, going through a day in which nothing that anyone says or does is predictable from the intentional stance. One might

still be able to predict what people were going to do next in some other way; suppose that one possessed a script for the day's events (issued, say, by God).¹⁰ But even if one got through this day making predictions as successfully as usual, the events would not *make sense* in the important way that human behavior normally makes sense. One's predictive success notwithstanding, the ability to render intelligible what has occurred would be lost. While, as I have argued, the intentional stance explanation of behavior is a causal one, not every causal explanation confers the same intelligibility as a belief-desire explanation. Physical explanations, for example, confer a certain rich intelligibility to events, but not the sort that we expect to gain from having someone's actions explained in terms of their reasons. While tune-memes are surely not reasons for actions, they are causes of the same general kind as reasons - taking "reasons" to be roughly synonymous with a belief-desire pairing. The ability to tell this sort of psychologically causal story about the behavior of a system is the greater part of what tune-meme ascriptions confer, for reasons that I outline below.

The more profound sort of constraint operating on the ascription of memes is one imposed by the demand for predictive reward. Ascriptions are gerrymandered specifically for the purpose of predicting patterns of behavior, and this is what prevents the underdetermination of mental states by behavior from resulting in fluctuating and wildly disparate ascriptions in accounting for similar behavior at different times. When a system's patterns of behavior persist and recur, one is best served by supposing that the underlying causes of that behavior persist or recur accordingly. Since bodies usually cooperate by behaving in a way that permits this sort of ascription, a pattern emerges over time; one may postulate putative causes in such a fashion as to permit the prediction of patterned behavior.

It is conceivable that Mary stops at one red light because she believes in halting briefly at 1:25 pm, and stops at a second red light because she wishes to protest environmental pollution in her own symbolic way, and stops at a third because she likes

¹⁰ And "events" is the right word, since it is not clear that we could consider any events to be actions (utterances or behavior generally) if those events were not intentionally characterizable.

the color red and wishes to stop and admire it briefly. This is conceivable, and as an explanation it conforms to the first constraint upon ascription. But if Mary's behavior in stopping at red lights is patterned and generally predictable, the profligate ascription of reasons to account for obviously similar behavior is a waste of the ascription game, since it gives one no purchase on the task of prediction. If one is trying to decide whether to walk across an intersection when Mary has a red light, the ascriptions given above are useless, for they utterly fail to capture the regularity of the relation between red lights and Mary's stopping.

These two fundamental requirements for the predication of memes subsume not only the rationality context of propositional attitudes, but also the ascriptive plausibility to which I have already referred. Thus, in ascribing tune-memes, one will base the attribution not merely upon the constraints of intelligibility and predictivity, but also upon considerations of which candidate ascriptions are most plausible given what one knows of the subject's musical background, tastes, exposure, and so forth.¹¹ The plain conditions of ascriptive plausibility are what my audience has to go by. This, I believe, is the extent to which non-representational memes are constrained in their predicability, mainly due to their atomistic nature as (abstract) causes.

When I refer to non-representational memes as "atomistic" causes of behavior, I mean that they may be invoked as causes independent of ancillary mental states. It has long been recognized, however, that beliefs and desires are only causally efficacious in certain appropriate combinations, and only with the appropriate background conditions. That is, we think of behavior as the result of a desire in conjunction with certain beliefs about the best way to attain the desired end, given the circumstances. I am suggesting that this twofold nature of propositional attitudes as recruited abstract causes imposes a more profound stricture upon their predication - namely, that the ascription be sensitive to belief-desire combination and to the subtleties of the situation. Not only must a

¹¹ This is an elaboration on the point I made above, when describing the sorts of cues at the disposal of my audience as they decide whether I have inaccurately reproduced "Ave Maria" or accurately produced some other meme.

propositional attitude predication conform to the two fundamental constraints, and moreover ascribe states that the subject's history and environment suggest are plausible states to ascribe, but the predication must propose that the agent has put together a belief with a desire in the right combination for this particular situation. It is this situational sensitivity that confers such enormous predictivity upon intentional stance ascriptions by limiting the number of situations in which a given belief-desire combination is plausible. Since non-attitudinal memes are atomistic, the situations in which their ascription yields predictive power are far rarer. They seem mainly to serve as sources of psychological intelligibility.

The particular reason for the greater predictive power of propositional attitudes over other memes seems to ground on conation. Practical rationality, as the faculty by which the right beliefs are assimilated to the right desires at the right time, is blatantly an instrument of hypothetical imperatives; given the knowledge of a behavioral system's broad goals, the task of ascribing specific desires becomes fairly simple. Equally important, one may infer specific beliefs from broad goals as well, since knowing those goals gives one insight into what elements of the world the behavioral system ought to find salient (and thus, which aspects the system will especially note and internalize as beliefs).

iii. Conclusions

I have argued that tune-memes are subject to all the same arguments as belief-memes with regard to the ontological status they ought to be accorded. While the meme may be abstracted from behavior, and although neural mechanisms are causally

implicated in such behavior, the meme is neither the neurons nor the behavior.¹² Bearing this in mind, let us then examine the use to which Dennett puts the meme in formulating his theory of the self.

¹² I will conclude this chapter by drawing out a ramification of what I have said. On my account, tunes, *inter alia*, are the same sorts of thing as beliefs, in many important ways. They, like beliefs, are abstractions that may be usefully ascribed as physically real and causally efficacious components of a system. Now, given the argument I have presented to the effect that memes are *mental objects*, one may conclude that tunes constitute non-representational mental objects. But this is a clear challenge to the widely held view that intentionality is "the mark of the mental," since tune-memes are not *about* anything.

Chapter Two: Dennett and Consciousness

Cartesian dualism helped itself to the most straightforward explanation of the mind imaginable: Everything, it said, is exactly as it seems. Does it seem as if the mind is nothing like the material of the body? Well, the mind *is* nothing physical. Do you seem to have authoritative access to the processes by which your mind functions? Well, you do indeed have such access. And so forth.

The history of materialism in the philosophy of mind may be viewed as an attempt to sell *seemings* at a much higher price than this. To a science that accepts materialism more or less as dogma, the task is to account for the seemings of the mind without recourse to immaterial substance, and considerable progress has been made in this regard (though not so much that an Occam's Razor argument against Dualism is particularly warranted). The first seemings undermined by a scientific approach to the mind were of the privileged access variety; the development of experimental psychology brought to light the many cases in which subjects either have no idea about how they accomplish some cognitive feat, or alternatively they give some confabulated account that is not borne out by available evidence.¹³ The theory I will exegetize in the following chapter is an attempt to explain the important seemings of consciousness within a materialist framework. Dennett finds fault - indeed, incoherence - in the traditional view of consciousness and self, a view that he thinks still pervades most theories of the mind and brain. But it is not enough merely to lay accusations of confusion and metaphysical profligacy when the criticized view is "the only game in town." So Dennett also presents a positive account of why consciousness seems serial, why selves seem real, why thought seems independent of language, *inter alia*.

¹³ I have in mind here the familiar experiments in which subjects are given a list of items to memorize, and are then shown an item and asked if the item was on the original list. Subjects typically insist that they perform this task by searching the memorized list until they find the probe item, at which point they report. However, the response time of subjects varies not with the position of the probe item on the original list, but rather with the magnitude of the list. In other words, subjects appear to engage in an exhaustive search of the list regardless of the position of the probe item.

1. The Cartesian Theater

Dennett breaks the ground for his model of consciousness by raising what he contends are commonly held yet erroneous notions about consciousness. Chief among these are, first, the appealing belief that there must be some point in neural processing at which one becomes aware of perception and, second, the sense that experience occurs at some one time between sensory perception and behavioral response. These may be loosely described as mistaken notions about the spatial and temporal features of consciousness, respectively.

The first of these beliefs implicitly postulates some spatial characteristic of consciousness: namely, that it occurs at (or is correlated with) a certain place in the brain. Dennett refers to this idea as *Cartesian materialism*, since it assumes that the content of consciousness is eventually assembled in one spot, the spot in the machine where the ghost of Ryle's satire once sat. This spot is the Cartesian Theater. The argument is that many theorists, despite having rejected the idea of a substantive homunculus in the brain, persist in imagining that inbound neural representations somehow must be modified into experiences before they reach an alleged crucial boundary between pre-conscious and post-conscious. This must be the case, it seems, in order for the representations to make sense to... well, to whom? A "Central Meander." While the explicit idea of the homunculus has been abandoned, Dennett claims that many or most theorists of mind implicitly posit a Cartesian Theater - a sort of hangover from substance dualism.

...materialists today often forget that once Descartes' ghostly *res cogitans* is discarded, there is no longer a role for a centralized gateway, or indeed for any *functional* center to the brain. The pineal gland is not only not the fax machine to the Soul, it is also not the Oval Office of the brain, and neither are any of the other portions of the brain.

...*Cartesian materialism* [is] the view you arrive at when you discard

Descartes' dualism but fail to discard the imagery of a central (but material) Theater where "it all comes together" (1991:106-7).

It is easy to see why Cartesian materialism is a tempting view. Consider, for example, a situation in which light waves reflect from a coloured object to impinge upon my retina, subsequent to which I report seeing the object. Surely I am not conscious of the experience at the exact moment of the relevant photons striking my retina; the awareness must occur later in the process of neural excitation. But neither does it seem right that I am conscious of the colour when my stimulated retinal cones themselves excite or inhibit neurons in the bipolar layer beneath the anterior surface of my eye. But clearly I am eventually aware of the colour. Thus, it is tempting to infer that *somewhere*, at some precise point or other, the representation of colour borne by my neural tracts becomes conscious. The Cartesian Theater is an epithetical reference to the view that there must indeed be a place in which neural firings coalesce and transform into the unified consciousness that we all (profess to) know so well. Modern sophistication rejects the little man in the head, but surreptitiously retains the show staged for the little man, and the venue of its presentation.¹⁴

The second commonsense idea of consciousness attacked by Dennett is one that assigns an independent time of occurrence to experience. An example of this would be the belief that, when I stop for the red light, my experience of redness must occur *sometime* between the stimulation of my retina and my stopping. More precisely, the Cartesian Theater account supposes that the experience occurs roughly equidistant (temporally speaking) between these two events. Once again, one's intuition is not that the experience occurred at exactly the same time as retinal stimulation, nor as late as the subsequent behavior, but rather that some amount of time elapsed while the stimulus was "interpreted," and that some further amount of time must have elapsed while the behavior

¹⁴ To be fair, however, one need not be as epithetical as Dennett. What he refers to as Cartesian Materialism may also be described as the view that there must be an integrator or an excogitator to account for the way that the diverse informational inputs to the brain get integrated and synthesized in inferences.

- motivated by the interpretation - was put into effect.

Dennett introduces the color phi phenomenon as a counter to this intuition (1991:120). This phenomenon is generated as follows: Two differently colored lights, placed close together, are flashed in quick succession. Not only do subjects report seeing only a single light that *moves* smoothly across their field of vision, but they furthermore insist that the light changes color at a point between its origin and its termination. How, one asks in amazement, did the brain experience the light changing color when it was impossible for it even to "know" that the light at its terminus was a different color? Dennett suggests two possible ways that this seeming miracle could have occurred. One is termed *Orwellian* (after the historical revisionism depicted in Orwell's *1984*), the other *Stalinesque* (after the meticulously planned show trials of Stalin's Russia, designed to falsify observers' perceptions of events from the outset) (1991:116-117).

The Orwellian explanation is essentially that color phi subjects have faulty memories of a correct perception; the original conscious experience was of two differently colored lights flashed in sequence, but the brain records the experience as a single moving light. The Stalinesque explanation proposes a correct memory of a faulty perception; the single light is originally experienced - the doctoring of the data happens at a pre-conscious level - and this flawed experience is correctly preserved in memory. The important point here is that these two explanations agree (*ex hypothesi*) on every issue except the veracity of the original unremembered conscious experience; that is the problem they are recruited to settle in their own respective ways. But it turns out that neither subjective reports nor behavioral analysis can provide good grounds for choosing one of these two explanations over the other. By the very nature of the problem, the subject has no internal method of deciding (and, therefore, none of informing us) whether the memory or the eyes are playing tricks. And any *behavioral response* could be unconscious, "except for subsequent telling" (1991:124).

This sets the stage for Dennett's standard move: If two, or three, or many explanations can account for all the data, and if there is no way in principle to gather the kind of facts that would settle the debate over which explanation is the true one, then

there is no debate, for there is no reason to suppose that an explanation is required. The difference between the Orwellian and Stalinesque accounts, writes Dennett, "is a difference that makes no difference" (1991:125).

This is his familiar instrumentalism, the kind which Dennett goes on to argue "makes manifest good sense." To persist in imputing a substantive difference to the Orwellian/Stalinesque distinction is absurd, since "it creates the bizarre category of the objective subjective - the way things actually, objectively seem to you even if they don't seem that way to you" (1991:132)! But since the Cartesian Theater view proposes that representations become conscious when they reach the right spot in the brain, or cross the proper line, it must insist that the content-tampering of the color phi phenomenon occurs on one side or the other of the divide between what is pre-conscious and what is post-conscious. We may reject the Cartesian Theater view on the grounds that it necessarily posits such pointless distinctions. Thus Dennett's alternative to the Cartesian Theater, the Multiple Drafts model, is introduced as "*first-person operationalism*" (1991:132).

II. Multiple Drafts

Suppose a madman were to claim that there were no such things as animals. We might decide to confront him with his error by taking him to the zoo, and saying, "Look! What are those things, then, if not animals?" We would not expect this to cure him, but at least we would have the satisfaction of making plain to ourselves just what sort of craziness he was spouting. But suppose he then said, "Oh, I know perfectly well that there are these things - lions and ostriches and boa constrictors - but what makes you think these so-called animals are *animals*? In fact, they are all just fur-covered *robots* - well, some are covered with feathers or scales." This may still be craziness, but it is a different and more defensible kind of craziness. This madman just has a

revolutionary idea about the ultimate nature of animals (Dennett 1991:43).

i. Introduction to the model

As the first step in the construction of Dennett's "empirical theory of consciousness", the Multiple Drafts model attempts to debunk the intuition that there may be facts about consciousness that neither study of the brain nor subjective reports, nor any other study, could reveal. The challenge, though, is to effect this debunking by providing a plausible alternative to it. Dennett offers the assumption that, by definition, "a conscious mind is an observer," but he subsequently claims that "there is no observer in the brain" (1991:101, 106). Taken together these statements suggest the first kind of craziness, for they seem a fairly direct denial of the existence of consciousness. But the Multiple Drafts model provides the basis for Dennett's claim to be arguing for "the second kind of craziness" - an alternative conception of our mental lives. There *are* observers, on Dennett's account, and there are therefore conscious minds, but they are nothing like our traditional conceptions of them. Consciousness properly construed, argues Dennett, is ultimately unmysterious, in the sense that it is scientifically explicable.

Multiple Drafts is Dennett's alternative to the Cartesian Theater. Its fundamental tenet is that the brain does not first manipulate neural content, then interpret it; rather, the manipulations *are* the interpretations. For example, the effect of the bipolar layer of retinal neurons in exaggerating edges in one's visual field is not an operation performed in order to make visual edges susceptible to one's apprehension; that operation, and others like it, *constitute* the apprehension of edges. They must, because nothing else in the brain *could*, on pain of postulating miracles or "Wonder Tissue."

As the final component in the assembly of the Multiple Drafts model, and in keeping with what is known about the brain's *modus operandi*, Dennett proposes that these discriminations are effected by specialist systems in parallel operations across the brain, and not in a serial sequence. Since representations pass along neural tracts in a massively redundant fashion, and since there is no "charmed circle" within which a

representation is conscious and without which it is not, we are better off to think of such representations as "multiple 'drafts' of narrative fragments at various stages of editing in various places in the brain" (1991:113).

The analogy with publishing an article is well suited to capture Dennett's view of consciousness. Since an article, in its early stages of drafting, may circulate simultaneously to many different readers in considerably different forms, there may be no saying which draft is *the* article. It is only when one version of the article is published that that version is seen as canonical. But the authority of that version is fixed only by convention, and may still be revoked as updated and revised versions are reprinted. What fixes one of the neurally-borne representational "drafts" as "what was the content of consciousness" is memory, on Dennett's account. The color phi phenomenon is recruited to illustrate the absurdity of unremembered moments of consciousness; consciousness, says Dennett, is criterial upon memory, on pain of implying the category of the objective subjective.

ii. A digression for the purpose of ungarbling

Now, before moving on to outline the relation between Multiple Drafts and the meme theory of the self, I propose to straighten out a rather glaring inconsistency between Dennett's general view of the mental and his way of explaining the manipulation of representations within the brain. The task of ungarbling what Dennett seems to have garbled involves introducing the notions of *sub-personal intentional psychology* and *homuncular functionalism*.¹⁵

The function of what Dennett calls sub-personal intentional psychology is to help explain larger scale mental phenomena like selves by breaking those rich and homuncular entities down into less complex "specialist systems." Intentional stance psychology holds that the predictive power conferred by postulating mental states is bought at the cost of

¹⁵ I stole the delightful term "ungarble" from Gilbert Ryle (1964:24).

positing some single complex homunculus within a behavioral system. On Dennett's homuncular functionalist method of doing cognitive science, one dismantles this problematic homunculus by proposing to divide its job description among numerous, less intelligent, specialist homunculi. Each such specialist system demonstrates some specific cognitive capability that was originally ascribed to the mind broadly. Moreover, each specialist may be further broken down into subsystems of still less impressive intelligence, and so on until one is positing intentional entities of such limited cognitive capacity that a mechanism could perform the same task. Thus the intentional stance helps us to explain the *appearance* of an intelligent Central Meaner by positing a collection of unmysterious mechanisms.

Unfortunately, Dennett's distinction between abstract memes and the physical seems to fall by the wayside as his focus moves from macro to micro intentional characterizations. While he is adamant that there is no intentional-physical identification to be made when one treats mental entities such as selves and beliefs and desires, he is constantly referring to the *content* of neural tracts, the *effects* of representations on brains, and neurally borne *information* (1987:56-7). But if the grand phenomena of self and consciousness really cash out as micro-intentional entities that are in turn instantiated as the firing patterns of neurons or neural tracts, then surely one may conclude that the higher level postulata of beliefs, desires, and even selves, are - via their constituents - at least token identical to brain states. This conclusion, of course, flies in the face of the argument from misrepresentation. Assuming that Dennett cannot mean to support such a conclusion, I wish to outline a way of thinking about homuncular functionalism that permits the dissolution of complex mental entities without employing Dennett's dangerously misleading language. All one needs, I think, is a vocabulary that lets one easily distinguish the posited properties of representations from the real properties of neurons.

Such a vocabulary is provided by Zalta (1988). In the context of his attempt to develop an intensional logic, Zalta proposes that we say of abstract objects that they *encode* their properties, while ordinary physical objects *exemplify* their properties. Taken

in the right spirit, this way of speaking can be illuminating for homuncular functionalism - the right spirit being one in which Zalta's way of framing this distinction implies nothing more than the familiar distinction between the abstract and the concrete.

Things that exemplify the property of being a detective exist, have a location in space and time, are made of flesh and bones, think, solve crimes, and so on, whereas things that just encode the property of being a detective are abstract and do not exemplify any of these characteristics... [T]hink about your mental representation of Mark Twain. It might involve the property of having a walrus moustache, and maybe the properties of being white-haired and wearing a white suit and bow tie... However, the representation itself doesn't really exemplify these properties. The representation does not have a walrus moustache, the representation is not white haired; etc. But it does involve these properties in some crucial sense. This sense of "involve" is what we mean by "encode" (Zalta 1988:17-18).

Let us say, then, that the self encodes such rich and diverse properties as language comprehension, language production, formal reasoning, affect, and any other mental trait generally predicated of persons. This is the starting point for "top-down" approaches to studying the mind. What requires explanation is how we can end up ascribing these capacities to a system comprising nothing more than neurons and electrochemical events. The homuncular functionalist approach dictates that we distribute the tasks of this single self to various less capable homunculi; now we have one homunculus encoding the ability to understand language, one encoding the ability to recognize faces, and so forth. This process continues, yielding from the single "language homunculus" a homunculus encoding the ability to comprehend the written word, plus a homunculus capable of understanding the spoken word, *inter alia*. Each of these homunculi will comprise semantic and syntactic homunculi themselves. Ultimately, after a long and demanding

process of breaking up these job descriptions in rewarding fashions, we end up positing myriad "stupid" homunculi whose *encoded* abilities are no better suited to fulfilling their job descriptions than are the *exemplified* abilities of neurons or neural tracts. In fact, I think we may be quite precise on this point, since we know that neurons are capable only of responding to stimuli. If the ability of each neuron is to fire when stimulated, then we know in advance that the job description of *each and every* lowest level homunculus will be: Speak up when you notice something.

One may now diagnose Dennett's misleading talk of representations *being* neural activity; he fails to distinguish the "stupidest" level of abstract homunculi from the neurons that actually do the job. There is a good reason to think that this ultimately is a harmless tactic, though. Even Dennett thinks that there are unproblematic cases of belief-desire ascription; such cases are, he claims, instances in which a system's connections to its environment are sufficiently rich, and its function or goals sufficiently obvious, that it would simply not make sense to attribute to the system any representational state besides the obvious ones. That is, the constraints on the plausibility of ascriptions, discussed above, support specific ascriptions as clear winners. Since we define the agenda of homuncular functionalism, the goals of the stupidest homunculi are carefully gerrymandered to coincide with the capacities of the neurons we have to work with, and thus we bestow unproblematic belief-desire ascriptions upon those neurons. Nevertheless, the individual neurons will be nothing more than systems for which a "unique semantic interpretation is practically (*but never in principle*) dictated" (1987:31, my italics). This explains how Dennett may rather incautiously equate representations with neural tracts - or the activity of neural tracts - without insinuating any sort of reducibility thesis, let alone identity. For whatever reason, Dennett prefers to speak of lower level homunculi in a manner that equates them to "wetware," but clearly this is a mere rhetorical device and not a necessary feature of homuncular functionalism.

III. Evolution, Serial Consciousness, and Parallel Brains

As it stands, the Multiple Drafts model leaves unanswered two enormously important questions. First, we are left to wonder what exactly is swirling around in our brains (however abstractly) when various "drafts" are in circulation. That is to say, what are the building blocks of these drafts? And second, the ways and means of the brain notwithstanding, consciousness is serial rather than parallel. How does an explanation in terms of massively redundant drafts account for the one-thing-at-a-time nature of my awareness? As usual, we can answer most questions by asking the right one first. In answering the question: Drafts of *what*? What is being built? Dennett's evolutionary account of the origins of serial consciousness also suggests an answer to the first two questions. Therefore, one ought to outline this evolutionary account in some detail.

This endeavor must begin with an important point of interpretation. As Ned Block (1993:181) and J.A. Brook (1994:47) have pointed out, one the great failings of Dennett's account of consciousness in *Consciousness Explained* is that he never adequately describes what he trying to account for. The assumption of a univocal sense of "consciousness" is utterly unwarranted, and has the unpleasant side effect of seeming to show that most animals are not conscious. This is because Dennett's account seems to be about very sophisticated and thoroughly linguistic phenomena - the misplaced belief in the self as a sort of "mind-pearl," or the illusion of qualia as real things. Surely other species of animal - say, dogs - are not conscious in the sense of having these sorts of beliefs. But it is equally certain that dogs are indeed conscious in some important way. They are clearly aware of their environment, and their behavior merits the ascription of an enormous number of complex beliefs about their surroundings.

What I propose is that Dennett means to discuss self-consciousness, and specifically the sense of a self *as a self*. That is, he wishes to tell an evolutionary story about the emergence of systems capable of having beliefs about their own beliefs, systems that can represent certain of their own representations, including the subject of those representations. When Dennett speaks of consciousness, then, let us take him to

mean this kind of self-consciousness; this interpretation is the most consonant with the case he presents and the misconceptions he is concerned to undermine.

The story of the evolution of consciousness, Dennett suggests, may be picked up at the point when creatures emerged whose survival depended on *acting*.

In order to cope, an organism must either armor itself (like a tree or a clam) and "hope for the best," or else develop methods of getting out of harm's way and into the better neighborhoods in the vicinity. If you follow this latter course, you are confronted with the primordial problem that every *agent* must continually solve:

Now what do I do (1991:177)?

The evolutionarily cheap solution to this problem is to equip such creatures with simple all-or-nothing reflex response systems. The simplest nervous systems are stimulated to feed or flee by direct physical contact with some important features of their environment - predators and prey, for example. More advanced reflex systems, when they evolve, will permit an organism to respond to distal aspects of the environment. Dennett gives two examples of this latter sort of alarm: vertical symmetry detectors and looming-object detectors (1991:178-9). But these reflex systems, both proto-somatic and proto-visual, function in isolation as "specialists," each monitoring the environment for signs of whatever salient feature it was selected to detect.

The next step up from such organisms are animals whose nervous systems are capable of responding to an alarm from one specialist by going into an *orienting response*. Organisms like some fish species will, when faced with something threatening or unusual, call a sudden halt to their activities while all the specialist systems broadcast their status.

A *temporary* centralized arena of control is established through heightened neural activity - all the lines are open, for a brief period. If the result of this update is that a "second alarm" is turned in, the animal's whole body is mobilized, through a rush of adrenaline. If not, the heightened activity soon subsides, the off-duty crew goes back to bed, and the specialists resume their control functions (1991:180).

But once such orienting responses are installed as reactions to alarms, they are no doubt very cheap (evolutionarily) to employ as active information gathering tools. One ought to expect that an animal will begin to provoke its own orienting response frequently - indeed, constantly - given this tactic's power to keep the organism informed of relevant environmental features on an ongoing basis. This recourse to the value of turning preexisting structures and abilities to new purposes is the recurring twist in Dennett's history of consciousness. In this first appearance, it sets the stage for the evolutionary explanation of serial consciousness.

The story has moved beyond nervous systems in which every signal activates a hardwired response of evasion, aggression or mating. A level of complexity above this threadbare kind of representation are phenomena such as the hardwired tendency to duck when objects loom up in one's field of vision. Dennett ponders, "Does the 'something looming!' signal *mean* 'duck!'? Well, it proto-means it; it is wired directly to the ducking mechanism." Now, with the introduction of the orienting response, these hardwired specialists become active seekers of information and active constructors of representations.

Regular vigilance gradually turned into regular *exploration*, and a new behavioral strategy began to evolve: the strategy of acquiring information "for its own sake," just in case it might prove valuable someday. This marked a rather fundamental shift in the economy of the organisms that made this leap: the birth of curiosity, or epistemic

hunger... [T]hey began to become what the psychologist George Miller has called *informavores*: organisms hungry for further information about the world they inhabited (1991:181).

Once one introduces the element of neural plasticity into brains (the evolution of which is a fascinating but vast topic in itself), one has the tools with which to solve a new problem that arises with the ongoing stimulation of environment-feature detectors. Since the animal brain, as the story runs thus far, is a collection of tarted-up specialist systems cobbled together and switched on by default, the result ought to be utter chaos. As Dennett points out, we have exchanged the problem of what to do next for the "meta-problem" of what to *think* next.

It is all very well to equip oneself with an "All hands on deck!" subroutine, but then, once all hands are on deck, one must have some way of coping with the flood of volunteers (1991:188).

In less complex species, one assumes, the group of inputs from specialist subsystems simply "fight it out;" the strongest signal is the one that temporarily dominates subsequent behavior. We should expect that the specialist offering the most relevant input will usually win such fights. After all, organisms whose neural structure is such that the texture of the soil produces a more salient input than does the approach of potential predators, will soon be weeded out by natural selection.

Dennett enlists an extant model of such competing specialists when he refers to the content struggle as a *Pandemonium* architecture (1991:189). This was the term introduced by Selfridge (1959) to describe his model for contextually-sensitive mechanistic selection between various contentful states in a system. In current terminology, Pandemonium is a production system - a system for determining what word (language is the behavioral coin of choice in artificial intelligence) gets produced in a given situation. Now, there are two broad sorts of production systems, and it is

important that Dennett chooses Pandemonium rather than some other extant model. For Pandemonium is a *broadcast* system, as opposed to a *fetch-and-execute* system. The latter work by the fiat of some central decision-making system that activates the appropriate specialist subsystems when the situation requires it. Pandemonium, however, works by setting all its subsystems on "broadcast" all the time; the analogy is to a myriad of specialist demons all clamoring for attention. In fetch-and-execute models, the agency inheres in the central recruiting and deciding module, while in Pandemonium, the agency is spread around to the various content pathways, the various "word-demons" themselves. The analogy to Dennett's homuncular functionalism is clear. But if these specialists are competing in simple organisms for nothing more than temporary behavioral hegemony, what are the inputs fighting for in the case of human brains?

Dennett answers this question by turning to the evolution of language. Language as we know it, he argues, had its origins in hominid vocalizations whose value consisted in their tendency to stimulate conspecifics to attend to important aspects of the situation at hand. Dennett writes, "I am proposing ...that there was a time in the evolution of language when vocalizations served the function of eliciting and sharing useful information..." (1991:195). And while this in itself is perfectly consistent with language as we know it, Dennett's point is that language in its infancy *only* shared and elicited useful information. Gossip, disinformation, and debates about artistic merit are capacities imposed upon the original function of language, much like the sensory modalities are capacities imposed upon the original functions of ducking mechanisms, symmetry detectors, and the like.

Crucial to this account of the origin of language is the claim that language began as a purely social and public function. This becomes clear in a passage so frank that it is worth quoting in its near entirety.

Then one fine day (in this rational reconstruction), one of these hominids "mistakenly" asked for help when there was no helpful audience within earshot - except itself! When it heard its own request, the

stimulation provoked just the sort of other-helping utterance production that the request from another would have caused. And to the creature's delight, it found that it had just provoked the answer to its own question. ... All that has to be the case for this practice to have this utility is for the preexisting access-relations *within* the brain ... to be less than optimal. Suppose, in other words, that although the right information for some purpose is already *in the brain*, it is in the hands of the wrong specialist; the subsystem in the brain that needs the information cannot obtain it directly from the specialist - because evolution has simply not got around to providing such a "wire."

... Crudely put, pushing some information through one's ears and auditory system may well happen to stimulate just the sorts of connections one is seeking ...

... In particular, we can speculate that the greater virtues of *sotto voce* talking to oneself would be recognized, leading later to entirely silent talking to oneself. The silent process would maintain the loop of self-stimulation, but jettison the peripheral vocalization and audition portions of the process, which weren't contributing much.

... It would be slow and laborious, compared to the swift unconscious processes it was based on, because it had to make use of large tracts of nervous system "designed for other purposes" ... It would be just as linear (limited to one topic at a time) as the social communication it evolved from (1991:195-7).

In this linguistic bottleneck of representations, we have the outline of the origins of serial self-consciousness in a parallel brain. The inputs from the human brain's specialists compete for access to the slow, limited and serial subroutine that yokes the specialists together. We also have here the basic elements of Dennett's central hypothesis about consciousness, which illustrates the purpose behind the discussion of

memes that began this essay.

Human consciousness is *itself* a huge complex of memes (or more exactly, meme-effects in brains) that can best be understood as the operation of a [serial] virtual machine *implemented* in the *parallel architecture* of a brain that was not designed for any such activities (1991:210).¹⁶

The abstract entities thrown into the competition for space in the "subroutine," then, are memes. Recalling the linguistic origin of this subroutine, one concludes that being conscious (in the very important sense of "conscious" that Dennett is exploring) is nothing more than "telling oneself" a series of memes. But who is the self to whom the memes are told? This may be an account of the serial nature of consciousness, but surely it leaves out everything else associated with the experience of being the single subject of multiple representations, both synchronically and diachronically. We may take from Dennett's account thus far that self-consciousness is a succession of subjective seemings (no objective seemings allowed!) and that "I" just *am* a sequence of memes. But this leaves an embarrassing question: Why doesn't it *seem* that I am a succession of memes? Dennett offers only the barest beginnings of an answer to this question, but it is an answer borne out by another compelling view of the self, as we shall see. The seeming unity of consciousness, I will contend in the final chapter, is the result of the systemic hunger of a system's specialists being directed toward the internal states of the system itself. In other words, not all of the representational memes competing for functional hegemony represent the states of the organism's environment; some represent the states of the meme-generating subsystems within the organism itself. This move from

¹⁶ Why is it more exact to refer to "meme-effects in brains" than merely to memes? Even supposing that Dennett has not considered the application of indeterminacy to tunes, *inter alia*, he is clearly aware that beliefs and desires are memes. Shouldn't it be precisely the opposite - that talk of meme-effects is strictly inaccurate? Similarly, Dennett writes elsewhere that the existence of a meme depends on its "physical embodiment in some medium" (1991:204).

It should be clear by now how this sort of careless talk may be unproblematically cashed out. Consciousness is not meme-effects in brains, nor are memes physically embodied in *anything*.

representation to self-representation is at the heart of our sense of unity.

IV. User Illusions and Virtual Captains

i. You're only fooling yourself

Recently the telephone company where I live implemented a computerized system for processing calls that had previously been operator-assisted. For example, suppose Mary makes a collect call. A computerized voice asks her to speak her name after a tone. The device records her voice, and when the person she is calling picks up the telephone, a voice explains that a collect call is being placed by... and Mary's recorded name is inserted into the dialogue. One could easily imagine a situation in which the computerized voice sounds sufficiently like that of Mary that the person receiving the call is fooled into thinking that a human operator is passing on the information. The receiver of the call (the user of the system) would be tricked into thinking that something with *real* intentionality - a human operator - was speaking, as opposed to a computer with mere *as if* intentionality. The user of the system would be falling victim to a variety of what is called the *User Illusion*.

Dennett recruits the notion of the User Illusion to account for our sense of self. The idea, roughly, is that the brain monitors the workings of its own subsystems in addition to monitoring the events of the environment. Now, Dennett's entire approach to the mind is predicated on the claim that a real "semantic engine" - a physical system with real, intrinsic intentionality - is an impossible miracle machine (1987:61). Not even telephone operators, strictly speaking, are intrinsically intentional. But at least some of the brain's subsystems are of sufficient complexity to *simulate* the workings of a semantic engine - in other words, to make it seem as if there were intrinsic intentionality in the brain.

The main thrust of Chapter One was that this semantic approximation achieved by the brain as a whole is what prompts observers to advance useful, though

conflationary, ascriptions of intentional states to a system. Dennett's point in bringing up the User Illusion is to propose that the brain attributes intentional states to itself in exactly the same fashion. One subsystem of the brain monitors another subsystem and is fooled, just like the receiver of the telephone call, into thinking that it has encountered a "real" intentional system as opposed to a well-designed but "stupid" mechanism (1991:311). The report of this encounter, one supposes, is then forced into the content-dogfight with all the other subsystem reports, and becomes, on a recurring basis, a feature of serial consciousness.

This accounts for the bare possibility of a sense of self by defending the possibility of unconscious ascriptions of consciousness. That is to say, without postulating that some part of the brain (or even the whole brain) *is* conscious matter - which would be a miracle - we are able to suppose that some subsystem(s) may render the *purely unconscious* judgement that some other *purely unconscious* subsystem, or the brain as a whole, is conscious. Dennett has already allowed that to be conscious is to comprise a "skein of narratives," but now we see how that comment makes sense; if each subsystem is, in effect, telling a story about that which it monitors, the internal monitors will tell stories about the systems to which they attend under which those systems are genuinely intentional and conscious. In fact, this account turns on the claim, central to Dennett's approach, that the dichotomy between "real" intrinsic intentionality and "mere" derived intentionality is ultimately a false one.¹⁷ Thus, it is as great a mistake to attribute real intentionality to the telephone computer as to predicate it of a telephone operator - or more appropriately, there is no mistake in either attribution, although presumably the richness of the ascriptions will differ.

Still, it is not clear that this account explains the sense of *persistence* that accompanies our conscious reflections on ourselves. After all, it seems to us that we exist over time as entities whose continuance - at least, over the short term - is unquestionable. This appearance hardly follows necessarily from what has been said of

¹⁷ cf. "Evolution, Error, and Intentionality." (1987:287-322).

the User Illusion. Another key concept is needed to explain persistence, namely, that of the *virtual captain*.

ii. Meme, meme, meme and me, me, me

The User Illusion, Dennett argues, operates upon the succession of memes in serial consciousness to produce a separate illusion of selfhood that he terms a "virtual captain" (1991:228). This is simply another variety of seeming, with the "virtual" component of the term borrowed from the computing sciences. In its original context, this term applies mainly to machines; the "virtual machine" is the hardware of a computer plus a modicum of operational software. It is virtual because it is merely a generic machine whose application software enables it to impersonate an application machine *simpliciter*.

It is useful to note that a feature of a machine's being virtual consists in the impossibility of telling, from the outside looking in, whether it is a real function-*P* machine or a generic machine running abstract software *P*. The attraction of this claim is its applicability to the self-monitoring brain. It captures Dennett's point to say that the monitoring system mistakes what the "program" produces - some meme that appears to be in charge - for a real Central Meaner that is genuinely in charge.¹⁸ Now, there is a sense in which each successive victor of the content-dogfight is a virtual captain, since none is real and each is reported by the internal monitors to be *the* real subject. But Dennett also seems to employ the term "virtual captain" to refer to the appearance of a single persisting seat of observation and volition.

On this latter reading of Dennett's virtual captain, or what I will call the

¹⁸ But note that in the case of computers, the only difference between the two kinds of machine is function. When one says that the opacity of the virtual/real distinction is a matter of being "outside looking in," one presupposes that looking inside the two machines could settle the debate, presumably by revealing design features of the hardware that indicate either a specific function or a generic software compatibility. This, however, is precisely the sort of distinction that we could not expect to settle by examining the brain, since there is no owner's manual for the brain by which we might tell the specific from the generic.

captaincy, it is more or less equivalent to the notion of a center of narrative gravity (1991:410). Returning to the talk of skeins of narrative, we are now prepared for the hailstorm of terminology by which we may summarize Dennett's position: The internal monitors operating on the succession of virtual captains will, via a User Illusion of a captaincy, result in reports of a central persisting self. Because these reports will be woven around this central nugget of intentionality, one may conceive of it as a sort of center of gravity. Just as if an object's mass is said to inhere in its center of gravity, so a mind's thoughts, memories, perceptions and experiences will be assimilated to its *center of narrative gravity*. The referent of almost every narrative occurrence of "I" will be this center, this captaincy. When we ascribe mental states to other systems, we ascribe them - generally - to a single narrative core, argues Dennett. Our brains, when weaving our selves, follow exactly the same tactic.

As a positive account of the seemings of consciousness, what I have outlined is, I think, mostly successful. That is, if what Dennett has said is plausible then the seeming continuity of consciousness and singularity of subjectivity can be accounted for without supposing that there is a separate substantial (or homuncular) subject of presentations, or an order of presentations, or even presentations in the standard sense, in the brain. The important qualification, of course, is that of plausibility. Therefore, to the end of vindicating the plausibility of Dennett's meme theory of the self, I will now outline Derek Parfit's theory of personal identity. Parfit's theory, I believe, is congenial to the meme theory of the self, in that their association ultimately benefits both projects; Parfit's theory is deepened and underwritten by a meme view of the self, while Dennett's theory gets tied into a familiar and generally well-received conception of identity and selfhood. Moreover, in the process of drawing out the connections between Parfit and Dennett, the final chapter will provide a forum for a richer account of virtual captains and their attendant illusions than Dennett (or my exegesis) has given thus far.

Chapter Three: Parfit and Identity

Was the vegetating body that survived another decade and more, until August 1900, still the poet and philosopher, artist and Antichrist?

Walter Kaufmann, in the Editor's
Introduction to Nietzsche's
Ecce Homo

I. Deeper Facts, Revisited

i. The problem

Kaufmann answers No to the question he poses in this quote, reasoning that "portraits of Nietzsche in the eighteen-nineties ... show no glimpse of Nietzsche's vanished spirit" (1967:202). But while Kaufmann's purpose in this passage is more to convey the tragedy of a great mind destroyed than to address the question of personal identity, his words illustrate an issue central to diachronic identity: How *does* one decide if temporally distinct selves are identical? Assuming that searching portraits for evidence of the appropriate spirit is a bad method for settling questions of identity, what might qualify as a *good* method?

At least as far back as Locke's prince and peasant, philosophers have used a wide variety of thought experiments as a way of teasing out the essence and implications of personal identity. Such thought experiments, however, have typically agreed in assuming that there must indeed be an answer to be discovered in even the most intuition-defying scenarios. In this respect, Derek Parfit's approach to personal identity represents a dramatic break from the intuitive and traditional notion of a persisting self.

Normally, there are several considerations involved in our judgements of diachronic personal identity, including overall bodily causal continuity (including brain-body weddedness), continuity of memory, behavioral continuity, *inter alia*. Since these relations are rarely divorced from one another in common experience, if indeed they ever are, they are often conflated or overlooked. And when we do consider them in isolation, we tend to regard these relations as actually indicative of some deeper fact of the matter regarding identity.¹⁹ Parfit explicitly states how his approach is directed against this supposition.

Does personal identity just consist in bodily and psychological continuity, or is it a further fact, independent of the facts about these continuities? Our reactions to the 'problem cases' show, I think, that we believe the latter. And we seem inclined to believe that this further fact is particularly deep, and is all-or-nothing - we believe that in any describable case it must hold either completely or not at all. My main claim is *the denial of this further fact* (1976:162n).

Furthermore, the supposed deeper aspect of personal identity must, if we are dealing with identity, be what Parfit calls "one-one". It can only be the case that x is identical both to y and to z if y is identical to z, on pain of denying the principle of transitivity of identity. Parfit argues that what is important in persistence is neither necessarily one-one nor all-or-nothing. The crucial relation, sloppily approximated by talk of identity, admits of both duplication and degrees. Since the alleged deeper fact of selves is postulated as that of which "one-one" and "all-or-nothing" are predicated, the argument that these predicates do not strictly apply to selves is also an argument against

¹⁹ Locke, for instance, thought that memory was a sufficient condition for identity (indeed, he thought it was the whole story), and many would agree, but on the other hand, we generally accept bodily continuity as a sufficient condition as well. We must be confronted with bodies in deep, long-term comatose states, or with truly catastrophic cases of cognitive breakdown, before we are prepared to believe that bodily continuity is an insufficient criterion for identity.

the deeper fact, in that it undermines the motivation for talking of such a fact of the matter.

ii. Parfit's solution

The thought experiment from which Parfit's argument derives much of its force is one proposed by Shoemaker (1963) and expanded by Wiggins (1967). One is asked to imagine half of my brain being transplanted into a new body, retaining all memories and behavioral dispositions. Parfit assumes that we will all agree that the resulting person is still me. Now, he asks, suppose that my brain is divided into two portions and both halves are transplanted into two separate bodies, such that each body now has my memories and personalities. There are, Parfit suggests, only three ways to describe the results. "(1) I do not survive; (2) I survive as one of the two people; (3) I survive as both" (1976:144).

Parfit rejects the first possibility on the grounds that we allowed that I survived when half of my brain was moved into a single new body. "How could a double success be a failure?" he asks (1976:144). Anyone claiming that I do not survive this operation would be like someone

who, when told of a drug that could double the years of his life, regarded the taking of this drug as death. The only difference in the case of division is that the extra years are to run concurrently. This is an interesting difference. But it cannot mean that there are *no* years to run (Parfit 1976:147-8).

Since (2) and (3) are, technically speaking, consistent with each other as Parfit has written them, it is worth noting that Parfit actually takes (2) to mean that I survive as *only* one of the two resulting people. (2) cannot be the best way to describe the case, he argues, because there is simply no more reason to think that just one of the two

people, and not the other, is really me than there is to think otherwise. There are no grounds for making a principled distinction between the two (1976:144). At this point, it begins to emerge that the interpretationist approach is one that Parfit shares with Dennett, whose appeal to a lack of principled distinctions is a familiar theme.

This leaves the third option, that I survive as both of the two resulting persons. But this option conflicts with one of the demands of identity - that identity be "one-one" - since the two resulting persons can be supposed to diverge in their experiences, memories and personalities following the fission procedure. So the most coherent way of describing the relations obtaining between myself and the two products of the operation is incompatible with identity. The solution, argues Parfit, "is to give up the language of identity. We can suggest that I survive as two different people without implying that I am these people" (1976:146).

II. Memory, Continuity, and Similarity

As Parfit notes, the rejection of identity as what matters in persistence has little or no practical effect on what does matter. In other words, fission into indistinguishables never occurs, nor do brain-body switches. But the divorce *in principle* of identity from what matters means that there is no reason to suppose that what matters is all-or-nothing. This enables Parfit to suggest that *psychological continuity*, which, unlike identity, is a relation admitting of degrees, is what matters in judgements of the persistence of the self.²⁰

Psychological continuity is defined as comprising a set of relations, the most important of which is memory. However, Parfit recognizes the danger that perhaps memory implies identity - i.e., that perhaps my remembering having had an experience

²⁰ Parfit claims "that we use the language of personal identity in order to imply such continuity" (1976:149). But actually, this is a stronger claim than he needs. Perhaps we commonly use the language of identity to imply identity! Such reportive dealings are of no concern to Parfit; all he needs is the stipulative claim that judgements of identity are replaceable *without loss of anything important*, by judgements of continuity.

implies that it was *I* who had the experience (1976:152). There is, after all, the powerful feeling of being a self unified over time to be explained, and memory "from the inside" seems like compelling evidence to the effect that something substantive does persist and thus accounts for the apparent diachronic unity. The tendency to take memory as proving persistence is, I believe, what most tempts us to propose some deeper fact of selfhood in virtue of which selves persist. Parfit's attempt to undermine this temptation begins with his definition of a new relation "*q*-memory." *Q*-memories need only feel like memories, and represent experiences that *someone* had. In other words, they are memories of experiences that need not have originally been the experiences of the remembering subject.²¹

Parfit uses the concept of *q*-memory to urge that the apparent memory of one's having an experience does not imply that the recalling subject is identical to the original subject of the recalled experience. He writes that

it cannot be a part of what I seem to remember about this experience that I, the person who now seems to remember it, am the person who had this experience. That is something that I automatically assume (1976:152).

Now, I happen not to remember experiences originally had by a subject in a different body than my own - that is, a body not spatio-temporally continuous with my own. As Parfit points out, if we did have such memories we would be less hasty to assume that every experience we remember "from the inside" is an experience originally had by a

²¹ Although Parfit sticks with this terminology throughout his discussion of memory, I believe he need only show that the concept of *q*-memory is coherent. For if *q*-memories are indistinguishable from memories save that they do not presuppose subject identity, then clearly memory itself need not presuppose subject identity. It might be argued that assimilating *q*-memory to memory is a violation of the ordinary language use of "memory," since *q*-memory permits one to err about whose experience is being remembered, while memory might be supposed a factive faculty. It seems to me that there is something to this objection, but consider other factive capacities: while I can neither "misknow" nor "misrealize," I can certainly misremember. Fortunately, for my discussion all that matters is that such errors are possible. Nothing much hangs on what we call them.

subject identical to the current self. But in the event of the different original subject being "located" in an earlier instantiation of my own body, how could I ever distinguish my *q*-remembering that subject's experiences from my actually remembering my own earlier experience? Since there would be no telling apart the two sorts of memory if the experiencer and the rememberer shared the same body, the intuition of persistence in memory does not actually imply persistence of the self. The same data are compatible with a succession of selves accessing their predecessors' memories from the inside; the feeling of a self unified over time would be preserved in such an arrangement.

Relations such as memory exhibit continuity when they form "overlapping chains," Parfit tells us. This is to be contrasted with *psychological connectedness*, which he offers as another important factor in assessing persistence. Connectedness is the relation that obtains between an experience and the *q*-memory of the experience. It is these direct connections that Parfit supposes can overlap and give rise to continuity.

While I have no quarrel with the emphasis Parfit places on continuity and connectedness as being what matter in judgements of "identity," I would like to suggest that his account gives short shrift to the aspect of *similarity* that generally obtains between temporally distinct selves. I am imagining similarity as a component relation of psychological continuity, one which usually weakens with temporal distance. This allows that there is an important difference between the me-today/me-tomorrow relation and the me-today/me-in-fifty-years relation, without even insinuating that this difference is more than a matter of degree.

While neither continuity nor similarity implies the other, they are related in at least this fashion: a story must be told to account for an assertion of continuity in the face of clear dissimilarity. Similarity between temporally distinct selves in a spatio-temporally continuous body is, I submit, generally taken to be indicative of whatever matters in persistence, which Parfit has identified as psychological continuity. But in the absence of similarity, continuity may yet consist in some relations that explain the features of a later self in terms of transformations that occurred to an earlier self. These relations, I think, are part of what Parfit endeavors to express as "overlapping chains."

But what sort of chains? For Parfit, as indeed for most philosophers of the mind and the self, the obvious answer would be *causal chains*; how else to explain the transformation of an earlier self into a later?

III. The Answer is: No Answer

Parfit has presented a view of the self in which identity has been replaced by psychological continuity. I submit that the three noteworthy features of diachronic continuity of the self are: (1) that the temptation to posit a deep fact about persistence is the result of a misinterpretation of a certain kind of memory; (2) that what matters in personal "identity" is a matter of degree; and (3) that the particular degree of continuity between selves, if any, is a matter of interpretation - there are no further facts by which to settle ambiguous cases. Parfit argues that the adoption of his view of the self enables one to see that, "for quite unpuzzling reasons, there *is* no answer to a question about identity" in any absolute sense (1976:146). Parfit is a Physicalist about psychological states in the sense given in Chapter One. Yet one his attitude toward selves may be categorized as "interpretationist." That is, even if one holds that the mental states of a system are facts to be discovered, one may yet hold that listing those facts does not answer the questions of whether the system is a self, and whether it is the "same" as some earlier self. The important issue upon which Parfit may *not* be seen as an interpretationist is that of continuity, since psychological continuity, on his account, comprises overlapping causal chains.

It should be clear that one of the first steps in applying a meme theory of the self to Parfit's account of identity will be a denial of causal psychological continuity. But this is not to deny the importance and utility of the Parfitian account. Our investigation of his position has left us two important themes that form a nexus with meme theory: first, continuity and similarity capture what matters about persistence, and second, insofar as diachronic identity *seems* to obtain, there *seems* to be some deep objective fact about selfhood. Parfit suggests that our commonsense intuitions tell us that identity is predicated of this deeper fact about persons, independent of the obviously different properties of a young boy and an aged grandfather, but every bit as substantive as considerations of bodily continuity, memory, and the like. His major diagnostic insight is that we need not think of the diachronic relation between selves as being one of

identity, and thus we need not posit any further objective fact about selves.

But recall also Parfit's claim to be presenting the *denial* of this further fact. We are now in a position to see that Parfit does not actually render untenable the view that there is such a deeper fact about identity. Rather, he has succeeded in showing that what seems to matter about this alleged further fact can be captured by the uncontested facts of psychological continuity. Now, this is an interesting thing to show, but if our pre-critical notions of a self continue to promote at least the suggestion that selfhood is indeed something deeper than body, age, or memory, then Parfit's argument is inconclusive. (His argument from Wiggins' thought experiments is also open *prima facie* to a generalization objection like the one described in Chapter One above). In other words, Parfit has not said anything about the alleged further fact of the self that would force one to reject any account that proposes such a fact. Indeed, one might concede that psychological continuity seems to capture the important aspects of personal identity, yet still claim that what matters in persistence over time is the identity of some buried nugget of selfhood.

It is a flaw of Parfit's treatment of diachronic identity, though not a fatal flaw, that he does not ground his view on an underlying account of the self, and thus does not present a conclusive argument for denying what he wishes to deny. Parfit, it seems, is content to perform an analysis of the way we talk about personal identity, concluding that talk of the relations commonly thought to imply identity is sufficient to capture what matters to us about persistence. If we fail to recognize that these relations fundamentally differ from that of identity, "we shall be led into quite ill-grounded attitudes and beliefs" (148).

There are two reasons why Parfit's omission of a theory of the self is not lethal to his position. First, his view enjoys the benefit of utility, proving itself useful in defusing the paradoxical violations of transitivity common amongst the thought experiments of modern personal identity debates. And second, there is indeed a theory of the self that lends compelling support to Parfit's view; specifically, the meme theory of the self.

Chapter Four: The Meme Theory of the Self

For my part, when I enter most intimately into what I call *myself*, I always stumble on some particular perception or other of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch *myself* at any time without a perception, and never can observe anything but the perception.... If anyone, upon serious and unprejudiced reflection, thinks he has a different notion of *himself*, I must confess I can reason no longer with him. All I can allow is, that he may be in the right as well as I, and that we are essentially different in this particular. He may, perhaps, perceive something simple and continued, which he calls *himself*; though I am certain there is no such principle in me (Hume 1739, Selby-Bigge, ed. 1967:232).

The purpose of this chapter is to enhance the plausibility of the meme theory of the self by showing its useful relations both to the views of Parfit discussed above, and to the metaphysical issues surrounding a certain psychological disorder. Above I identify three features of particular importance to be taken from Parfit's approach to personal identity: namely, the centrality of a certain sort of memory, the claim that what matters in identity admits of degrees, and the further claim that the assessment of these degrees is a matter of interpretation. I will begin this final chapter by illustrating how these most important aspects of Parfit's theory are underwritten by the meme theory of the self. To this end, I will first present a working summary of the meme theory of the self, followed by a closer analysis of what I see as different levels of description within the theory. This will lead into a discussion of how the meme theory specifically underwrites the most crucial components of Parfit's theory, and thus may be closely associated with that influential and widely accepted view. Finally, I will present the relation between the meme theory of the self and the fascinating psychological condition called *Multiple Personality Disorder*, arguing that the meme theory gains plausibility from its capacity

to make metaphysical sense of the appearance of multiple selves in one body.

I. Description of the Theory

As we have seen, Parfit's analysis places the blame for our mistaken intuitions about the persistence of the self upon a tempting, but ultimately false, belief in some deeper fact of personal identity. I believe that the meme theory of the self both diagnoses our adoption of this belief, and bears out Parfit's rejection of this supposed deeper fact. Here is a summary of what I will defend as the meme theory of the self:

The self is an abstract homuncular construct, comprising at one level of description a constellation of memes, and at another level a succession of memes that serially dominate as the virtual captains of the organism. The self is seemingly unified over time in virtue of the relations between present virtual captains and past ones, and in virtue of a certain kind of compatibility obtaining between temporally distinct meme constellations in a single body.

The development of this summary hinges upon the explication of virtual captains and their relation to unity. This explication will lay the groundwork for a discussion of the synergetic connections between the meme theory and the first crucial theme of Parfit's theory - that of memory.

i. Memory, memery and virtual captains

The identity of the consciousness of myself at different times is therefore only a formal condition of my thoughts and their coherence, and

in no way proves the numerical identity of my subject. Despite the logical identity of the "I", such a change may have occurred in it as does not allow of the retention of identity, and yet we may ascribe to it the same-sounding "I", which in every different state, even in one involving change of the [thinking] subject, might still retain the thought of the preceding subject and so hand it over to the subsequent subject (Kant 1781:A363, Kemp Smith trans.).

Recall Dennett's account of the meme-struggle, the "content dogfight" for functionally salient space, with the temporary victor occupying the role of a *virtual captain*. In this fashion the traditional notion of a central subject and agent perusing the content of consciousness has been replaced by a succession of homunculi, each comprising the content of consciousness. Now, this way of expressing the meme theory is plainly unhelpful if one is trying to develop Parfit's view of the way selves *change* over time. Taken at this level, the meme theory becomes an account of the mutability of persons that describes consciousness as little more than a diaspora of psychological states - and recalling Chapter One, *abstract* states at that. This is clearly a far more radical conclusion than Parfit intends to advance; he thinks that selves change with time, but obviously has in mind a timescale on the order of months or years rather than seconds or even milliseconds. What this aspect of the meme theory *does* provide, though, is an excellent account of the apparent unity - synchronic and diachronic - of the self. And since my concern is mainly (but not exclusively) with diachronic unity, I will emphasize the diachronic issue and argue that the notion of a virtual captain seems tailor-made to explain the special sort of memory central to our sense of persistence.

In the introduction to virtual captains given in Chapter Two, I presented the claim that the "epistemic hunger" of representational systems could be recruited as the beginnings of an explanation for the phenomenon of the unity of consciousness. This is especially clear if the sort of unity involved is synchronic, by which I mean the unity carried in the experience of being a single subject of compresent representations. If we

posit specialist subsystems whose representational domain includes various other meme-bearing subsystems, we open the door to the possibility of memes that act to capture in themselves multiple other memes.

One might conceive of this as analogous to "multiple tracking" in the recording of music. In this process, a musician first records one instrument onto a channel of audiotape and then, while playing back that recording, plays another instrument through another channel. Both the recorded track and the live performance are imparted to a third channel, and this third track, when played back, creates an illusion of two instruments being played simultaneously. In this fashion a single musician may impersonate an entire band by creating a single recording that ultimately captures many separately recorded instruments in one channel. Similarly, the meme that eventually gains temporary dominance in the struggle for functional space may well contain at least partial information from various sensory modes, from memory traces that have been activated (for whatever reason), and indeed, from the linguistic bottleneck occupied by the previous virtual captain. In other words, we have here a snowballing process in which representations are captured in more complex representations. A central subjective faculty simultaneously perusing each of the meme pathways is unnecessary on this account, since all we need posit is a collection of subsystems structured such that this snowballing may occur.

Furthermore, there is nothing to say that the meme that does become virtual captain must be one that represents any other meme, although unless the task at hand requires all of one's attentional resources the advantages of "keeping tabs" on a variety of internal and external states are clear. There is certainly no reason to suppose that every victor of the meme-struggle must be one of the snowballed memes; the meme theory's explanation of synchronic unity may be made to account for those occasions that find us utterly absorbed by a single aspect of the environment, or absorbed by an internal state (such as a memory) to such an extent that our surroundings are ignored. One need not, therefore, posit any phenomenologically implausible continuous awareness of unity, since the meme theory permits us to describe why we experience compresent multiple

representations *when we in fact do*, and is congenial to the possibility that we often do not. This excursion into synchronic unity provides the basis for a discussion of diachronic unity, and particularly for a view of diachronic identity that supports Parfit's comments on the role of memory "from the inside" in promulgating the conviction that the self persists - although it has other diagnostic value, as well.

When Dennett presents consciousness as criterial upon memory (by arguing against unremembered moments of consciousness), it is plainly short-term memory to which he refers; his arguments proceed from situations in which the relevant time spans are too short for the subject to report what seems to be the case, and Dennett argues that there is no seeming without the saying. And uncontroversially, some - not all - of the contents of short-term memory are stored as long-term memories. Since consciousness, on this account, just is a series of virtual captains, this means that some of the virtual captains of the self leave long-term memory traces.

I propose that experiencing a memory from the inside consists in the current virtual captain's having as one of its contents a memory meme that was once a virtual captain itself. In getting incorporated into the new virtual captain, the traces of the old one may be synthesized with other competitors in the meme-struggle (some of which may themselves be former virtual captains). The new captain, having donned the garments of its predecessor, supposes itself to have been the original subject of the remembered experience, and thus assumes its own persistence through time. This, of course, is strikingly similar to the position presented by Parfit on the relation between this sort of memory and the belief in persistence, but I am presenting much more than a restatement of Parfit's views. His conclusion, after all, was that memories from the inside could not discriminate between a genuinely persisting substantive self and a series of selves capable of accessing some common memories. But I am contending that the former option is impossibly mysterian, that the latter option is the best way to conceive of the self, and that a story may be told about the specific dynamics by which a succession of virtual captains of the self uphold an illusion of diachronic identity.

There is something more to be said about the value of the meme theory in

permitting Parfit to diagnose the mistaken (as he sees it) belief in a persisting self. At the other end of the spectrum from the diaspora of competing memes, we find the functional position *occupied* by the virtual captain - that is, we find the captaincy. This, I submit, looks like a promising candidate for what gets mistaken for a deeper fact of identity, since even the meme theory proposes that there is a persisting captaincy even though the particular meme holding the position may change from moment to moment. This is what I had in mind when I (or someone, anyway) wrote above that the meme theory allows us to appeal to more than just memory from the inside when we attempt to explain why the belief in a persisting self is so tempting.

Since each consecutive virtual captain finds itself in the captaincy, the first-person referent of each consecutive captain is "I." But to what could "I" refer? The sense generally attached to "I" is that it refers to one persisting thing, namely a captain. But we have seen that captains do not persist. This means that it refers to different agents, different memes, on every occasion of its use. But if the theory is taken at the level of the captaincy, "I" may be said to refer to the functional position itself. Insofar as the captaincy may be said to persist, then, one might suppose that there is indeed some deeper fact of persistence of the self. The problem with this supposition, and the reason why Parfit would reject it, is that a captaincy is precisely the sort of "thing" about which *substantive* selfhood over time could not be predicated.

The problem is not that the successive constituents of the captaincy are virtual, nor that they are abstractions themselves. Take the example of an actual military unit. We must realize that while there may be more or less straightforward conditions for the diachronic identity of the position itself, this is of no use in determining whether the same captain or a series of captains have occupied the position. Captaincies and transcendental unities of apperception, *inter alia*, are not the sort of "thing" capable of meeting or failing tests for the identity over time of a substantive subject, since positions and logical unities can be neither substances nor subjects. Whatever it is that we imagine persisting through time, we suppose that that thing is the subject of our experiences and the source of our volitions. If we purchase the persistence of the self at the cost of

abandoning the agency and subjectivity of what persists, the price has clearly been too high.

ii. The self as a constellation

I have not yet made clear how Parfit's conception of the self changing over time maps onto this model of the self. On a level falling somewhere between the "diaspora" of Pandemonium-stricken memes and the logical construct of the "captaincy," we may treat the self as a more or less stable *meme constellation* that generally changes only slowly. In so doing, we refer not just to the fleeting contents of consciousness, but to the recurring and identifiable memes for the behavioral patterns typical of the particular body in question. Crucially, this is the level that captures the aspect of the mutability of the self in a fashion apt to express Parfit's notion of psychological continuity - or, as I will argue below, at least *some* notion of psychological continuity. A remembered experience, taken individually, is only a meme like any other, incorporated into the current virtual captain. But we may speak also of the store of memories, meaning thereby the set of memes representing past experiences that tend to achieve functional dominance.

This constellation facet of the theory of the self gives us reason to believe that another way to express what matters in persistence is to speak of psychological *consistency* rather than merely of continuity. To understand this, consider the problem of persistence framed in terms of the ownership of actions: from the first person, one might say that one is no longer the same person as a *q*-remembered person if one denies ownership of actions that one *q*-remembers the earlier self performing. This is a useful way of framing the problem, since it allows us to bring in familiar examples that do not necessarily involve a mutation from one self into another. Often I do and - especially - say things that shock or horrify me even as the words leave my mouth. My immediate response, and the response of most people in similar situations, is simply to deny ownership of the offending action or utterance. "That wasn't me!" is an excuse many

of us have employed - not meaning that some other body produced the behavior, nor even that some other self owned it, but rather that the offending behavior is without an owner altogether. The meme ascribable to account for such behavior is inconsistent - in some broad, not necessarily logical, sense to be discussed below - with the constellation of memes constitutive of the self as seen from this level.

This, we may suppose, is what happens slowly as the elements, key and trivial, of a meme constellation become replaced by new memes. Gradually, in a process that will clearly admit of degrees, a constellation changes such that it may, as a whole, grow increasingly inconsistent with some of its memes of remembered experience. These memories may either be rewritten to be consistent with the new constellation, or be retained while yet being disowned. "The person I used to be did those things," one would say under the latter option, "but I am no longer that person."

An emphasis on consistency need not be a challenge to continuity as the hallmark of what matters in persistence. Rather, the two may be seen as converging upon a common idea. Moreover, the notion of *similarity* within the self converges upon the same idea of some special sort of compatibility; whether we name continuity or consistency or similarity to be the fundamental concept is immaterial.

This account of the relation between similarity and continuity helps to undermine an objection to Parfit's view raised by Patricia Kitcher, who champions Kant's notion of causal continuity as the whole story in identity.

Modern mentalists [like Locke and Leibniz] take memory, belief, and desire continuity to involve causal connections among mental states. Current accounts [like Parfit's] differ in stressing the *similarity* of a person's beliefs and desires across time. What justifies the claim that mental unity requires similarity of beliefs and desires, in addition to causal connection, however? No one knows how much similarity in belief and desire people have over a lifetime (1990:129).

Kitcher's criticism proceeds implicitly from a presumed tension between mere continuity and continuity plus similarity as what matters in persistence. But Parfit's message in its most basic form is merely that diachronic mental unity is a matter of degree. This hardly entails (whatever else Parfit says) that degree of *similarity* is all that matters, although similarity certainly admits of degrees. Rather, as we have seen, the property admitting of degrees that Parfit stresses most in his characterization of mental unity is psychological continuity. Psychological similarity, one might argue, is important precisely because (given bodily continuity) it often engenders a judgement of continuity. Of course, Kitcher is right to wonder why causal continuity should be thought to imply, or even suggest, diachronic similarity. Causal continuity carries no such implication. One recognizes, however, that similarity is crucial nevertheless.

The tension between these two concepts may be resolved by effecting a fundamental shift in the way one views psychological states; it is the shift proposed by Dennett in regard to propositional attitudes, and expanded by myself to the class of memes. I, unlike either Parfit or Kitcher, have it at my disposal to claim that the relevant continuity is not causal in nature, but something altogether more abstract. The task, then, is to explore what sort of continuity, expressing the idea of compatibility bordered by causal continuity, consistency, and similarity, obtains between temporally distinct meme constellations when a judgement of identity seems warranted. The relation proposed by the meme theory is one that bears out the latter two of the three crucial Parfitian themes discussed above: that persistence is a matter of degree, and that our judgements of what constitutes sufficient degree are strictly a matter of interpretation.

II. Continuity, Causality, and Narrative

For the Physicalist, the mental states between which relations of psychological continuity obtain are realized as brain states (or, at least, as physical states). But, given the meme theory's rejection of Physicalism, Dennett and I owe an answer to the question: Continuity of *what*? The answer suggested by Dennett is: The continuity

conferred by being part of a narrative. Postponing the issue of how literally I intend the "self as narrative" to be taken, I will begin by drawing out the properties shared by narratives and Parfit's persisting self.

Parfit's interpretationism toward the self may be made more palatable to sceptics if one is careful to point out that, in general, this approach preserves most of our practically important views about personhood. Indeed, we have already seen that this is one of Parfit's central themes. *Identity* is not crucial, but what *is* crucial is the significance that we attach to judgements of the psychological continuity mistaken for identity in everyday life. He claims that

judgements of personal identity do derive their importance from the fact that they imply psychological continuity. ... when we can, usefully, speak of identity, this relation is our ground (1976:150).

It is only when we cannot speak of identity usefully that the interpretationist underpinnings of the self as narrative make themselves known. With regard to the narratives of the self, there will almost always be consensus amongst me and my peers as to whether I am identical (in the loose and mistaken sense) to the self using this name, body and social niche yesterday. But, as the thought experiments surrounding personal identity indicate, situations could conceivably arise in which it is unclear that psychological continuity obtains between me and some future self - conceivably despite that future person's protests that it *does* obtain.²² In such a case, it would be especially clear that whether the relation obtains is a matter of interpretation, not of fact. Some examples from literary narratives may help illustrate this point. What happens, for example, when one tries to ascertain the continuity of a narrative?

Here is a less abstract way of asking the same question: Which exhibits a greater continuity of narrative, Burroughs' *Naked Lunch*, Ondaatje's *Coming Through Slaughter*,

²² Indeed, one needn't plumb thought experiments for such cases; recall Kaufmann's puzzlings about (who may or may not have been) Nietzsche.

or Heller's *Catch-22*? This is a question to which there is no objective answer. One might define some set of criteria by which continuity is measured, but such a set would, within broad parameters, be arbitrary and subjective. For which is more important: continuity of theme, of character development or of plot? And even this question presupposes agreement on what constitutes continuity in these facets of narrative. Like beauty, continuity may be in the eye of the beholder.

Consider the results of appending *To Kill a Mockingbird* to the end of *All Quiet on the Western Front*. Published in one volume under the name of one author, is it an objective fact that there are or are not two distinct narratives here? The fact that this hybrid is contained in one volume is loosely analogous to bodily continuity, while its having credited only one author loosely corresponds to a claim from a current self to be identical with a remembered self.²³ Here is a more difficult case: Is Plato's *Republic* one book (as publishers evidently think) or ten (as some argue) or some number in between? This is a less certain example than my absurd constructed one, with various people's intuitions evidently leaning in different directions, but the method for answering the question is the same for both examples; one can only appeal to a consensus among interpretations. Certainly I am not claiming that the continuity of persons is as grossly disjointed as these literary examples. But then again, the practical consequences attendant upon judgements of personal identity are often so grave that we do not need a particularly disjointed case to knock our intuitions askew. My point is that, when we are uncertain about the continuity of persons, we have as little recourse to disambiguating facts as when we are uncertain about the continuity of literary narratives.

While the author's opinion is given as much weight as anyone's in reaching this consensus, no mysterious power accrues to the originator of the narrative to dictate what the narrative *really* means, as opposed to how it is conceived. Were Dostoevsky to return from the grave and inform us that *The Brothers Karamazov* was intended to be

²³ The correspondence is loose since two distinct narratives do not imply two separate authors in the literary realm. Even if a text comprises two narratives with crashing clarity, one author might have created both. In other words, literature permits a distinction between author and narrative that I do not recognize (except in a special and strained sense that I discuss below) in discussing the self.

printed in the middle of *The Double*, he would not automatically create narrative continuity in the resulting unwieldy tome, assuming anyone would even provide it with bodily continuity by publishing it. There is no such authorial veto power over *what is the case*, because there *is* no case, no fact of the matter, about how many narratives there are. There are only interpretations that are more useful or less useful - and even these rankings depend on what the interpreter finds useful.

Narrative thus seems to capture the right brand of continuity to underwrite the interpretationist variety of persistence demonstrated by the Parfitian self. Or to state my position more clearly: the self, taken as a persisting constellation of psychological states, is a kind of narrative. Does this come to grief on the obvious objection that narrative continuity is not causal while psychological continuity is? No, because that objection is not compelling. Causal continuity cannot be the relation obtaining between the mental states of a self, since such states, as I have argued, are not causally efficacious. Nor, as Kitcher points out, does similarity of mental states seem a compelling necessary condition for unity, since it appears at least possible that sharp dissimilarity could obtain between temporally distinct instantiations of causally connected selves. Yet both criteria seem intimately related to our conception of an enduring or unified self, insofar as dissimilarity over time of meme constellations within one bodily continuity must be explained in terms of psychological continuity if a judgement of persistence is to stand.

Both psychological continuity and similarity, on my account, are captured by the notion of narrative continuity. This notion expresses the relation between two constellations of psychological states when one seems to follow from the other, in the same sense that a section of a narrative seems to follow from the previous section: neither logically, nor causally, nor necessarily by more than superficial similarity, but by the fact that one is predictively and descriptively better served in regarding the two sections as part of a unified narrative than in supposing otherwise. While these sorts of relations may obtain *within* the narrative, they do not characterize the relations between components of the narrative. For example, within a certain text, a direct causal chain obtains between Eilbo Baggins' theft of the Ring and Frodo Baggins' losing a finger, but

in no sense does the narrative of *The Hobbit* cause the narrative of *The Lord of the Rings*. That any continuity relation at all obtains between these two narratives (or are they one?) is a matter of interpretation. Thus, when intuitions are split into incompatible camps, or when nobody has any particularly firm sense of clear continuity or discontinuity in the narrative, the lack of a settled interpretation is neither puzzling nor *conceptually* problematic (although in the case of personal identity it may be legally or emotionally puzzling, to say the least). If this sort of continuity is the sort by which the psychological states of the self are related, it would bear out Parfit's claim that questions about tendentious cases of personal identity are *empty* in the same sense that the question, "Was England the same nation after 1066?" is empty (1976:146).

III. Ironing Out Some Wrinkles

i. From whence the self?

One awkward aspect of this treatment of the self is the apparent implication that the self is a self-construction - as tidy an infinite regress as one might hope to never encounter. Dennett, however, offers what I think is exactly the proper avenue to defuse this problem before it becomes an embarrassment. The self, he argues, should be treated as part of the *extended phenotype* of a contemporary human being - something which, given the appropriate developmental conditions, humans create as a method of modifying their environments for various reasons relating to selective advantage. The extended phenotype is a concept first introduced by Dawkins (1982), who argues that the phenetic effects of genes should be examined beyond the scope of the physical development of the organism. Genes, he suggests, may be selected for on the basis of the behavior they typically engender; such behavior and the objects it employs ought to be considered almost as much a part of the organism's phenotype as the organism's body. Thus, nests are part of the extended phenotype of birds, as are shells for crabs (even hermit crabs, which are genetically programmed to scavenge a shell) and dams for beavers.

The value of this view lies in its placement of the construction of selves (as opposed to self-construction) in the realm of unconscious actions, in the same camp as the construction of webs by spiders (Dennett 1991:416). *Humans*, or human brains, and not (necessarily) persons, spin the narratives of selves on this account. The author of the self, therefore, need not be an author in the conventional, richly intentional sense.

Of course, one might think that the brain merits sufficient attributions of beliefs and motives in its construction of selves that it too deserves to be called a self. And Multiple Personality Disorder, as will be discussed below, gives us reason to impute motives to the brain regarding its *manipulation* of (something approximating) different selves, since some conation underlies the way that various personalities dominate according to situational demands. But the homunculus hiding in this imputation is no more problematic than the brilliant engineer lurking in a similarly functionalist explanation of beaver behavior, or the artistic genius ascribable to spiders. We have already seen (in Chapter 2 above) how such homuculi are to be cashed out.

ii. Narrative and agency

One need not be blessed with a spectacular imagination to anticipate the shock that will emanate from some quarters over the theory I have explicated thus far. Perhaps the most counterintuitive aspect of the claim that the self is a narrative, and the aspect apt to generate the most resistance to the theory's acceptance, is the disanalogy between the passivity of narratives as we generally conceive them - i.e., as literary constructs - and the volitional character of selves as we experience them. I propose to nip in the bud this source of resistance to the theory.

Dennett does not say much about agency as the meme theory presents it, but I think the way to rectify this is to stress the deep-running parallels between what meme theory does for subjectivity and what it may do for agency. On Dennett's Pandemonium model of language and behavior production, recall, Dennett maintains that we must literally propose words *wanting* to get themselves said (1991:242). A developed account

of the Multiple Drafts model has not merely the capacity for meaning but the capacity for willing, spread around to a mass of homunculi or word-demons competing for functional saliency. My contention is that these lesser homunculi provide the appearance of central agency in precisely the same fashion that they provide the appearance of central meaning. Their agency drives the engine of an illusory will.

This approach fits with the phenomenology of volition and it is worth taking a moment to reflect on how nicely. The struggle between content-demons provides as straightforward an account of deliberation as one could desire. Furthermore, the notion that a succession of memes are the serial executors of the will gives us a way of conceiving of "being torn" between two desires. When two memes are nearly evenly matched, the temporary victory of one (and consequent behavioral effect) will engender regret - perhaps even loathing - when the other subsequently dominates. Finally, my account admits the data regarding the long term objects of our conation. We have already explored the level of description from which the meme theory presents the self as a constellation of memes that may change only slowly over time. Given the relative stability of this constellation, many intents will recur and synthesize with other goals and representations. Thus we may account not only for the mass of beliefs and desires that we typically hold for most of our lives, but also for the fact that we often have long term - perhaps lifelong - goals and plans.

IV. MPD as a Test

While I take it to be a credit to Parfit's view that it resolves certain difficult thought experiments in the personal identity literature, this is unquestionably a limited endorsement. I have argued that, when bolstered by the meme theory, Parfit's view furthermore accounts for the dramatic changes in personality that can occur over long periods of time. This is a more practical, and thus a more ringing, endorsement of both Parfit's theory of identity and the meme theory of the self. Similarly, the meme theory that unites the views of Dennett and Parfit gains credibility from the way in which it

deals with one of the most intuitively unlikely phenomena in clinical psychology: Multiple Personality Disorder, or MPD. I propose to develop the plausibility of the meme theory by putting it to work in making sense of MPD. This will involve first deflecting the criticisms of those who deny the legitimacy of any trade between theories of the self and the phenomena of MPD, and then describing how the meme theory applies to MPD. In particular, I argue that the meme theory of the self dissolves the appearance of a metaphysical problem with the notion of more than one self in a single body.

MPD is characterized by sudden changes between recurring and relatively stable patterns of behavior within one body, such that the body can appear to be "occupied" by more than one self or person. Bolstering this appearance may be phenomena as suggestive as a body's referring to itself using different (recurring) names, demonstrating unique and reidentifiable personality types, and generally different behavioral patterns across the board. Since these shifting patterns of behavior engender incompatible intentional ascriptions, one winds up attributing different broad likes and dislikes to the various "alters" of the MPD victim's body; indeed, these likes and dislikes may be deeply incompatible across alters. The difference may run to skills as well, in that an alter may demonstrate abilities that others deny having altogether.

Now, "alter" is technically a neutral term with respect to the potential selfhood or agency of whatever subsumes the different behaviors (as is "personality" when used in the label of the disorder, one supposes). But the most extreme cases of MPD certainly move one to wonder about the conceivability of different selves existing in one body. Is it miracle-mongering to suggest such an explanation of MPD phenomena? Hacking reports that the consensus within "the multiple movement" (meaning those psychotherapists who recognize MPD as a genuine disorder rather than a theorist's artifact) is that "killing off" or simply excising alters is wrong at least partly because one would be destroying a real person (unpublished:20). But is it intelligible to speak of multiple selves in a single body? What metaphysical contortions does this put one through?

On the meme theory of the self, none. We have seen in Chapter One the constraints operating upon the ascription of intentional entities to physical systems, with the ends of predictive power and descriptive intelligibility subsuming all. But if selves and their constitutive memes are pragmatically expensive, they are metaphysically as cheap as can be. Once we resort to the intentional stance to make predictions about a physical system, we have multiplied entities in a strong metaphysical sense, proposing abstractions as causally efficacious things (for it is only as causally efficacious that they are worth ascribing). How many such entities we postulate thereafter is a function of what requires explanation and prediction - a pragmatic matter - but metaphysically the damage has been done, although my point once again is that it need not be a problematic move if we understand it when we make it. Dennett writes:

The idea of MPD strikes many people as too outlandish and metaphysically bizarre to believe - a "paranormal" phenomenon to discard along with ESP, close encounters of the third kind, and witches on broomsticks. I suspect that some of these people have made a simple arithmetical mistake: they have failed to notice that two or three or seventeen selves per body is really no more metaphysically extravagant than one self per body (1991:419).

But some philosophers have other reasons for treating MPD as a dubious phenomenon. Before I embark on my account of how the meme theory of the self makes sense of MPD, it is incumbent upon me to examine the views of two prominent writers who deny that there is anything there for a theory to apply to. Stephen Braude's argument proceeds from an apparent unity that underlies the collection of alters in a multiple. Ian Hacking argues that MPD cannot serve as support for any theory of the self because the phenomenon *as described* is therapy-induced rather than natural.

i. Braude and apperceptive unity

As I have mentioned above, the actions of the brain in weaving selves are characterizable from the intentional stance. In a similar vein, the way that alters emerge and retreat is contextually sensitive, and seems to proceed from a set of beliefs and desires about which alter is best suited to deal with a given situation; there seems to be some underlying manipulator of alters. I contend that Braude misinterprets the notion of the self being composed of a series of virtual captain-memes to be the claim that normally selves are unified - that is, purposefully organized and manipulated - groupings of entities every bit as encapsulated as alters in a multiple (1991a:164-191). Thus, he thinks that the distinction between normal and dissociated systems *on the meme theory* must be as follows: the sub-agents of a normal self are transcendently unified by an apperceptive conative force, while dissociated alters are not so unified. But, as I have said, there do indeed seem to be examples of such a unity across alters within a multiple. The fact that therapy is often able to integrate alters suggests some pre-existing unity, while the very fact that the right alter is produced to deal with the situations for which it is more or less specifically tailored strongly supports the idea that some deeper "observer" is making decisions about the best way to negotiate the world. Braude takes this to mean that Dennett is wrong to propose that alters could be selves, since the actual differences between pre- and postdissociative cases are not those that an intentional stance view, as Braude (mis)interprets it, predicts would obtain in a case of different selves. The conditions that Braude thinks Dennett assigns to cases of multiplicity are not met in cases of (what we call) multiplicity. This is because some aspect of unity, as Braude imagines it, is in fact demonstrated by many or most multiples.

But consider this analogous argument: The government of a nation is clearly a candidate for treatment from the intentional stance. It makes decisions about which internal "agents" will have some sort of hegemony over others, and in dealing with other nations. decides which internal "agents," or "citizens," are best suited to speak for the government given the situation. That is, sometimes a businessperson must speak to outsiders for the government, and sometimes an academic, sometimes an advocate, and so forth. Therefore, the "citizens" of a nation cannot be selves in their own right, since

their actions are partially explicable, and perhaps only explicable, from the point of view that treats the government as an agent.

Braude fails to recognize that one can find intentional systems almost anywhere, that we do in fact find them practically everywhere, and that they very rarely amount to anything remotely so rich or complex as human selves. That the switching between various alters seems governed by a desire to deal with the world in a certain fashion and by certain beliefs about how best to accomplish this feat, is probably a reasonable conclusion. But we may equally well ascribe intentional states to thermostats. While the unifying manipulator of alters is doubtless more complicated a system than a thermostat, there is no reason to think it closer to the complexity of a normal self than to the thermostat. This becomes clear if one imagines replacing the alters manipulated by the underlying apperceptive unity - alters possessed of rich memories, personality-dominated, language-driven - with some other merely identifiable features, like different shapes or colors. Even if the underlying switching mechanism were to switch which shape it presented to the world with the same alacrity by which it changed between deeply homuncular alters, I assume that nobody would want to suggest that the underlying unity was the unity of a *self*. And even if the unifying apperceptive force were a self (which, I repeat, it clearly is not) there is absolutely no *a priori* reason to conclude that the alters cannot be genuine selves as well.

Braude's mistake is only a more specific version of an error made by philosophers as distinguished as Hilary Putnam. Putnam too seems to think that the functionally organized constituents of a system that is a mind, cannot themselves be minds. We must specify this, Putnam argues, in order "to rule out such 'organisms' (if they can count as such) as swarms of bees as single pain-feelers" (1975:432). But what is the motivation for ruling this out *a priori*? It might be to block any sort of Searlian objection to functionalism based on the ludicrous nature of the *mere possibility* that a group of persons (say, all the people on Earth) standing talking amongst themselves, with the right sort of functional organization, could be a single mind. But why rule this out in principle when it has already been ruled out contingently? We have no reason to treat

the behavior of swarms of bees as though it were caused by a pain properly predicable of the swarm itself. But it just seems obvious that if we had such a reason, there would be nothing ludicrous about the ascription. Similarly, if we have independent reasons to ascribe selfhood to alters, the possibility that they might be unified within some other self cannot be a reason to renege on our initial ascription. Braude thus fails to show that the meme theory is wrong to describe some alters as selves.²⁴

ii. Hacking and evidence

Hacking, in a piece unsightly titled "Why Multiple Personality Disorder Shows Nothing about the Soul/Self/Mind/Person/Subject/Consciousness," distinguishes between two ways that a theory of the self might appeal to the phenomena of MPD.

I assert that [MPD]... does not furnish any evidence for any substantive philosophical thesis about the mind (or self, etc.). The phenomena may nonetheless illustrate some claim about the mind that is held for reasons quite independent of the phenomena. If so, would not the phenomena be

²⁴ Braude takes the Dennettian view of the self to claim that the way the self comes apart in MPD reveals its predissociative structure. Braude argues convincingly against the tenability of this *compositional reversibility*, writing that

[t]he self is like a pie in that... [w]e can slice a pie any number of ways... And even if we select a general method of slicing the pie...we still have to decide how large to make the pieces. The self, too, can be divided functionally in a vast number of ways (1991b:138).

Therefore, the way a self dissociates says nothing about how it was loosely divided predissociatively; there just is no non-arbitrary way to decide which of the (potentially infinite) functional divisions within the predissociative self is the relevant one.

But the relevance of this argument to Dennett's view (or the meme theory as I have described it) is doubtful. Dennett and I claim neither that the self has "a grain corresponding to situation-specific multiple personalities," nor that we may view alters as indicative of *the* functional or affective typological divisions within the predissociative self. My point would rather be that, if we imagine the collection of intentional states predicable of a body as a pie, the way we divide the pie up into selves is metaphysically ambivalent.

supporting evidence for the philosophical claim? No. I maintain that they furnish no evidence at all. They are only illustrations (unpublished:5).

I believe that the second sort of appeal to MPD - about illustration and, thus, application - actually deals with the possibility of MPD phenomena playing a roundabout evidentiary role. Thus Hacking's distinction is between two ways of recruiting MPD as evidence for a theory of the self. However, he argues only against the first way of potentially employing MPD in one's theory of the self, while the way I would employ MPD is precisely by using it to illustrate claims "held for reasons quite independent of the phenomena." But before engaging the tenuous difference between illustrating and evidencing, one should be clear on why Hacking rejects the idea that the phenomena of MPD can serve as supporting data for a theory of mind.

The claim is roughly that what count as MPD data are not solely, perhaps not even mainly, a function of the traumatic events proposed to cause dissociation. Hacking argues that the way patients describe themselves and their condition pretheoretically "has changed radically in the past two decades." More important, and more dramatic, has been the change in the way patients have been taught to remember what they did in childhood when they began implicitly to dissociate. The way alters refer to themselves, the way they engage in the politics of behavioral hegemony, and the way they describe their origins and relations are, Hacking argues, artifacts of the way they have been pressured by therapists to do so.

Dennett speaks of '*the terrible experiments that nature conducts.*'

What exactly are these experiments? It is not that nature produces for us adults who say the things that Dennett says they say... The experiment consists of a patient in therapy often lasting several years, coming to say the things she says. The experiment is so strongly controlled that if she does not say those things, she will commonly be released from therapy, for being too resistant, for denial.

The question is not whether children are abominably treated. The question is not whether they will grow up with grave psychological difficulties if their childhood is vile. The question is, is the subsequent stereotypical MPD behavior and therapy one of nature's experiments? Or is it rather the way in which a certain class of adults in North America will behave when treated by therapists using certain practices, and having certain convictions (unpublished:18)?

Hacking is raising the reasonable concern that, by the time a case becomes a legitimate diagnosed MPD case, the data are so infected by the therapist's conception of the mind and self that it is pointless and perhaps even irresponsible to accept the data as something that *nature* is revealing about the mind. But does the meme theory lose much, or anything at all, by giving up the claim that MPD is an experiment?

This depends upon what an illustration, in Hacking's sense, is allowed to be. Hacking refers to a passage from *Consciousness Explained* in which Dennett claims that if there was reason to think that a single person occupied two bodies (Fractional Personality Disorder, or FPD), the meme theory could accept that in stride as easily as it accepts MPD.

Now the power of FPD as an illustration in no way depends on whether the report of such [shared bodies] is true or false. Dennett's view of the person allows such a description to make sense... [but] Wittgenstein observes that if a picture of an experiment is just as compelling as the actual experiment, then the event is not functioning as an experiment at all... It is not as experiments that FPD or MPD teach anything for Dennett's philosophy of mind (unpublished:16).

But this clearly leaves the door open for MPD to teach us something about the meme theory in some fashion other than as an experiment; indeed, it allows that MPD

may render the meme theory "compelling." Hacking suggests that MPD cannot play an evidentiary role because it functions only as a picture rather than as something about the world requiring explanation. If this were true, then the most the meme theory could say about MPD is that *if there were such phenomena* the theory could account for them. Hacking's argument from the theory-infectedness of therapy, however, does not give us any reason to think that there is no real-world correlate of the "picture" of MPD. At most, it gives us reason to think that a different story about the *origin* of such phenomena is required.

Suppose, for example, that one grants (an uncertain claim) that treatment thoroughly infects the data that one subsequently garners from multiples. Indeed, suppose that therapy is what *creates* the appearance of multiple selves in one body, that before therapy there is no *x* such that the excising of *x* would constitute the murder of a morally significant agent. That the existence of post-therapy multiple selves is not a "natural" phenomenon would be an important fact and worthy of note. But surely there would still be a fundamental level of analysis at which, origins aside, we have multiple selves in one body and our theory of the self had better be able to account for this. First, our theory had better allow that such cases are coherent, and second, our theory will hopefully be able to offer some story about how these multiple selves came to be. Once again, the story about abuse and subsequent dissociation may not be the whole story; Hacking has argued that it is not. But once again, the story that *is* required can be told easily from the meme theory of the self - this time by dividing up the stages of dissociation between the trauma and therapy, and by incorporating some element of outside suggestion into the account of how alternative narratives are formed.

Nobody, not even Hacking, would suggest that one can create multiple selves just by telling any person that they are a multiple. The predisposition toward multiplicity conferred by childhood traumata is at least a crucial enabling element in the onset of MPD. But even if one could indeed create multiples by browbeating average persons into saying that they are multiples, the task would remain of explaining the conceivability and the psychology of multiple selves in one body. If we have a model of the mind,

formulated on considerations divorced from MPD, that performs this task, then we may conclude that *ceteris paribus* we are better served in regarding the mind as conforming to our model than in supposing it to be modelled in a fashion under which the phenomena would be miraculous. Thus the meme theory of the self can appeal to MPD for support, in that MPD comprises counterintuitive phenomena - whatever their origin - in need of explanation, and those phenomena are at least coherent under the meme theory.

iii. Meme theory and MPD

The account provided by Dennett for the origins of MPD proceeds from the observation that the overwhelming majority of multiples experienced severe abuse or some other variety of betrayal or trauma in childhood. The idea is that these traumatic events were so incompatible with the nascent self being woven by the brain that a new and temporary "Head of Mind" was launched in which the horrible events would have a place (Humphrey and Dennett 1991:147). This preserved the original, and predominate, self intact, but opened the door for the alternative self(ves) to aspire to functional saliency on occasion. This, argues Dennett, is what in fact happens when different alters move to dominate behavior. As criteria for evaluating whether there are indeed different selves in a single body, Humphrey and Dennett propose five guidelines. Genuinely different selves will be characterized by reciprocal opacity (or at least translucency) of thoughts and memory, by claims to consciously dominate behavior when present²⁵, by assertions of "integrity and personal importance," and by behavior such that alters are plausibly thought by others to possess such integrity and be "interesting different" (1991:153). Now, how may this account be developed in light of the meme theory I have synthesized?

The meme theory gives an account of the self according to which it is a narrative.

²⁵ Interestingly, Humphrey and Dennett choose to frame this criterion for multiplicity in terms of a self's willingness to assert *ownership* over the actions performed during its "rule." Cf. my comments above on this phenomenon as a criterion for *persistence* (Chapter 4).

It is a narrative in virtue of certain important relations between *q*-remembered experiences and present experiences, and in virtue of an interpretable compatibility obtaining between temporally distinct constellations of memes. In applying this meme theory to MPD, it is crucial to begin by recalling that the brain weaves these narratives.

When a person is said to have changed into a new person after a long span of time and/or some profound change, the meme theory explains this in terms of the current narrative being importantly discontinuous or incompatible with the former narrative. In recognizing this possibility, one allows that the brain may weave more than one narrative - usually over long periods. But given this, who is to say that the brain cannot weave different selves over shorter time spans?

I submit that the brain may, under special circumstances, have multiple works *in progress* at a given time. Perhaps these special circumstances will have only one narrative being written at a time, with the brain switching back and forth between stories; perhaps the brain writes two or more at the same time. My contention is that this is the way to describe cases of multiplicity from the meme theory of the self. The task, as I see it, is to explore what it means to say that multiple selves are in progress at one time.

One might suppose that an intentional stance view of the mind will have trouble explaining the shift from a single center of narrative gravity in a predissociative self to several such centers in a multiple. Specifically, a problem might arise from the intentional stance's assumption that rationality is a constitutive principle for the ascription of intentional states. On that assumption, if a system ever demonstrated anything greater than a relatively minuscule degree of practical irrationality, we would abandon the assumption that the system was an intentional system before we would advance the hypothesis that it was an intentional system choosing to defect from some Gricean convention of rationality. So when we are confronted with a system which, like the body of a multiple, warrants deep and widespread attributions of utterly contradictory beliefs and desires, why should we ever reach any conclusion other than that the system is too unreliable to be the subject of any intentional ascription? This conclusion is what Dennett's approach seems to dictate *prima facie*. So why aren't multiples just *mad*?

In other words, the meme theory owes an account of the escape from the Scylla of proposing one ideally rational intentional system and the Charybdis of concluding that one observes a single, altogether non-intentional system. A third option, to a first approximation, consists in the way that observers, and the brain of the heretofore single system, succeed in cutting a system's mental state pie in a such a fashion as to render up a multiplicity of compatible sets of beliefs and desires.

What will distinguish madness from multiplicity - what will make a multiplicity of compatible constellations an option - is the way that the *prima facie* inconsistent behavior of a genuine multiple may be subsumed under separate *encapsulated narratives*. What the alters say about themselves and their histories, the opacity of awareness across alters, and the complementary gaps in their respective memories are explainable in terms of the dynamics of virtual captains, past and present. All one need suppose is that the set of past virtual captains available for entry into the meme-struggle varies in some regular and recurring fashion. This means that the set of memories "from the inside" available to the ongoing succession of captains will vary correspondingly; the self that the virtual captain *remembers itself being* will change according to which set of memories is active.

Moreover, the sets of memories must be extensive and rich in order to comprise a self that a captain may recall being, and thus justify our proposing multiple centers of narrative gravity. To see this, imagine taking two stories and scrambling them together at the level of individual words. The resulting word-salad would hardly qualify as two intermingled narratives, which is to say, one could hardly explain the incoherence by saying that there are actually *two* stories. But if the stories were mixed by alternating a page of one with a page of the other, the continuity of each individual page would draw one's attention to the two encapsulated narratives running the length of the book. One might conclude that something out of the ordinary was going on with the structure of the book as a whole. But once a rule for understanding the book was established ("*read only every second page, then return to the beginning and read the remainder*"), the existence of two perfectly intelligible narratives in one book would be no more remarkable in this

form than if the two had been printed standardly, with first one appearing unbroken and the second following unbroken.

This, then, is how the meme theory makes sense of MPD. Alters are or are not separate selves on the basis of considerations like: the relations between the body's virtual captain and a set of memories from the inside; the richness of the constellation of memory to which the virtual captain has access at any given moment; and the distinctness of behavior by which one might feel impelled to ascribe multiplicity in the first place, *inter alia*. The meme theory permits one to claim that there is no metaphysical problem of MPD. Certainly the account given above the does not resolve the important questions about the recurring inability of sets of memories to surface, but these are psychological questions if ever there were. On the meme theory, the sort of mystery involved in questions about MPD is the same sort involved in questions about the formation of supernovae.

I conclude this section by way of mentioning some ideas which, though relevant, would require another thesis to develop adequately. One might suppose that the traumata of abuse somehow violate some constraints of what can and cannot happen in the nascent narrative of a child's self. That some horrible events simply cannot fit into a narrative, once begun, seems to me suggestive of some more or less hardwired imperative of trust for parents (or parent-figures) that the brain will go to great lengths to avoid breaking. A stronger claim, though equally interesting, would be that humans come equipped with basic templates for how a life should be, and that the betrayal and hurt of abuse is so incompatible with these templates that the results do not "make sense" to the victim's brain.²⁶ This latter account is supported by the fact that alters tend to be dominated by a single emotion; the claim, one supposes, would be that these selves do not have the option of passing on the undesirable aspect of their narrative, since it is their defining feature.

²⁶ The idea of this sort of template, though not necessarily applied in this way, finds a supporter in Brook (personal communication).

2 of/de 2

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

Conclusions

i. What has been done?

There are, I suppose, two ways to look at what has been presented above. I might be taken to be arguing in favor of the meme theory of the self by showing its positive connection both to Parfit's philosophy and to the phenomenon of MPD. Alternatively, I might be taken to be presenting a single approach to doing philosophy of mind and self - specifically, an interpretationist one - and to be defending it by showing the coherence between its different levels and facets, and by outlining its ability to accommodate various problems: misrepresentation, the evolution of consciousness, thought experiments of identity branching, and the possibility of multiple selves in a single body. I believe the latter is the most useful way to regard my project, but if anything is a matter of interpretation, this is. Take your pick.

Now, without wishing to claim that there is some deep fact in virtue of which this work is *one work*, I want to examine the important connections between the diverse themes of this essay.

My central claim about Parfit's theory of personal identity is that it cries out for synthesis with a theory of the self that will support his claims about what matters in persistence. In structuring this essay, I took my own advice. It is no accident that my thesis begins with a discussion of the nature, not merely of the mind, but of its constituents. Like Parfit's theory of identity, the meme theory of the self demands a grounding in a certain view of things as elementary to personal identity as propositional attitudes. Once the interpretationist view is pulled into the ring at these basic levels, its application to selves, and then to persistence, is much more coherent.

Of course, coherence is not cognate for palatability. There is no question of someone who already finds Parfit's theory overly interpretationist suddenly finding it plausible when it is shored up by a still more interpretationist view of the mental. Nevertheless, there is much to be said for coherence; it is at least a necessary condition

for plausibility. This is particularly true when the opposition to an interpretationist view argues from being *the only game in town*. The more topics and issues to which my theory demonstrates its relevance and its ability to resolve philosophical difficulties, the less compelling the claim that only a Realist approach can serve to explain things.

ii. The construction of the self

The self, it may seem, has vanished on the meme account. Constructed out of fictional memes, headed by an abstract captain occupying a logical captaincy, and bound together as a story by an illusion, the theory hardly seems to describe what we suppose ourselves to be. How, one might wonder, could a materialist theory take what is obviously real and portray it as fictitious?

But selves are not obviously real, or not as anything material; Parfit argues that there is no need to suppose they are, and Dennett argues that it is false to suppose that they are. Moreover, the project of materialism as I conceive of it is not to show that what *seems* real in the mind *is* physically real. Rather, it is to explain the *seemings* in terms of what is real physically. That, I submit, is precisely the meme theory's mandate. And in the end, one must give up very little when one adopts the pragmatic approach toward mental objects, the mind, the self and identity.

It is, after all, very much *as if* there were substantive beliefs, and selves. The intentional stance allows us all the *abstract* beliefs and desires we need to make sense of behavior; this has been Dennett's point all along. I consider myself to have argued in support of this point, and furthermore explored the important *seemings* of the self to show how the interpretationist meme theory deals with them. I conclude that even such phenomena as the synchronic unity of consciousness, the diachronic unity of the self, the capacity to "lose one's self" in an absorbing task, our ways of dealing with *fauxes pas*, and the etiology of Multiple Personality Disorder, *inter alia*, are explicable from the meme theory of the self. Indeed, not only are they explicable, they are philosophically resolved; the only task remaining is one for psychologists, or perhaps sociologists. The

value of the meme theory of the self is precisely its ability to demystify those aspects of the mind and self long thought ineffable and beyond the reach of scientific study.

Bibliography

- Block, N. "Review of *Consciousness Explained*" in *Journal of Philosophy*, Vol. XC, No. 4, April 1993. 181-193.
- Braude, S. - *First Person Plural*. London: Routledge. 1991.
 - "Multiple Personality & Structure..." in *Self and Identity*. Kolak and Martin, eds. Toronto: Collier Macmillan. 1991. 134-144
- Brook, J.A. *Kant and the Mind*. New York: Cambridge. 1994.
- Churchland, P. *Matter and Consciousness*. Cambridge: Bradford/MIT Press. 1986.
- Davidson, D. *Essays on Truth and Interpretation*. Oxford: Oxford University Press. 1984.
- Dawkins, R. - *The Selfish Gene*. Toronto: Oxford University Press. 1976.
 - *The Extended Phenotype*. Toronto: Oxford University Press. 1982.
- Dennett, D. - *Consciousness Explained*. Toronto: Little, Brown. 1991.
 - *The Intentional Stance*. Cambridge: Bradford/MIT Press. 1987.
 - "Back from the Drawing Board" in *Dennett and His Critics*. Dahlbom, ed. London: Routledge. 1993.

- Fodor, J. *Theory of Content*. Cambridge: Bradford/MIT Press. 1990.
- Fodor, J. and Lepore, E. *Holism, A Shopper's Guide*. Cambridge: Blackwell. 1992.
- Hacking, I. "Why Multiple Personality Shows Nothing about..." Unpublished.
- Hume, D. *Treatise on Human Nature*. Selby-Bigge, ed. Oxford: Clarendon. 1967.
- Humphrey, N. and Dennett, D. "Speaking for Ourselves" in *Self and Identity*. Kolak and Martin, eds. Toronto: Collier Macmillan. 1991.
- Kant, I. *Critique of Pure Reason*. Kemp Smith, trans. London: Macmillan. 1990.
- Kitcher, P. *Kant's Transcendental Psychology*. New York: Oxford. 1990.
- Nietzsche, F. *Ecce Homo*. Kaufmann, ed. New York: Vintage. 1967.
- Parfit, D. - *Reasons and Persons*. Toronto: Oxford. 1984.
 - "Personal Identity" in *Philosophy of Mind*. Glover, ed. Oxford: Oxford University Press. 1976.
- Putnam, H. "The Nature of Mental States" in *Mind, Language and Reality; Philosophical Papers*. Cambridge: Cambridge. 1975.
- Ryle, G. "Ordinary Language" in *Ordinary Language*. Chappell, ed. Toronto: Prentice-Hall. 1964.

Quine, W.V.O. *Word and Object*. Cambridge: MIT press. 1960.

Zalta, E. *Intensional Logic and the Metaphysics of Intentionality*.
Cambridge: Bradford/MIT Press. 1988.

END

'20-07-94

FIN