The Many Worlds Interpretation of QM: A Hylomorphic Critique and Alternative *

Robert C. Koons

1 Introduction

The so-called Many Worlds Interpretation of quantum mechanics has been extant now for nearly sixty years, beginning as H. Everett III's doctoral dissertation in 1957 [**?**], with further contributions by B. DeWitt and N. Graham in their 1973 book, *The Many Worlds Interpretation of Quantum Mechanics* [**?**]. The Everett approach takes quantum mechanics both realistically and as a stand-alone, autonomous theory of the world, not in need of a separate theory of measurement to bridge the apparent gap between the deterministic evolution of the wave-function in a highly abstract, probabilistic space and empirically observable statistics in the laboratory. Instead, Everett proposed that all of the apparently contradictory macroscopic results assigned some finite probability by the theory are equally real, coexisting in distinct sets of *relative states*. DeWitt and others later identified these clusters of mutually consistent relative states with distinct and co-existing *worlds* or *branches* of the world.

These early versions of the interpretation faced a huge problem: there were no *worlds* or *branches*, describable in macroscopic terms, to be found in the formalism of quantum mechanics itself. We can find within the formalism something called *superpositions*, which are states that seem to attribute to particular systems (like particles) a plurality of mutually inconsistent properties, each with a certain amplitude, but there

*I would like to acknowledge the support during the 2014-15 academic year of the James Madison Program in American Ideals and Institutions at Princeton University (for a Visiting Fellowship) and the University of Texas at Austin (for a faculty research grant).

seems to be no way to recover macroscopic instruments and determinate measurement relations from these isolated superpositions. This problem is often described as the problem of finding a *preferred basis*, since the decomposition of the world into discrete branches can only take place relative to a selection of a certain set of orthogonal parameters. Any selection of such a basis seemed arbitrary and unprincipled, and so the objectivity of the co-existing branches was thrown into doubt. In addition, there is nothing in the wavefunction that corresponds to the *persistence* or *splitting* of branches. Probabilities of various states simply fluctuate over time: there is no way to *trace* where the probability that once belonged to a given state has moved (either as a unified packet or through fission).

In the 1970's, 80's, and 90's, a great deal of theoretical work commenced on the problem of giving a fully quantum-theoretic account of measurement. This work comprises the programs of *decoherence* of W. Zurek [**?**] and H. D. Zeh [**?**] and the *consistent histories* approach of R. Griffiths [**?**], R. Omnès [**?**], and Gell-Mann and Hartle [**?**, **?**]. The decoherence results show that under favorable circumstances a stable, approximately classical domain can be expected to emerge from the quantum-mechanical descriptions of a measuring system, its object, and the surrounding environment.citeWallace11 What decoherence left unsolved was why we see the emergence of just *one* such quasi-classical domain when interacting with quantum superpositions. A marriage of decoherence with the Everett interpretation was inevitable, with the Everett interpretation explaining the apparent uniqueness of result as a product of the relativity of our perspective in this or that branch, and the decoherence providing the missing preferred basis and explaining how to extract persistent and apparently "splitting" quasiclassical domains from quantum descriptions.

The consistent histories approach was even closer to the spirit of the Everett interpretation, since it sought to extract approximately classical domains from the quantum function for the entire cosmos, rather than looking at particular instrument-object-environment arrangements. Here again, the two approaches seemed designed to resolve each other's deficiencies: with consistent histories providing the preferred basis, and the Everett interpretation dissolving the worry about what to do about certain regions or phases of the cosmic history in which there are no consistent histories at all. On the Everett interpretation, only the quantum wavefunction describes fundamental reality, so only it can be expected to have universal validity. Consistent histories simply describe the approximate emergence of quasiclassical branches under favorable circumstances, including, presumably our own.

In recent years, the Many Worlds Interpretation has found a new home in Oxford, among both physicists and philosophers of science, including David Deutsch, Simon

Saunders, David Wallace, Christopher Timpson, and Harvey Brown. The Oxford group has developed the idea of using decoherence and consistent histories approaches to solve the preferred basis problem, explaining the emergence of approximately classical "domains" from the wavefunction. They have also, building on seminal work by Deutsch [**?**], attempted to solve the other central problem of the interpretation, which is that of making sense of the precise probabilities ascribed to different outcomes by applying Born's rule to the wavefunction.

In the next two sections, I will raise two objections to the new, Oxford-style Everettian interpretation. First, in section 2, I will argue that Deutsch's strategy cannot make sense of the probabilities that play such a central role in quantum mechanics. The many-worlds interpretation cannot explain the rational necessity of one of the crucial axioms (Savage's Sure Thing Principle) upon which the modern theory of subjective probability depends. Then, in a much longer section 3, I will argue that the Oxford Everettians attempt to use the philosophical framework of *functionalism* to elucidate the relation between the manifest world of scientific experiment and observation and the underlying, fundamental quantum reality ends in failure. Specifically, I will idenify four failures of this account:

1. I will use Putnam's paradox to demonstrate a radical indeterminacy of content that would afflict all of our scientific theories.
2. I will demonstrate that any consistent story of the world (no matter how fantastic) would count as equally *real*.
3. As a consequence, it would be impossible for any of our scientific theories to be wrong, making it equally impossible for them to be empirically confirmed.
4. This failure of empirical testability would deprive us of any reason for believing in quantum mechanics in the first place.

In section 4, I will critically examine seven possible strategies that Oxford Everettians might rely on to solve the Putnamesque problems identified in section 3. These strategies are (1) appealing to the concept of *emergence*, (2) appealing to "our" actual language and theories, (3) relying on causal constraints to fix the interpretation, (4) appealing to *natural* or *eligible* properties, (5) using realism about spacetime, (6) appealing to simplicity, and (7) relying upon decoherence to solve the problem. I will argue that all seven strategies fail, although there is a version of the appeal to simplicity that might turn out to provide at least a partial solution. However, even this appeal to simplicity (if it were ultimately successful) would leave us without an adequate account of the *reality* of the manifest, macroscopic world, and it would still

have the consequence that none of our theories in the special science (theories of the emergent, macroscopic world) can ever be false, with all of the epistemological catastrophe that such a result would bring

I turn in section 5 to sketching a neo-Aristotelian alternative to the Oxford Everettian interpretation. This new interpretation adds the additional metaphysical constraints needed to solve the Putnamesque paradoxes in the form of a set of *essences* of macroscopic substances. This also enables us to use the actualization of these essences as a way of distinguishing the one *actual* branch from all the *merely possible* ones. Alex Pruss and I call the resulting interpretation *the traveling branches* interpretation. This interpretation builds on both the realism about the quantum wavefunction and the results of decoherence theory, in exactly the same way as these are treated by the Oxford Everettians, and yet it ends up in a metaphysical and semantic position that is much more defensible.

2 Probability and the Oxford-Style Everett Interpretation

The basic problem can be stated quite simply: since all possible outcomes will in fact occur with probability *one*, what meaning can be assigned to the varying strengths of probability assigned by Born's rule to different outcomes? On the Oxford interpretation, it makes no sense to count branches, since the very existence of branches is only a non-fundamental and inherently vague phenomenon, resisting any perfect precisification. Deutsch's answer, developed further by Saunders and Wallace [**?**, **?**, **?**], is to use the pragmatic approach to subjective probabilities developed in the early 20th century by Frank Ramsey, Leonard Savage, John von Neumann, and others in order to argue that perfectly rational agents must, given certain constraints including perfect knowledge of the quantum state, *act as if* they assigned the appropriate probabilities to the various branches.

In a paper entitled "Truth and Probability," [**?**] Frank P. Ramsey sought to provide an operational or behavioral definition of the notion of *degree of belief* or *degree of confidence of truth*, as well as the correlative notion of *desirability* or *utility* or *subjective value*. Ramsey imagined an idealized experimental set-up in which both the degrees of belief in various propositions of the experimental subject (i.e., the subject's *subjective probability function*) and the degrees of desirability that the subject attaches to the states of affairs represented by those propositions (i.e., the subject's *utility*

function) may be measured. Ramsey imagines that the subject is confronted by what he (the subject) believes to be an omnipotent and totally trustworthy Bookie, who offers the subject a series of choices between two options. Some of these options come in the form of simple bets: e.g., an option that might be offered to the subject could take the form: α if p is true; otherwise, β. If the subject's choices conform to certain principles of mutual coherency, then there exists a unique representation of the subject's state of mind in terms of a probability function taking as its values real numbers in the interval from 0 to 1 (inclusive) and a utility function taking real numbers as values.[1]

In 1954, Leonard Savage [**?**] provided a slightly different axiomatization, from which he was able to prove a representation theorem of the appropriate kind. Savage's set-up included the following elements:

1. A set of *states* of the world, S, with elements $s, s', s'', \ldots$ and subsets (the *events*) $E, E', E'', \ldots$.
2. A set of consequences or outcomes C, with elements $c, c\prime, c\prime\prime, \ldots$.
3. A set of acts $\mathcal{A}$, with elements $A, A', A'', \ldots$.
4. an assignment of a consequence from C to every act-state pair (A, s), designated $A(s)$.
5. A binary relation $\succeq$ between pairs of acts that is interpreted to mean *is preferred or equal to*.

His axiom system included the following axioms:

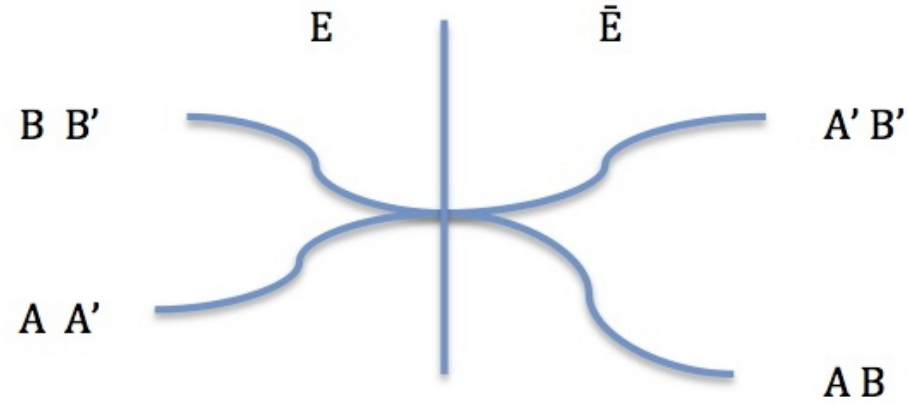
Axiom 1 *The relation $\succeq$ is a weak ordering of the acts: the relation is transitive, and any two acts are comparable (either $A \succeq A'$ or $A' \succeq A$, or both).*

Definition 1 *$A \succeq_E A'$ if and only if: if acts A and A' are modified so that their consequences are the same for every state not included in E, and they are not modified for any states in E, the resulting modification of A is preferred or equal to the modification of A'. The propriety of this definition depends on the next axiom, Savage's Sure Thing Principle.*

Axiom 2 The Sure Thing Principle *If acts A, A', B, B' and event E are such that (i) A and B agree on all states outside of E, (ii) A' and B' also agree on all*

[1]Strictly speaking, the function isn't unique, but any two acceptable functions will be such that each is a linear transformation of the other.

states outside of E, (iii) A and A' agree on all states within E, (iv) B and B' agree on all states within E, then: $B \succeq A \leftrightarrow B' \succeq A'$.



Savage's Sure Thing Principle asserts the irrelevancy of non-discriminating possibilities. That is, suppose that two actions A and B are known to have exactly the same consequences (as far as things are concerned that matter to the agent) on the assumption of E: then, if B is preferred to A, then B would still be preferred to A regardless of what those irrelevant consequences would be (regardless of what the common consequences are that would follow from either A or B on the supposition of E). The Sure Thing Principle (in conjunction with the other axioms) has a further consequence: the agent will also prefer B to A regardless of whether the agent knows E to be true or false or is left in some state of ignorance about E. The probability of E, given its irrelevance to the comparison of A and B, must be irrelevant to the preferability of B over A.

On the basis of his axioms, Savage was able to prove the following representation theorem, demonstrating that any reasonable agent must act as if guided by both a utility and probability function (with maximizing expected utility as the decision criterion):

Theorem 1 Representation Theorem. *There exists a unique real-valued function P defined for the set of events, and a unique (up to positive linear transformations) real-valued function u defined over the set of consequences such that:*

1. $P(E) \geq 0$ *for all* E.

2. *$P(S) = 1$.*

3. *If E and E' are disjoint, then $P(E \cup E') = P(E) + P(E')$.*

4. *E is not more probable than E' if and only if $P(E) \leq P(E')$.*

5. *If the E_i's are a finite partition of S, and A is an act with consequence c_i on E_i, and if the E'_i's are another finite partition on S, and A' is an act with consequence c'_i on E'_i, then $A \succeq A'$ if and only if:*

$$\sum_{i=1}^{n} u(c_i)P(E_i) \geq \sum_{i=1}^{m} u(c'_i)P(E'_i)$$

2.1 The Failure of Savage's Principle in a Many-Worlds Setting

Deutsch [**?**] and Wallace [**?**] argue that the formal results of Ramsey and Savage can enable the Everettians to make sense of the varying probabilities of the various branches of the wave function. If we assume that our ideally rational agent must satisfy Savage's axioms (with states now identified with branches) and must also respect certain physically symmetries, then we can prove a representation theorem similar to Savage's, but with the added feature that the probabilities assigned to the branches must mimic the probabilities assigned to those branches by Born's rule. In particular, Wallace incorporates Savage's Sure Thing principle into his condition of *diachronic consistency*.

I share the doubts about the cogency of the argument expressed by Huw Price [**?**]. Price points out that by adding multiple worlds to our inventory of reality, we give rational agents *new things to care about*, things that can't be captured by any probability-weighted averaging of world-bound utilities. For example, a rational agent might assign a certain finite amount of utility to the degree of *equality of outcome* enjoyed or suffered by persons across the span of branches. Wallace [**?**, pp. 256-7] argues that we can capture such cross-world utility considerations by tinkering with the various world-bound measures, but it is easy to prove that giving a positive weight to transworld equality will necessarily violate Wallace's diachronic consistency condition (by violating Savage's Sure Thing Principle), despite the obvious rationality of that principle in a one-world setting.

A core principle of Savage's axiomatization of decision theory is his *Sure Thing*

Principle, which is incorporated into Wallace's axiom system (in the form of his *diachronic consistency condition*). As we have seen, the Sure Thing Principle asserts the irrelevancy of non-discriminating possibilities. That is, suppose that two actions A and B are known to have exactly the same consequences (as far as things are concerned that matter to the agent) on the assumption of E: if B is then preferred to A, then B would still be preferred to A regardless of whether the agent knows E to be true or false. This makes sense in a normal, one-world decision setting: if I know that the world will end up either in E or in E' (exclusively), and I am indifferent between A and B on the assumption of E (because I know that A and B would produce exactly the same results in that case), then my preference for A over B must be concerned exclusively with what would happen on the assumption of E'.

However, in a many-worlds setting, I have to imagine that both E and E' will become actual (albeit in different branches), and I might care about certain trans-branch facts. We can no longer assume that my utility function is simply a weighted sum of intra-branch values. For example, suppose that there are two people affected by my action, person 1 and person 2. In the E' branches, person 1 and person 2 will each receive zero units of value, regardless of whether I perform A or B. However, in the E branches, persons 1 and 2 get one unit of value if I do A, but if I do B instead, then person 1 gets 0.4 units of value and person 2 gets 2 units. Now suppose that I value outcomes on the basis of the total utility (across all branches) minus the average deviation from the mean (representing my dislike for interpersonal inequality). My desire to avoid overall inequality applies equally well to inter-branch inequality as to intra-branch inequality. Here is the resulting table:

	E	E'	Expected value
A	$\langle 1, 1\rangle$	$\langle 0, 0\rangle$	$2 - 0.5 = 1.5$
B	$\langle 0.4, 2\rangle$	$\langle 0, 0\rangle$	$4.4 - 2.9/4 = 1.7$

Note that if we eliminate E' from consideration, then I would definitely prefer action A, since it provides perfect equality and that outweighs in this case the somewhat greater total utility of action B. In contrast, however, if we take branch E' into account (as a genuine part of reality), the total value of action B now outweighs its somewhat lower average deviation from the mean, and B is now preferable to A. This is all perfectly reasonable in a many-worlds setting, in which all of the outcomes are equally real (enjoyed or suffered in reality in one branch or the other), but it clearly violates Savage's Sure Thing Principle (and, consequently, also Wallace's even stronger diachronic consistency condition). No amount of tinkering with utility

functions can overcome this difficulty.

We can see that, in the Many Worlds setting, we can turn this example into a straightforward violation of Savage's Sure Thing Principle (and thereby, also, a violation of Wallace's diachronic consistency axiom). Let's add two new actions A' and B' to the setting. A agrees perfectly with A' on condition E, and B agrees with B' on that same condition. The two actions A' and B' produce exactly the same result under condition E'. Therefore, the Sure Thing Principle requires that the agent prefer A to B just in case she prefers A' to B'. However, considerations of equality that are germane in the Many Worlds setting lead the agent quite reasonably to prefer B to A but also to prefer A' to B'. The Savage-style justification of the Born rule collapses.

	E	E'	Expected value
A	$\langle 1, 1\rangle$	$\langle 0, 0\rangle$	$2 - 0.5 = 1.5$
A'	$\langle 1, 1\rangle$	$\langle 1, 1\rangle$	4
B	$\langle 0.4, 2\rangle$	$\langle 0, 0\rangle$	$4.4 - 2.9/4 = 1.7$
B'	$\langle 0.4, 2\rangle$	$\langle 1, 1\rangle$	$4.4 - 1.8/4 = 3.95$

Why doesn't this possibility of violating the Sure Thing Principle apply with equal force to those who reject the Many Worlds interpretation? Because it is irrational to let what might have happened but did not actually happen influence one's evaluation of a particular possible outcome. We could express this idea by means of a slogan: *Value Supervenes on Being*. The value of an outcome is a function of what does or does not happen if that outcome were actual. It cannot depend on what happens in other, mutually incompatible outcomes. In the many-worlds interpretation, all of the branches are equally real, equally partaking of existence. Hence, Deutsch and Wallace have no grounds for excluding trans-branch values and so no way to validate Savage's axioms.

Ironically, the technical results of Deutsch and Wallace do provide strong grounds for accepting a One World interpretation of quantum mechanics. Since any One World interpretation will satisfy Savage's axioms, the symmetry considerations cited by Deutsch and Wallace make it reasonable to suppose that any rational agent in a one-world setting must attribute objective chances to events in a way that corresponds to the Born rule, given certain knowledge of the actual quantum function. Thus, Deutsch and Wallace's hard work was not wasted: they simply deployed it on behalf of the wrong interpretation!

2.2 A Second Failure: Retrospective Probability and Anti-Darwinian Branches

There is a second critical gap in the Deutsch-Saunders-Wallace program: namely, its inability to justify purely *retrospective* uses of probability, including the grounds for our conviction that we do not inhabit any of the low-probability anti-Darwinian branches (branches in which highly maladapted populations have managed to survive despite that maladaptiveness). An anti-Darwinian branch is one in which the typical populations of organisms are *very* badly adapted to their environment, a world in which Empedocles' half-ox-half-man and other unfit monstrosities predominate. There is good reason to think both (i) that there are very many such anti-Darwinian branches lurking in the world's quantum-wave function, under some suitable interpretation function, and (ii) that all such anti-Darwinian branches have extremely low quantum amplitudes.

One-world theorists can hold that all branches with extremely low amplitude are *approximately impossible*—that is, close enough to being impossible as to be practically negligible. However, such a notion of "approximately impossible" is unavailable to many-world theorists. Each branch is, for a many-worlds theorist, just as real as any other. They are all *equally far* from being impossible, since all actual events are equally possible. It makes no sense to say that some are so close to being impossible as to be practically indistinguishable from it.

Many-worlds theorists (Deutsch, Saunders, Wallace) have focused their attention on two issues concerning probability: (i) arguing that it could be a constraint on rational action that the rational weight expected utilities of future branches proportionally to the square of the branches' amplitudes, and (ii) arguing that it could be rational for an investigator to treat sequences of observations that match those expected on relatively high-amplitude branches to count as confirming a statistical theory. Both of these issues concern agent-bounded uses of probability, in the sense that it is only the probabilities of events actually observed by the agent, either in his or her past or expected in the future, that are relevant.

However, when we use natural-selection arguments to exclude the possibility that we and other extant organisms are terribly maladapted to our environment, we are using probabilities in a way that extends far beyond our own past or present experiences. The arguments about decision theory and about statistical inference are irrelevant. We would be faced, on the Many-Worlds interpretation, with a problem of pure self-location, and there would seem to be no grounds for rationally assuming that we

must be *ab initio* located in one sort of branch rather than other. I cannot say that it is virtually impossible for me to be located in a low-probability branch when there are in fact many counterparts of mine who are in fact located there. For example, Richard Dawkins has often stated, quite reasonably, that he is certain a priori that natural selection will have shaped the evolution of living organisms on any planet in the cosmos, no matter how remote. To have any validity, such a priori appeals to natural selection must embrace the whole of reality, or none of it. The alternative branches of the Many-Worlds interpretation are simply remote regions of reality and yet there is no doubt that such a priori confidence in the validity of natural selection in all such branches would be profoundly misplaced. But charity begins at home: if we cannot apply the principle of selection to remote branches a priori, we would be unjustified to do so for our own world.

What cost do we pay if we forego such a priori appeals to natural selection? A very profound cost to the foundations of epistemology. Our natural reliance on our senses and memories is undermined if we cannot count on natural selection to ensure their reliability. Such undercutting of confidence in empirical knowledge will, in turn, deprive scientific theory (including the theory of quantum mechanics) of objective warrant.

3 Recovering the Manifest Image through Ramseyfication

Leaving probability aside for the moment, the central problem for the Many Worlds interpretation is that of bridging the gap between the scientific image of the quantum wavefunction and the "manifest image" (to use Sellars's phrase [**?**]) of our approximately classical, macroscopic world, the world occupied by all experimenters, their instruments, and the results of their experiments. David Wallace's solution is an admirably simple one: all features and entities of the "manifest image" (macroscopic objects, organisms, sensible properties) are to be reduced to functional roles realized in some way by the quantum wavefunction—where these functional roles can either be identified with second-order, functional properties, or with the quantum-mechanical role-fillers corresponding to those properties (in style of David Lewis [**?**, **?**, **?**]).

The functional properties can be identified with the result of Ramseyfying ([**?**]) our ordinary folk ontology and our special sciences (including, perhaps, classical mechanics) in the language of pure quantum-mechanics (infinite Hilbert space, unitary

Schrödinger evolution). There are three historical precursors to the kind of functionalization of the manifest image that Wallace has in mind: the phenomenalistic project, as typified by John Stuart Mill and the early Carnap, Bertrand Russell's functional-structural account of physics in *The Analysis of Matter* [**?**], and the late-twentieth-century, behaviorism-inspired accounts of the mind, especially the Analytic Functionalism of David Lewis [**?**, **?**, **?**]. The ideal formal machinery for each attempted functional reduction is F. P. Ramsey's account of scientific theories [**?**], well explained by Lewis in [**?**].

In both the phenomenalist project and in Russell's 1927 structuralism, there was a significant epistemological dimension, based on the idea that we have privileged and certain access only to our own conscious states (a kind of Cartesian starting point for knowledge). That epistemological element is much reduced in Lewis's functionalism, and entirely absent from Wallace's project, so I will, at the risk of some anachronism, present all four programs as if they were concerned solely with ontological issues, that is, with identifying the correct truthmakers or truth grounds for the reduced theory.

In this paper, I will use a model-theoretic version of Ramseyfication, in which, instead of introducing second-order variable and quantifiers for the predicates, we simply extend the interpretation function of a given model in order to turn a model of the original, base language into a model of the emergent theory in an appropriately expanded language.

We must also make use of a set of *possible worlds*, because all versions of functionalism require that we make some reference to the *dispositions* of things to respond or behavior in specified ways, even if the things never actualize these dispositions. The simplest formal semantics for such dispositions makes use of the subjunctive conditional: if P *were* true, Q *would be* true. We can represent the truth of such a subjunctive conditional at the actual world $w*$ by supposing that w is surrounded by a system of spheres of worlds, representing degrees of closeness or similarity of those worlds to w. The subjunctive conditional $(P \mathrel{\square\!\!\rightarrow} Q)$ is true at $w*$ if the material conditional $(\neg P \vee Q)$ is true in all of the worlds contained by some P-permitting sphere.

For the sake of simplicity, I will assume that each of the individuals exists in only one world. Our interpretation function must assign to each constant (proper name) an individual in each world, and to each n-ary predicate, a set of n-tuples of individuals from that world. This will enable us to assign truth-values at each world to all logically complex formulas, using the usual clauses of Tarski's truth definition. That

is, a negation $\neg\phi$ is true in a world w just in case ϕ is not true there, and a disjunction $\phi \vee \psi$ is true in w just in case either ϕ or ψ is true there. We can interpret the existential and universal quantifiers in the usual way, using at each world the domain of individuals that exists there.

1. A model frame F consists of a set of worlds W, a designated actual world $w^* \in W$ and, as in David Lewis's semantics [**?**], a system of spheres S, consisting of nested subsets of W, centered on $w*$. I will assume that the sphere-system S is dense (between any two concentric spheres there is always a third), and that the number of worlds in every sphere-membership equivalence class is equal to the number of sets of atomic formulas of the language.

2. A model M consists of a model frame F plus a domain of "worldbound" individuals D (each existing in just one world) and an interpretation function I, which is used to interpret the predicates, function symbols, and simple singular terms (names or constants) of the language.

3. The set of individuals D is partitioned into disjoint cells, one for each world in W. We can think of D as a function from W into a set of disjoint sets, with $D(w)$ designating the worldbound individuals of world w.

4. For any n-ary predicate F, $I(|F|)$ is a function whose domain is W, and for each world $w \in W$, $I(|F|)(w)$ is a set of n-tuples of the members of $D(w)$.

5. For any constant c, $I(|c|)$ is a function whose domain is W, and for each world $w \in W$, $I(|c|)(w)$ is a member of $D(w)$.

6. For any atomic sentence $F(c_1, c_2, ..., c_n)$, $I(|F(c_1, c_2, ..., c_n)|)$ is a set of worlds in W, where each world w belongs to $I(|F(c_1, c_2, ..., c_n)|)$ if and only if the n-tuple $\langle I(c_1)(w), I(c_2)(w), ..., I(c_n)(w)\rangle$ belongs to $I(|F|)(w)$.

7. $I(|(\phi \,\&\, \psi)|) = I(|\phi|) \cap I(|\psi|)$, and similarly for the other sentential connectives.

8. $I(|\exists x \phi(x)|)$ = the infinite union of the sets $I(|\phi(c)|)$, for each constant c in the language L. (We'll assume that the language L has been enriched with enough constants to provide a witness for every existential generalization true in M.)

9. $I(|(\phi \,\Box\!\!\rightarrow \psi)|) = W$ if there is an $I(|\phi|)$-permitting sphere s in S such that every world in $I(|\phi|) \cap s$ is also in $I(|\psi|)$). Otherwise $I(|(\phi \,\Box\!\!\rightarrow \psi)|) = \emptyset$.

A model $M = \langle F, D, I\rangle$ is a model of a theory T just in case, relative to I and D, the actual world w^* belongs to $I(|T|)$, where $I(T)$ is the intersection of the sets $I(|\phi|)$,

for each formula ϕ in T. As usual, a theory is defined as a set of formulas closed under logical implication.

Let's suppose that we start with a model $M_{base} = \langle F, D, I \rangle$, defined for our base language L_{base}, which represents the fundamental level of reality. Now suppose that we extend the language L_{base} to a language $L_{base+emergent}$, by adding constants, function symbols, and predicates that signify an emergent, non-fundamental level of reality. A theory $T_{emergent}$ of this emergent world is realized in our base model M just in case the interpretation function I can be extended to a new function $I_{realizer}$, defined for $L_{base+emergent}$, such that the model $M_{extended} = \langle F, D, I_{realizer} \rangle$ is a model of $T_{emergent}$. In such a case, we can say that the function $I_{realizer}$ is a *realization* of the emergent theory $T_{emergent}$ in the original base model M_{base}. This model-theoretic version is a generalization of Ramsey's original idea, since it applies even to theories that are not finitely axiomatizable. Instead of taking a single formula that axiomatizes the emergent theory and replacing all the emergent terms and predicates with first- and second-order variables, we extend the interpretation function of the original model in order to provide extensions to all the terms and predicates of the emergent theory. In cases in which a theory can be axiomatized by a single formula, the two methods are exactly equivalent: the base model will verify the second-order Ramsey formula if and only if the model's interpretation function can be extended to produce a model of the corresponding theory.

3.1 Classical phenomenalism and Russell's structuralism

Using this model-theoretic approach to realization, classical phenomenalism could be seen as postulating that all truths about the existence and characteristics of physical objects are realized by truths about the private and subjective sense-experience that human observers have or would have under specified, counterfactual conditionals. As Mill put it, physical objects are "mere permanent possibilities of perception." So, we start with a base language L_{phen}, which includes terms for subjects of experience, terms for sense-data, predicates that define sense experiences in terms of the locations of sense data in the egocentric spaces of subjects (with properties like *up* and *down*, *left* and *right*, *forward* and *back*) at times in private, egocentric time lines. The language will also include the subjunctive conditional. We will then consider a class of models for this language, consisting of a set of worlds W, a designated actual world w^*, an interpretation function I for evaluating atomic sentences in each world, and a system of concentric spheres S for the interpretation of subjunctive conditionals. For simplicity's sake, I will treat all sense-data and subjects as worldbound individuals

(in Lewis's sense). We can then select the model $M_{true-phen}$ that incorporates all the actual truths about actual and counterfactual experiences. The set of formulas true in $M_{true-phen}$ is the set $TRUE_{phen}$, the set of all truths expressible in the vocabulary of L_{phen}.

Throughout this paper I'm going to assume that the structure of the true model of fundamental reality is rich enough that there is a homomorphism from any canonical model for Lewis's subjunctive conditionals into that model, which in the case of phenomenalism we'll call $M_{true-phen}$.[2] If this were not the case, we would have little reason to believe that any of our theories of the emergent world have even approximate models in the true model of fundamental reality. Furthermore, we have good reason to think that the model of fundamental reality is very rich representationally, with a very large number of worlds, with a very rich set of relations of comparative similarity. There is every reason to think that the canonical model for any language using the subjunctive conditional can be mapped into such a rich model.

We now enrich the language by adding terms referring to physical objects, which will now be assigned locations and trajectories in a single three-dimensional (public) space, indexed by universal time. Since we still retain the subjunctive conditionals, we can now express conditional relationships between sentences expressed in purely phenomenal terms and sentences expressed in purely physical terms, and between pairs of sentences both of which are purely physical in form, as well as between sentences that mix both vocabularies.

- Call the resulting language $L_{phen+phys}$.
- Consider each theory expressible in $L_{phen+phys}$ that is consistent with the set of phenomenal truths, $TRUE_{phen}$.

[2]Let A be a set of formulas that is logically consistent (in Lewis's conditional logic). Extend A to a maximum consistent set C. Given our assumptions about the model (namely, that the system of spheres is dense, with a sufficient number of worlds in each sphere-membership equivalence class), we can find an interpretation function I that verifies all of C (and, therefore, also A) in M_{base}. We can use C to impose a partial ordering on the formulas of the language: $\phi \leq \psi$ iff '$((\phi \vee \psi)\Box\!\!\rightarrow \phi)$' $\in C$. We can then use the Axiom of Choice to find a 1-to-1 function from this ordering into the system of spheres S in such a way that the smallest sphere containing a ψ-world contains the smallest sphere containing a ϕ-world iff $\phi \leq \psi$. Now take every set G of atomic formulas such that for every χ consisting of a conjunction of members of G and of negations of atomic non-members of G, the formula '$\neg(\phi\Box\!\!\rightarrow \neg\chi)$' belongs to C. Select a world from the designated set of closest ϕ worlds and place it in the extension of exactly the members of G. By repeating this for every formula χ and every set of atomic formulas G, we will build the appropriate interpretation function I.

- Let T_0 be one such a theory.
- Since T_0 is consistent with the set of phenomenal truths, we can extend the interpretation function I_{phen} to a function $I_{phen+phys}$ in such a way that theory T_0 is true in the model $M_{true-phen}$ relative to $I_{phen+phys}$.

The extended interpretation accomplishes exactly the same thing as would be accomplished by Ramseyfying the physical vocabulary in a finite axiomatization of T_0, if there is a such a thing. That is, $I_{phen+phys}$ assigns some property-intension or individual-concept-intension in $M_{true-phen}$ to every predicate and individual constant in the physical vocabulary of T_0 in such a way as to verify T_0. The model-theoretic approach that I've sketched is actually more general than Ramseyfication, since it will apply to any consistent theory, whether or not that theory can be finitely axiomatized. In addition, it means that we can keep everything in first-order logic. Each interpretation function I relative to which T_0 is true in $M_{true-phen}$ constitutes a distinct *realization* of T_0.

Russell's structuralist program in *The Analysis of Matter* is exactly isomorphic to the classical phenomenalist program. The only difference is that Russell does not use subjunctive conditionals, as Mill did, but instead speaks of *causal* relations, both in the phenomenal and in the physical world. However, he does not offer a substantive account in the 1927 book of what causation consists in, so this is a difference we can, at least for the moment, set aside. In addition, of course, instead of speaking about phenomenal sense-data, Russell in 1927 speaks instead about *perceptions*, which he takes to be events in the brain with which we are immediately acquainted.

In general, there will be many realizations of any theory T_0 in the model $M_{true-phen}$, and there will be many other theories in the enriched language besides T_0 that are consistent with the set of all phenomenal truths (and which therefore have realizations in $M_{true-phen}$). In order to cut down the number of theories and realizations, we need some further constraints both on our theory T_0 and on the permissible realizations of that theory. We can accomplish both of these at once simply by restricting the interpretation function. We can then hope to pick out the one true theory of physics that has a unique permissible realization in the model $M_{true-phen}$.

In the case of both the phenomenalist and Russellian-structuralist program, these constraints consist in the *laws of perspective* that link geometrical properties described in terms of public four-dimensional spacetime with properties described in terms of egocentric phenomenal space and time. We can put a constraint an any acceptable interpretation function, requiring that when it identifies a physical object in a world with a set of sense data associated with subjects in that world, the

interpretation function must assign a shape and size to the physical object that corresponds to the shape and size of each of the corresponding sense-data, with the correspondence relation fixed by the laws of perspective as applied to the physical location assigned to the relevant subject of experience. That is, the physical primary qualities of bodies must correspond to sense-data and subject-locations in such a way that each sense-datum accurately records the shape of the body, as it would appear to a subject at the location to which the subject is assigned.

This is quite a severe constraint—in fact, too severe, since it fails to take into account the existence of illusions and hallucinations. It is reasonable to suppose that only one theory-interpretation pair will maximize the degree of fit between the bodies and the corresponding sense-data, and we can take this pair to give us both the set of truths about the physical world and the corresponding truthmaker in the phenomenal world for each truth.

3.2 Analytic Functionalism about the Mind

David Lewis's version of Analytical Functionalism is exactly isomorphic to the phenomenalist or structuralist model sketched in the preceding subsection. The differences are these: first, the base model with which we begin is a model of something like classical physics and chemistry, including facts about overt behavior, sensory-organ stimulations, and neural structures and patterns of firing. The true model of the world $M_{true-phys}$ yields a set of physicalistically acceptable truths, $TRUE_{phys}$ in a language of purely physical (and chemical, biological, and neurological) vocabulary L_{phys}. We want to extend this language to a language $L_{phys+psy}$ that includes the vocabulary of psychology, with predicates that assign beliefs, desires, and sensory experiences to a class of sentient and rational bodies (the human beings). Lewis assumes that we are already given, not only the vocabulary of $L_{phys+psy}$, but also a fairly rich theory of *folk psychology* T_{folk}, that specifies a large number of connections between psychological and physical states. This will include facts about the sensory experiences resulting from sensory-organ stimulations, coordinated in such a way that experiences are veridical under normal conditions. It will also include connections between belief-desire pairs and overt behavior, and certain kinds of overt behavior that results directly from certain experiences or desires, like wincing from pain.

- Let's assume that T_{folk} is consistent with the set of physical truths, $TRUE_{phys}$.
- If so, we can find an interpretation function $I_{phys+psy}$, relative to which T_{folk} is

true in the true model of the physical world, $M_{true-phys}$.

- If there is such a function, it will be a *realization* (in Ramsey's sense) of the folk theory of psychology.
- If there is a unique such function, then we can use it to define the set of *all* psychological and psychophysical truths bysimply identifying it with the set of sentences $TRUE_{phys+psy}$ in the language $L_{phys+psy}$ that are verified by the model $M_{true-phys}$ as extended by the interpretation function $I_{phys+psy}$.

Lewis is entitled to help himself to the psychophysical language $L_{phys+psy}$ and the folk theory T_{folk} in that theory, since the facts about what language humans speak and what sentences in that language they assert can be recovered with a high degree of determinacy from the physicalistically and behavioristically acceptable set of facts, simply by consulting users' overt verbal behavior (including their counterfactual behavior under all possible circumstances). This is the sort of task that Donald Davidson described as *radical interpretation.*[**?**] In any case, overt linguistic behavior (as described in $TRUE_{phys}$) would place very severe constraints on acceptable candidates for the language $L_{phys+psy}$ and the folk theory T_{folk}.

In addition to or as an alternative to reliance on the folk theory T_{folk}, we could rely, as Donald Davidson recommended, on a Principle of Charity, which could serve as a constraint on acceptable interpretation functions. We could require that the interpretation of sentences that attribute the belief with content ϕ be assigned intensions in which ϕ is also verified, at least to as great an extent as possible. We could also apply a similar Principle of Charity to the assignment of sensory and mnemonic contents to human subjects, along with a Principle of Humanity or Reasonableness that requires that beliefs be reasonable, given a subject's sensory and mnemonic information.

3.3 Wallacian Functionalism

In a sense, Wallace's functionalism, inspired by Daniel Dennett's *Real Patterns* [**?**], is a combination of phenomenalism and analytical functionalism, with mental properties reduced to macroscopic (and chemical and biological) properties in something the form of Analytical Functionalism, and macroscopic properties reduced to states of the quantum wavefunction, in something like Russell's structuralism. The difficulty with this strategy, as we'll see, is that this leaves us trying to lift ourselves by our own bootstraps, with too little basis for constraining the kinds of emergent

domains that can emerge.

It's reasonably clear what the reducing or fundamental model is supposed to be. We can take the language of pure quantum mechanics (with its description of the cosmic wavefunction and its deterministic Schrödinger evolution) and supplement it with a counterfactual or subjunctive conditional. This will require a model that contains a domain of worlds, each of which consists of a single quantum wavefunction evolved through time, one world designated as actual (which picks out the world's actual wavefunction), and a system of spheres S for the evaluation of subjunctive conditionals (the *worlds* of these models will not be Everettian branches, but different versions of the underlying quantum wavefunction). The system of spheres could be based, as in David Lewis's semantics[**?**], on a relation of comparative similarity between quantum worlds. This would require something beyond pure quantum theory, and in that sense Wallacian functionalism does, like other interpretations of quantum mechanics, require some substantial supplementation to the theory. However, we might hope that the mathematics of the Hilbert space would provide us with a unique, natural measure of distance between possible wavefunctions, or at least a fairly small family of such measures. We would still have to decide whether to follow David Lewis's proposal, in which we must include among the possible worlds those with small, localized "miracles," to be preferred in closeness to worlds that verify the antecedent of the conditional that are non-miraculous but otherwise quite far from the actual world. Deciding how many such miraculous worlds to include and how to weigh their comparative similarity will introduce a large measure of subjectivity or conventionality to the project. However, for the sake of argument, I will waive objections along these lines.

In Wallace's proposal, the only constraint on the Ramsey realization of the emergent theory is this: all fillers of functional roles in the emergent theories must be entities and sets of entities to be found in the correct model of the formal language of pure quantum mechanics. In particular, there are no constraints on the extended interpretation function that can be expressed in terms of *causal connections* between emergent and quantum-mechanical entities or *pure semantic conditions* (such as metaphysically correct reference or truth-conditions for the emergent language) or *metaphysical priority* (no degrees of *naturalness* or *eligibility* that apply to sets of n-tuples quantum-mechanical entities), as I will argue in Section 3 below.

There is, however, another difficult choice for the Wallacian functionalist to make: is the high-dimensional space ($3N$, where N is the number of particle-systems) of the quantum wavefunction the whole of fundamental reality, or is there in addition a 4-dimensional spacetime upon which the higher-dimensional space is defined? Wal-

lace and Timpson [**?**] prefer the latter, but this seems ad hoc and artificial, given their wavefunction Puritanism. Why posit the 4-dimensional manifold, if there are no fundamental entities located there? The Schrödinger dynamics doesn't depend in any way on the familiar four-dimensional structure. In addition, this position seems inconsistent with the attempt to use decoherence to explain all of classical physics: see Halliwell's attempt to generate three-dimensional space as an *emergent* by-product of quantum cosmology.

In fact, by moving from a pure quantum wavefunction (in its $3N$-dimensional state space) to the wavefunction plus a four-dimensional manifold, Wallace and Timpson are moving in exactly the right direction. I will simply argue that they should move still further in that direction, admitting still more to the fundamental ontology of the world. By embracing spacetime realism, Wallace and Timpson are admitting that there is at least one entity, spacetime, with fundamental existence and a real essence that stands over and above the austere mathematics of pure quantum theory. Since this is a move in the right direction, I will allow the inclusion of spacetime in the base model $M_{true-QM}$.

So, let's turn now to the emergent domain. Our first problem is a very basic one: What language do we use, and what theory in that language? In the case of phenomenalism, we had the common vocabulary of geometry and the necessary laws of perspective to constrain the language and theory of the emergent domain of physical objects. In the case of Analytical Functionalism, we had a folk theory of psychology and psychophysics that could be recovered from, or at least powerfully constrained by, the overt verbal behavior of human beings, all of which was contained within the base model of fundamental things. In addition, the beliefs and sensory states attributed by the emergent theory have contents that match the vocabulary of the base theory. Now, we have only the language and theory of pure quantum mechanics to begin with, which by itself tells us nothing about the languages and beliefs of the denizens of an emergent world, and which lacks the direct access to our beliefs and concepts of the physical environment, as was available for the phenomenalist.

So, it seems that we must use *every* possible language and *every* possible theory. There are no languages or theories and no language users or believers explicit at the level of quantum reality. Any constraints we place on these theories (besides their sheer interpretability in the model of quantum mechanics) are going to be constraints of internal coherency. That is, we might reasonably demand of any theory $T_{emergent}$ of the emergent world that, according to $T_{emergent}$ itself, the human beings speak the language of $T_{emergent}$ and have beliefs and sensory and mnemonic experiences that mostly accord with $T_{emergent}$. We can also require that $T_{emergent}$ have the theoretical

virtues valued by most people (as depicted in $T_{emergent}$), and that $T_{emergent}$ be well-confirmed, according to itself. Call the theories that meet these constraints the *internally ideal* or *coherent* theories.

3.4 Putnam's Permutation Argument for Semantic Indeterminacy

I will argue, in a way inspired by Putnam's argument for metaphysical anti-realism [**?**, **?**, **?**, **?**], that there is a radical indeterminacy of meaning and intension for all the names, predicates, and function symbols of the languages of our emergent theories. This isn't surprising, since all of the entities and properties posited by such theories are, from the point of view of Wallacian functionalism, mere useful fictions. In Wallace's picture, all that matters is that we find an interpretation of those theories in the true model of quantum mechanics that makes all of the formulas of that theory come out true (or at least approximately true) under that interpretation. The meaning of the emergent theories, the theories of the world's manifest image, is utterly holistic in character.

Suppose that $T_{emergent}$ is a theory of a world that is emergent relative to the model $M_{true-QM} = \langle F, D, I\rangle$. That means that there is an interpretation function, call it $I_{intended}$ that extends I to the language of $T_{emergent}$, resulting in a new model $M_{QM+emergent} = \langle F, D, I_{intended}\rangle$, with the theory $T_{emergent}$ true in $M_{QM+emergent}$. It is immediately obvious that there are an infinite number of alternative extensions of I that will also produce an extension of $M_{true-QM}$ relative to which $T_{emergent}$ is true. Take any permutation $\pi(w)$ for any world $w \in W$ of the objects in $D(w)$. Now apply the permutation $\pi(w)$ to the interpretation $I_{intended}$ with respect to the interpretation of all constants and predicate symbols at w. The resulting interpretation $I_{intended-\pi(w)}$ will also be a realization of $T_{emergent}$. Apply similar permutations to every world in W, resulting in the thoroughly scrambled interpretation $I_{bizarro}$. The extension of $M_{true-QM}$ by $I_{bizarro}$ will also be a model of $T_{emergent}$, and so $I_{bizarro}$ will be a realization of $T_{emergent}$ in $M_{true-QM}$.

So, for example, it is completely indeterminate what a predicate like 'is human' or 'is conscious' is true of or realized by. In the interpretation function $I_{bizarro}$, the intension of *human beings* might be the intension of *kumquats* in $I_{intended}$, and the interpretation of *is conscious* might be *contains vitamin B*. In fact, as Alexander Pruss [**?**] has pointed out, all the predicates that apply truthfully to the emergent world as it exists today (including mental-property predicates) could be interpreted

in such a way that they apply truthfully only to the cosmos as it was 12 billion years ago.

> Any two worlds that are isomorphic under an isomorphism of the quantum structure (i.e., of the Hilbert spaces and the operator algebras) have the same functional properties. Now consider two worlds w_1 and w_2. Both are short-lived worlds: the temporal sequence of each is only a billion years long. Each world is an exact duplicate of a temporal portion of our world. Thus, w_1 is an exact duplicate of the temporal portion of our world from 13 billion years ago to 12 billion years ago, while w_2 is an exact duplicate of the temporal portion of our world from a billion years ago to the present. Then w_2 has the same kind of mental properties that obtained in our world over the last billion years. And w_1 has the same kind of mental properties that obtained in our world from 13 to 12 billion years ago. But there is a quantum-structure preserving isomorphism from w_1 to w_2. This isomorphism is simply given by the time-evolution operator U_{12} (where we measure time in billions of years). This operator is an isomorphism of the quantum structure. Hence w_1 and w_2 are exactly alike with respect to mental properties. Hence our world had exactly the same mental properties in the early 13-to-12 billion-years-ago period as in the last billion years. That's absurd. (For one, it makes us question how we could possibly know that the world is as old as we think it is.)[**?**]

Here is the key difference between Wallacian functionalism and the phenomenalistic functionalism of a Mill or Carnap, or the behavioristic functionalism of David Lewis. In the case of a phenomenalistic functionalism, the fundamental or base theory is a theory of our phenomenological experience, and the target or reduced theory is one of the "external" world. In this case, there is arguably some constraint on the content of the reduced theory that is non-holistic. For example, in the case of the primary qualities, we could insist that the geometrical properties assigned to physical objects in our external theory resemble the geometrical properties of the corresponding inner phenomena. So, if the external theory asserts the existence of something tetrahedral in shape, we could insist that the corresponding model of the phenomenal world include something that at least appears tetrahedral (perhaps, a two-dimensional projection in visual space of a tetrahedron). However, in the case of Wallacian functionalism, there are no phenomenal qualia on either side of the equation. There are only sentences in our folk psychology assigning certain geometrical experiences to subjects, and so long as the interpretation of these sentences preserves their truth and their counterfactual inter-connections, we have met every constraint on a successful

interpretation.

In the case of behavioristic functionalism, we have real connections between the subjects of psychological states on the one hand and the subjects of behavior on the other. We assign beliefs and desires to x in a way that corresponds rationally to the behavior of x. In the case of Wallacian functionalism, we have only the universal wave function on the side of the base theory. All real or fundamental behavior is ultimately behavior of that function, and so there are no localized constraints on the connections between belief and desire and behavior. It is the theory of the folkish world as a whole (with both human belief and behavior contained in a single package) that confronts the model of pure QM as a whole.

In addition, as I will argue in section 4 below, Wallacian functionalists lack any of the resources used by metaphysical realists to meet the challenge of Putnam's argument: causal ties between emergent terms and their quantum-mechanical referents, specially eligible or natural properties (at a phenomenological level), so-called "reference magnets", or metaphysically primitive facts about semantics or reference.

There is one particular case of referential indeterminacy that is especially devastating to the Everettian interpretation: namely, it is indeterminate whether a particular emergent world is assigned to quantum states with a high or low amplitude. Thus, there is no objective fact of the matter about whether the quantum probability associated with a given "branch" is high or low. Thus, the problem is not just that of finding a reason for believing that we are in a high amplitude branch. The problem is that we cannot give any real meaning to the question itself. If a given emergent world is realizable in a low-probability segment of the wavefunction at a given time, there is another interpretation function (and thus another realization of that world in that wavefunction) that assigns it to a high-probability segment, and vice versa. It is all a matter of performing the appropriate permutation of quantum objects.

If there is no objective matter of fact about the quantum probabilities corresponding to the various emergent realities, then the Deutsch-Saunders-Wallace strategy for defending the reliability of observed statistics is further undermined. Emergent realities in which the statistics radically disconfirm quantum mechanics will be, not only as real as our own, but possessing the same status in regard to the underlying quantum probabilities. They will have just as much right to claim to be realized in the high-amplitude sectors of the quantum wavefunction as do the QM-confirming branches.

3.5 Model-Theoretic Indeterminacy Guarantees Truth of our Emergent Theories

Let $T_{emergent}$ be one of our target theories of the world: folk psychology or a scientific theory of "emergent" phenomena. We can suppose that $T_{emergent}$ is internally ideal and that it has a realization in the model of quantum mechanics, $M_{true-QM}$. Let $I_{intended}$ be the "intended" interpretation of the theory $T_{emergent}$ in the model $M_{true-QM}$, with a domain consisting of the spacetime regions and quantum subsystems of the quantum world and with the predicates of the language $L_{emergent}$ assigned appropriate intensions in the corresponding model $M_{true-QM}$.

Now consider a theory $T_{bizarro}$, whose intended model includes the same interpretation function $I_{intended}$ but includes a different, counterfactual model of the quantum world, $M_{counterfactual-QM}$. Both $M_{true-QM}$ and $M_{counterfactual-QM}$ have infinite models, and both T_{emerge} and $T_{bizarro}$ are semantically consistent with the hypothesis of a domain of infinite cardinality. By the Skolem-Löwenheim theorems, there is an interpretation $I_{bizarro}$ of $T_{bizarro}$ in the actual model of the quantum world, $M_{true-QM}$. Thus, the bizarro emergent world represented by $T_{bizarro}$ is realized in the actual quantum world in just the same way as $T_{emergent}$ is.

In fact, *all possible theories of emergent domains are actually true*: if they are logically consistent (in the logic of quantified counterfactual conditionals), and they contain no quantum-mechanical vocabulary and make no claims about the finite size of reality, then (by the Skolem-Löwenheim theorems), they have a model that extends $M_{true-QM}$.[3] In fact, just this point was made by H. A. Newman in 1928, as a criticism of Russell's structuralism.[**?**]

In fact, the situation is even worse than this, since Wallace doesn't require perfect realization in $M_{true-QM}$—just a reasonable degree of approximation to such perfect realization. So, even *inconsistent theories* or theories that entail the existence of a finite domain or that entail falsehoods about the structure of spacetime will nonetheless have quantum realizations and so will be *actually true* theories of a world that emerges from the quantum world.

The upshot is this: we are free to believe and say *whatever we want* about the emergent world of macroscopic objects, and we are guaranteed to believe and speak the truth (so long as our stories are internally coherent and not massively inconsistent).

[3]Remember: I assumed above that $M_{true-QM}$ has a sufficiently rich structure that we can embed in it the canonical model for the logic of counterfactuals. That means that any consistent theory in that language has an interpretation in $M_{true-QM}$, aside from issues of the cardinality of the world.

As a result, every consistent story corresponds to a real, emergent world, on par with our own. This includes the world of Tolkien's mythology or that of H. P. Lovecraft, the world of Harry Potter or Greek mythology. They are all just as real as our own. And, even more importantly, our own theory of the emergent world is true by a kind of stipulation: true simply by virtue of satisfying our demands for its internal coherency.

But that is surely is wrong. If our theory of the emergent world is true by a kind of stipulation, then it can't be interpreted realistically. To interpret it realistically is to take seriously the metaphysical possibility that it could be wrong. For example, all of the evidence we have for classical mechanics could be misleading ('could' metaphysically, not 'epistemically'): it could have been produced by some other quite unknown mechanism. For example, the planetary orbits that led to Kepler's laws and ultimately to Newton's laws of motion could actually have resulted from the fact that the planets move on gigantic rails built by ancient aliens. For the theory of classical mechanics to be a substantive theory of the world, it must have the metaphysical possibility of being wrong. But Wallace's functionalism denies it that chance.

3.6 Epistemological and Pragmatic Consequences

If classical physics is understood as true by stipulation, this undermines any rational confidence we might have in quantum mechanics, since a large part of our evidence for QM consists in its agreement with classical mechanics when interference terms are small.

In fact, Wallace's functionalism leads quickly to an epistemological catastrophe: if we cannot interpret our theories of the emergent world realistically, then no belief in such a theory can count as objective knowledge. And yet all of our knowledge of the truth of quantum mechanics depends on our having objective knowledge of experimental data that belongs to an emergent domain. So, in the end, Wallacian functionalism is epistemologically self-defeating, destroying the only grounds we have for believing that quantum mechanics is true at all, to say nothing of believing that it exhausts the fundamental level of reality.

In fact, we couldn't even interpret our emergent scientific theories as *instrumentally* valuable in an objective way, since *any* theory of our future experiences would be equally true (just one more realizable emergent theory). In addition, what counts as *the same* qualitative properties of experience is itself up for grabs via the interpretation function. By choosing a suitable function, we can make any set of predictions

about future experience come out true.

Thus, pragmatism itself is inconsistent with radical indeterminacy of meaning, as Plato recognized in the *Theaetetus*:

> SOCRATES But, Protagoras (we'll say), what about the things which are going to be, in the future? Does he [the individual human being] have in himself the authority for deciding about them, too? If someone thinks there's going to be a thing of some kind, does that thing actually come into being for the person who thought so? Take heat, for example. Suppose a layman thinks he's going to catch a fever and there's going to be that degree of heat, whereas someone else, a doctor, thinks not. Which one's judgment shall we say the future will turn out to accord with? Or should we say that it will be in accordance with the judgments of both: for the doctor he'll come to be neither hot nor feverish, whereas for himself he'll come to be both?
>
> THEODORUS. No, that would be absurd. [**?**, p. 56, 178c1-10]

Every claim about the future, practical consequences about believing and acting on an emergent theory of the world will itself be part of some emergent theory of the world. I have shown that every such theory, so long as it is not massively inconsistent and doesn't entail the finitude of the universe, will be realizable in $M_{true-QM}$ and so will be true. Thus, we cannot appeal to pragmatic considerations (like avoiding being eaten by a tiger) as grounds for preferring some theories over others.

3.7 The Argument's Upshot in a Nutshell

To sum up, there are four disastrous consequences for Wallacian functionalism:

- Radical indeterminacy of content, via Putnam's paradox: there an infinite number of alternative interpretation functions mapping our actual theory of the world into M_{QM}. In particular, there is no fact of the matter as to the quantum probability associated with any given emergent world.
- Every consistent and internally coherent story (more precisely, every story consistent with an infinite domain) represents an emergent reality in M_{QM}, on a par with our current best theories about the macroscopic world.
- So, we can't go wrong in proposing theories about the emergent world we inhabit, so long as our theories are consistent with an infinite domain, and so

long as they are internally coherent from a semantic and epistemological point of view.

- These facts undermine any claim to know that quantum mechanics is true, on the basis of experiments and observations that depend in any way on the emergent. The impossibility of objectively false theories of the emergent domain makes objective knowledge of that domain impossible, including objective knowledge of the data and observations upon which we ground our claims to know the truth of quantum mechanics itself. Therefore, Wallacian functionalism is epistemologically self-defeating.

4 Putnam's Paradox: The Problem of the Missing External Constraints

The central problem for Wallace's proposal is that we are missing all the constraints that were available for phenomenalism or Analytic Functionalism. We have no counterpart to the laws of perspective that rigidly tied the physical world to the phenomenal data (in the case of Phenomenalism) or to the overt verbal behavior that rigidly tied (via a principle of charity) the acceptable psychophysical theories to the physical facts. All the constraints we have are constraints of consistency and coherency, but as Putnam's paradox shows, these are not sufficient to delimit the range of emergent worlds to any significant degree. These considerations demonstrate that no functionalist theory can bridge the gap between the quantum and emergent levels. What is needed is some additional, metaphysical constraint.

In this section, I will argue that the Oxford Everettian approach lacks the resources to build in such constraints. I will proceed by a process of elimination, showing that each of six plausible candidates cannot supply the necessary constraints on the interpretation of emergent theories in the Everettian setting. There are two possible constraints that we can eliminate quite quickly: an appeal to the concept of *emergence* itself, and a brute preference for the emergent theories that we actually endorse.

1. Can the concept of *emergence* fix the interpretation?

   Could we hope that the concept of *emergence* itself could somehow provide powerful constraints on what counts as an acceptable interpretation function for a candidate theory of an emergent world? It doesn't seem so: all that we

can say is that the emergent world must be realized by some such interpretation function. There are just aren't any a priori or analytic constraints on what that function must be like.

2. Can we use "our" actual language and folk theory?

   No, because first we would have to establish that such things exist, and that they are relatively determinate and well-defined. But that's just the problem. It seems that any coherent theory about what language we do in fact speak and what beliefs we do in fact hold will turn out to be equally true.

In the following subsections, I consider four additional candidates: (1) causal constraints plus the category of *natural* or especially *eligible* properties, (2) an appeal to facts about spacetime locations and trajectories, (3) the use of the *simplicity* of our emergent theories or of their quantum interpretations (or both), and (4) the use of facts about decoherence.

4.1 Can causal constraints or natural or eligible properties fix the interpretation?

One standard realist approach to the Putnam paradox, defended by Michael Devitt [**?**, **?**] and Hartry Field [**?**], is to appeal to a causal theory of reference. Such a causal theory could constrain the range of acceptable interpretation function, by requiring that the atomic truths including a given predicate be assigned to a property in the model of such a kind that an appropriate causal mechanism can be found between occurrences of that property in the world and uses of the predicate by speakers of the language. To apply this idea to Wallace's functionalism, we would have to require that there exist causal connections of the right kind between language use (or concept deployment) on the one hand and the emergent properties that we language users are supposed to be representing.

However, this strategy just won't work here, for three independent reasons, two having to do with the base theory and the other to do with the emergent theories. First, human language and concepts do not even exist in the quantum world to begin with, so the question of whether our language use or concept deployment is causally connected to anything at all cannot arise, independently of assuming the truth of a given emergent theory and the correctness of a given interpretation function. In contrast, classical phenomenalism and Russellian structuralism included concepts and concept-use within the base theory, and so the issue of what properties that concept

use could be causally connected to could constrain the interpretation of emergent theories. And, although Analytic Functionalism did not include concept-use within the base theory, it did include language users and their overt linguistic behavior, which could be used to tie concepts to fundamental causal connections. In Wallace's functionalism, any causal connections that involve concepts must be inextricably part of the emergent story itself. Hence, they can be of no help whatsoever in linking the language of the emergent theory to the underlying quantum world. The emergent causal connections that exist occur entirely within the story or theory that defines the emergent world. And so, at best we can simply add another coherence condition to our emergent-world stories: namely, that *in the story* there are the right sort of causal connections between (emergent) environmental conditions and (emergent) language- and concept-use.

Second, there are no causal connections between facts in Wallacian functionalism in the base model: all we have are counterfactual conditional dependencies. There is good reason for this, since the pure formalism of quantum mechanics contains no non-Humean information about *causation* over and above the facts about what counterfactually depends on what. But counterfactual dependence is not sufficient to fix reference determinately, as we have seen.

Third, emergent theories did not historically and still do not generally include any information about the quantum realm. Hence, we cannot even require ideal or coherent emergent theories to include causal connections between concepts and underlying quantum processes.

A standard approach to resolving Putnam's paradox, championed by David Lewis [**?**], is to appeal to *perfectly natural* or *eligible* properties that can serve as "reference magnets" for the interpretation of predicates. However, to qualify as *perfectly natural* a property must be one of the fundamental properties of the world. In the case of Wallacian functionalism, this would not enable us to move far enough away from the austere properties of pure quantum mechanics to define the properties and relations of macroscopic branches, to say nothing of biological and psychological properties.

4.2 Can we use spacetime to restrict the interpretation function?

As I mentioned above, I am willing to embrace, for the sake of argument, the Wallace-Timpson theory of spacetime realism.[**?**] So, there will be some common vocabulary

between many emergent theories and the theory of quantum mechanics: both will have the vocabulary and the axioms needed to characterize the spacetime continuum.

However, beyond the vocabulary of space and time, there will presumably be no other overlap between our quantum and emergent vocabulary. It's clear that there is no such overlap in "our" emergent world (if it exists). This still leaves us with a very weak constraint on true emergent theories: they must not entail anything false about the structure of spacetime. This is very unlike the situation in the case of classical phenomenalism, were we could assume a common vocabulary about what regions of space are *occupied* by bodies at what points in time. We could use simple laws of perspective to link bodies in public space with sense-data in private spaces.

Couldn't we restrict the interpretation function by requiring that macroscopic properties assigned to regions of spacetime must be interpreted by quantum properties assigned to the same region? But the problem of cosmic entanglement ensures that there are no quantum properties that are localized to any finite region. If we take spacetime seriously, we must understand the quantum wavefunction to be what Peter Forrest calls a *polyfield*, in which fundamental magnitudes are assigned to N-tuples of widely separated spacetime points (for a very large N, representing roughly the number of fundamental particles in the universe).[**?**] There is, therefore, no simple function from quantum events in a space-time region and macroscopic object-events or phenomenal appearances.

Furthermore, appeals to spacetime won't help to blunt the Putnam-style argument for referential indeterminacy. We can still get complete indeterminacy for every term and predicate except for the geometrical ones. For example, Pruss's argument based on the 12-billlion-year time-shift will still work. We can insist that the events of the emergent world be located somehow in the real spacetime continuum of $M_{true-QM}$, but there is no way to ensure that they line up in any fixed or determinately intended way with the pattern of quantum events in that same continuum.

Here again the contrast with classical phenomenalism is instructive. A classical phenomenalist could, in effect, locate phenomenal qualia in regions of public spacetime and then stipulate that the sensory qualities of any physical object located in that same region by the emergent theory correspond to those of the qualia. However, for the Wallacian functionalist qualia are themselves just parts of the emergent theory. Moreover, they are parts that presumably share no intrinsic features with the underlying quantum events.

4.3 Why not simplicity?

Wallace can legitimately complain that I have ignored the constraint that he mentions explicitly: the constraint of *simplicity*. Wallace could impose on the extended interpretation function a condition related to Lewis's condition of *naturalness* [**?**] by requiring that the function map emergent predicates onto relatively simple sets of n-tuples of relatively simple quantum entities.

But before we examine the utility of a simplicity requirement, we must ask a more fundamental question? Why should *simplicity* be of any relevance to the *metaphysical* and *ontological* question of defining emergent reality? Simplicity is plausible as an epistemological constraint: other things being equal, the fact that one theory with wide scope and a high degree of accuracy is much simpler than all of its competitors with comparable scope and accuracy seems a good (if defeasible) reason for thinking that the simpler theory is objectively true. Simplicity may be an indicator of *probability of truth* but it does not seem to be a criterion of existence or reality. How could the real existence of an entity be a function of its simplicity?

The metaphysical deployment of simplicity is especially implausible once we remind ourselves that the very definition of *simplicity* is difficult and contentious. Simplicity, like beauty, seems to be in the eye of the beholder. Different theories of the emergent world will attribute different standards of simplicity to the scientific community. More fundamentally we can ask: what is the truthmaker for the "correct" account of simplicity here? And what is the metaphysical ground for imposing any simplicity constraint at all?

There are two independent parameters of simplicity to consider: (1) the simplicity of the *theory* of the emergent world (is it, for example, finitely or recursively axiomatizable, or at least approximately so?), and (2) simplicity of the *interpretation function* that interprets the non-quantum vocabulary in the quantum model.

4.3.1 Maximizing simplicity

It might seem that simplicity provides a solution to Putnam's paradox: take the correct interpretation function to be the *simplest* extension of the base interpretation that verifies the emergent theory $T_{emergent}$. However, this will only work if we are given the complete theory of some emergent reality. The problem is that there is at the quantum level no set of privileged theories of emergent domains. And any

consistent theory, no matter how bizarre, will have *some* realization in the quantum model and so, in all likelihood, some simplest realization.

Can we pick out the privileged emergent theories by focusing on the *simplest* theories that are realizable in the quantum model? We can't maximize simplicity of both theory and interpretation simultaneously, since the simplest possible interpretation function is just the original interpretation function of the quantum model (with no addition), and the simplest possible theory is the just theory of the original quantum model (the totality of purely quantum truths, including the theorems of logic).

In other words, the simplest emergent theory is just the null theory: the theory consisting of nothing but the theorems of logic. That will obviously give us no help in fixing the interpretation function. The simplest extension of the interpretation function is just identity, and that will give us no help in fixing the emergent theory. Maximization is futile in either case.

Maximizing the simplicity of one parameter or the other might be useful if we could first fix the metaphysically correct value of the other parameter. For example, if we could independently fix the true theory (and language) of the emergent world, it might make sense to try to maximize the simplicity of the extended interpretation used to realize that theory in the model of quantum mechanics. Or, if we could independently restrict the set of interpretation functions to a very narrow class, we could then use the degree of simplicity of the resulting emergent theories as a way of selecting the "best" emergent world. However, we have been unable to find any independent constraint. How, for example, do we select the right emergent theory to use? For any consistent theory of the emergent world, no matter how bizarre, there will be some interpretation that maximizes the simplicity of its truthmakers in $M_{true-QM}$.

4.3.2 Setting a minimal degree of simplicity

Perhaps instead of *maximizing* the simplicity of one or the other, the Wallacian functionalist could just require that the simplicity of one or the other (or both) has to meet a certain fixed standard, allowing any theory whose realization meets that standard to count as really emergent. Let's focus on the simplicity of the interpretation function. I see three problems here.

1. Any requirement of relative simplicity will have to be quite loose and permissive, since we know that the entities and properties of the emergent manifest

image are, under the most optimistic assumptions, far from natural. See recent work on color and color-experience, for example ([**?**, **?**, **?**]). Neurological and other biological properties will be highly disjunctive and gerrymandered from the viewpoint of fundamental quantum mechanics, and phenomenological, intentional, and semantic properties even more so. If the requirement of simplicity is too strong, it would give us *no* emergent worlds at all; if too weak, it would give us far *too many*.

2. What could be the truthmaker or metaphysical ground of the correct standard of minimum simplicity? How are we supposed to explain the connection between complexity and unreality?

3. There has to be some counterweight to simplicity, or we should embrace an eliminativist theory (a no-emergence theory), in which our theory of the "manifest" world just is fundamental quantum mechanics. We need some reason not to set the standard at the maximum level of simplicity. So, what is the counterweight? Usefulness? Apparent truth? But these criteria only make sense given a manifest theory. We need people, organisms, perceptions, beliefs, purposes, etc. in order to make these judgments.

Again, Plato's *Theaetetus* point applies again. Usefulness cannot be both part of the emergent picture and the criterion for which picture is really emergent. If there's nothing to balance against simplicity, then simplicity becomes either too strong a constraint (eliminating all emergent worlds) or a completely vacuous one. If we are going to be pragmatists, there has to be some kind of bridge between the emergent world we posit and our actual experiences and beliefs (how the world *appears to us*—in both a non-epistemic and epistemic sense of the phrase). But the phenomenal world, as we might call it, is also *part* of the emergent world, and so just another part of the theory it is supposed to be used in evaluating.

4.3.3 Degrees of reality

Could we talk about degrees of reality, as measured by the simplicity of the isomorphism into the wavefunction? Instead of requiring maximum simplicity of our emergent reality, and instead of arbitrarily setting some minimum standard of simplicity, we could adopt a kind of sliding scale: the simpler a theory and the simpler its realization in the one true quantum model, the *more real* the corresponding emergent world would be.

But can we make any sense of one world being *more real* than another? Isn't reality (whether emergent or fundamental) always a simple matter of Yes or No? In addition, why should we accept any level of reality short of the highest one? What counter-pressure could make us prefer a theory that is to any degree unreal? As we have seen, neither conservatism nor pragmatism provide any such independent counter-pressure.

Finally, there would still be a huge number of alternate emergent realities that will count as equally real (by this standard) as our own world. In fact, each of the bizarro permutations of the "intended" interpretation function $I_{intended}$ will be just as simple, considered as mathematical functions, as the original interpretation function.

4.3.4 The best option for Everettian functionalists

There might be some emergent theories that have a uniquely simplest realization in the quantum model: a realization that is *much simpler* than the second-best realization. We could stipulate that it is exactly such theories that represent an emergent reality.

However, as Wallace has pointed out, it is unlikely that any of our quasi-classical branch theories have a uniquely simplest realization. They are all afflicted with a significant degree of vagueness and indeterminacy. Still, we might hope that such theories have a *relatively compact* class of simplest realizations, each of which is *significantly* simpler than any realization outside of the class. In other words, there might be a fairly sharp peak of simplicity associated with certain realizations of the theory. We could stipulate that only such theories represent emergent realities, and that only interpretations at or near the unique peak of simplicity count as correct interpretations of the theory. To be more exact, we should look at the class of simplest *near* realizations of the theory: extensions of the interpretation that verify most of the sentences of the emergent, or most of the most important ones.

There will be some considerable amount of work to be done to show that the quasi-classical theories of the Everettian branches actually meets this new, more stringent condition of emergence. First, the condition might be too broad. For all we know, there are many fantastic theories that have a simplest interpretation at the quantum level (nota bene: this simplest interpretation could be horrendously complex, so long as all the others are even worse). Imagine a panpsychist or even pan-voluntarist theory, according to which all physical entities have sensations and make free choices. This would count as a real emergent world, so long as there is one interpretation

function for the fantasy that is significantly simpler than the others. Second, the condition might be too narrow. We know very little about the class of possible interpretations of our emergent theories. We know that in many cases the intended interpretation of the emergent theory (e.g., geology, thermodynamics, psychology) is extravagantly complex. Is that intended interpretation always uniquely simplest of all the possible interpretations? It is hard to tell. In fact, Alexander Pruss's argument (recounted in section 3.4 above) shows that the intended interpretation for our theory of astronomy is not significantly simpler than an alternative interpretation that shifts our descriptions of the current state of the cosmos backward 12 billion years.

But perhaps the technical viability of this solution can be verified. There remain three philosophical difficulties. First, the criterion of simplicity is highly subjective and variable, probably even conventional. Simplicity is in the eye of the beholder, but the beholder is part of the emergent world, and so simplicity cannot provide an independent, exogenous constraint on the identification of real emergence. There are possible theories of the world in which the conscious inhabitants have radically different standards of simplicity from our own, and so the *simplest* interpretations of the emergent theories by their lights will be quite different from our own.

Second, this solution raises severe epistemological worries. How could we know that our current theories of the emergent world satisfy the constraint of having a uniquely simplest interpretation in the quantum field, without having independent access to the level of quantum reality? We can presumably know that the truths, both categorical and subjunctive, of our favored emergent theories are verified somehow or other by the quantum level, but how could we know whether they are verified according to an interpretation that is substantially simpler than any alternative?

Finally, the criterion of simplicity is inherently vague and indeterminate. How could reality itself select one precise version of simplicity as the ground of determinate interpretation? It seems that we would have to look for that property of simplicity that is uniquely *natural* or *eligible*. But, the simplicity of emergent theories is not a feature of the underlying quantum reality, and so the Everettian have no grounds for attributing extreme degrees of naturalness or eligibility to such emergent realities. We could always pick out that precise form of simplicity that favors our current theories of emergent science, but such a tactic would be evidently ad hoc and metaphysically unmotivated.

4.4 What about decoherence?

I haven't said much yet about decoherence. Surely that's a problem, given the prominent role that decoherence plays in Wallace's account of the emergence of the macroscopic world. What exactly does decoherence tell us? It tells us that under certain circumstances, quantum systems can mimic the dynamics of classical mechanics, because of the way in which environmental interactions suppress the interference terms in the systems' Hamiltonian operators.

How is that relevant to our theories of the emergent world? It's relevance seems to depend on two assumptions:

1. The dynamics of the emergent world must be (approximately) that of classical (Newton-Maxwell) mechanics.
2. The dynamics of the emergent world should closely mimic those of the underlying quantum reality.

Given these two assumptions, decoherence would indeed be crucial, since it would be needed to explain how and why the emergent world can exist, given that the underlying quantum reality does *not* generally obey classical mechanics. Decoherence gives us reason to believe that, under most normal circumstances, the dynamics of actual quantum systems will effectively *approximate* those of classical mechanics.

But what *metaphysical* grounds do we have for accepting either of these assumptions? It seems clear, in fact, that they are both false. It has never been obvious that the emergent world of everyday macroscopic objects obeys Newton-Maxwell dynamics. If it had been obvious, it would not have taken scientists millennia to transcend the limitations of Aristotelian, Archimedean, and Galilean mechanics. Even today, most physical phenomena do not apparently obey classical laws, as Nancy Cartwright pointed out in *How the Laws of Physics Lie*.[**?**] In addition, there are many special sciences, including especially social sciences like politics, economics, and sociology, in which classical mechanics plays little or no role.

In addition, why should we assume that *any possible* emergent world must be classical in its dynamics? What basis is there for such a stipulation? It doesn't seem to be built into the concept of *emergence* in any way.

Turning to the second assumption, there seems again to be no reason to suppose that the dynamics of the emergent systems must mimic those of the underlying physical reality. Surely it is sufficient if they are functionally realized by that reality.

In any case, even if we grant both assumptions and thereby limit the emergent world to conditions in which decoherence obtains, this is still going to result in constraints that are far too weak. Any theory whatsoever that is consistent with classical mechanics and which entails substantive content only under conditions under which decoherence would obtain will have an acceptable interpretation in $M_{true-QM}$ and so will represent a genuine emergent world, as real as our own world. This would include bizarre Dan-Brown or John-Bircher conspiracy theories, theories of Aryan racial supremacy, the alternative histories of Philip K. Dick, UFO realism–so long as these respect classical dynamics in ordinary conditions, there will be a real emergent world corresponding to each.

Of course, if we keep loading up conditions on a *proper* emergent world, we will eventually isolate the theory we want. The following five conditions might work:

1. Extend the model $M_{true-QM}$ of quantum mechanics to a model $M_{QM+branches}$ with a branch parameter, each branch being assigned a probability weight, a period of time, and a set of particles in each world.

2. Privilege the basis consisting of position and momentum by having the extended model $M_{QM+branches}$ assign definite but branch-relative position and momentum to each particle that belongs to the branch and at each time assigned to the branch.

3. Require that the sum of branch-probabilities corresponding to a set of particle positions (or momenta) be a good approximation to corresponding sum of probability amplitudes in the original model $M_{true-QM}$.

4. Require that the dynamics assigned to particles by branches approximate a dynamic theory that is both simple and relies on highly localized, separable quantities (i.e., very like classical mechanics) .

5. Require that the branch structure include as many particles and as much time as possible, given the other constraints.

We can then stipulate that the only real emergent worlds correspond to the set of truths verified by some such extended model $M_{QM+branches}$. In addition, we could count as an emergent theory a theory that is expressed in a reduced language of $L_{absolute}$, reduced by the replacement of each branch-relativized predicate with an absolute version of the same predicate (including location and momentum predicates). A theory $T_{absolute}$ in the reduced language could count as *realized* by $M_{QM+branches}$ just in case there is a consistent assignment of branch-parameters to the formulas of $T_{absolute}$ results in a theory that is verified by $M_{QM+branches}$.

However, such a move has four disadvantages:

1. The account is no longer tied to and no longer provides a *general theory* of emergence. Consequently, we would have to deny the emergent reality of other special sciences, like chemistry, thermodynamics, biology, psychology, and the social sciences.

2. We would be offering no account of why these conditions are of great *metaphysical* significance. Why must an emergent world satisfy just these conditions to count as real?

3. We would give up the claim that decoherence generates the privileged basis by itself. Instead, we would simply be stipulating what we shall count as the correct basis. If we try to get around this by deleting conditions (2) and (3), we will be unable to dissolve the Putnam paradox, since permutations of the intended model of the emergent world will meet the other three conditions. This would leave Pruss's time shift argument, with the superfluity of minds in obviously mindless regions, untouched.

4. We would be making the many-worlds or many-branches structure of emergent reality true by stipulation.

5 The Solution: Real Essences and Extra-Conceptual Grounding

5.1 Two Forms of Grounding

My real complaint is with Daniel Dennett's "Real Patterns" [**?**], which is the original inspiration for Wallace's functionalism. I believe, in fact, that my arguments in section 3 above would apply with almost equal force to *any* functionalist account of the emergent world, including Bohmian interpretations of quantum mechanics. And the deterministic version of Bohm's theory (in which everything occurs with probability 1 or 0) also has difficulty accounting for the quantum probabilities. (I leave this extension as an exercise for the reader.)

The problem for Dennett is this: What makes a pattern *real*? As we have seen, mere realizability in the true quantum model of the world does not suffice. Here is the crux of the problem: we have to construct both the theory and its semantics

simultaneously, with an aim toward maximizing simplicity (in both dimensions). What then keeps us from simply collapsing into the identity isomorphism and the trivial theory? If we had a fixed theory and had to find the simplest semantics, or if we had a fixed semantics and had to find the simplest theory, the problem would be well defined and constrained.

We need some *top-down* constraint. It also has to be an *ontological* constraint: a set of natural kinds of emergent entities, each with a fixed real essence. These essences can then constrain both the story and the story's semantics. But to carry this out, we can't be eliminativists or physicalists—we can't limit the fundamental structure of reality to what can be described exclusively in quantum-mechanical terms. When things have real essences, they must be real.

In our forthcoming book *The Atlas of Reality*, my co-author Tim Pickavance and I argue for a distinction between *conceptual* grounding and *extra-conceptual* grounding.[**?**, pp. 62-5] On our view, *grounding* is an explanatory relation between truths or facts. Along with Kit Fine[**?**] and Gideon Rosen[**?**], we take all grounding facts to be underwritten by facts about some real essence or essences of entities involved. When we say that one truth is grounded in another, there are two importantly different cases to consider. First, it could be that the truth of p is grounded in the truth of q because of relations between the essences of some of the concepts or logical operators appearing in the two propositions. So, for example, the truth of a disjunction $(p \vee q)$ is grounded in the truth of the atomic proposition p (if both are true) by virtue of the essence of the logical operator of disjunction, $\vee$. Similarly, the truth of *John is a bachelor* may be grounded in the truth of the proposition *John is a never-married adult male human being* by virtue of the essence of the concept *bachelor*. In other cases, however, we must appeal to the essences of extra-conceptual entities, entities that are not essentially part of some abstract object of thought. For example, the existence of the singleton {Socrates} is wholly grounded in the existence of Socrates, but this grounding relation depends on certain facts about the essence of a singleton set like {Socrates}. Thus, the grounding of the existence of the singleton set in the existence of its member gives us no reason to eliminate sets from our ontology. We explain the distinction further in *The Atlas of Reality*:

> The distinction between conceptual and extra-conceptual grounding turns on a very subtle difference. Compare the following two claims, where '$[Fa]$' abbreviates the proposition a is F and '$[Ga]$' abbreviates a is G:
>
> - $[Fa]$'s truth is grounded in $[Ga]$'s truth.
> - $[Fa]$'s truth is grounded in a's being F, and a's being F is grounded

in a's being G.

> In both cases, the truth of $[Ga]$ is in some sense prior to, more fundamental than the truth of $[Fa]$. In the first case, the dependency is propositional or conceptual, in the latter case, extra-conceptual. To distinguish between the two, we have to look carefully at what licenses or justifies the explanatory connection between a is F and a is G: is it licensed by the essence of the property designated by the predicate F or the concept the predicate expresses? Is the essence involved in something in the mind-independent world, or is it merely in the mind? [**?**, pp. 63-4]

The two kinds of grounding have very different ontological import. Conceptual grounding gives us reason to think that the entities putatively designated by the concepts in the proposition whose truth is grounded in other propositions (lacking those same concepts) do not really exist. We can safely eliminate them from our ontology. However, as Pickavance and I explain, the eliminativist option breaks down in the case of extra-conceptual grounding: if the entities in the grounded fact didn't exist, they couldn't have an essence, and without an essence, the relation of extra-conceptual grounding between that fact and its ground could not exist.[**?**, p. 64]

It would be incoherent to say both that all grounding is conceptual and that reality is purely quantum-mechanical, because concepts and propositions do not appear at the fundamental level of quantum mechanics. And yet they must have non-trivial essences if the project of conceptual grounding is to work. This simple fact explains the incoherence of Dennett's attempt to reduce our existence to the reality of *patterns* in things as they appear to *us*. It's impossible for us concept-wielders to be both the grounds of emergent reality and merely another component of that emergent reality.

What's needed in a coherent account of a world that emerges from QM is the extra-conceptual grounding of emergent entities in the model of quantum mechanics. And the very existence of extra-conceptual grounding falsifies the claim that the quantum-mechanical exhausts the fundamental structure of reality. The real essences of emergent entities must be co-fundamental with the quantum-mechanical facts. These essences must themselves be ungrounded or perhaps zero-grounded, to use Fine's term. [**?**]

Once we posit new entities with their own essences that can explain (in a top-down fashion) those entities' grounding in the quantum domain, we open up the possibility that these composite, macroscopic entities (and not just their essences) are also

ontologically fundamental: metaphysically dependent on but not *wholly* grounded in (not fully explainable in terms of) the micro-quantum realm. In particular, there could be a diachronic and causal component to the answer to van Inwagen's Special Composition Question [**?**, pp. 21-2]: the existence of a composite macro-object on a certain branch of the cosmic wavefunction might be causally dependent on the prior existence of composite objects in that branch in the immediately preceding period, that is, partly dependent on composite objects in that branch with the required fundamental causal powers. This would of course fit well with Aristotle's vision of the world, in which the generation of new composite substances is always the result of the "corruption" of pre-existing substances, with the processes of corruption and generation being explainable in terms of the exercise of active and passive causal powers by participants in the processes. On a *hylomorphic* interpretation of Everettian quantum mechanics, the state of the quantum wavefunction (in particular, the presence of decoherent branches) is the *material cause* of the existence of certain composite, macroscopic entities. But the quantum function by itself is not a metaphysically sufficient ground or explanation: in addition, there must be a *formal cause*, reflecting the real essence of a natural kind of macroscopic, composite object. The presence of such a substantial form in a branch of the cosmic wavefunction at a particular place and time would have a diachronic causal explanation, which could make reference to earlier facts at the emergent scale in the branch in the spatial neighborhood of the persisting or newly generated composite substance.

To put the point more formally, we must enrich our base model, representing fundamental reality, by supplementing $M_{true-QM}$ with a set of natural-kind essences K and a fundamental composition relation $COMP$. The new model, M_{QM+HM} (HM for 'hylomorphism') would be defined over a language that contains constants for each of the natural kinds in K along with a four-place *part-of* predicate P, where $P(k, p_1, p_2, t)$ represents the fact that both particles p_1 and p_2 are at time t proper parts of a substance of kind k. The truth-conditions for the P predicate will be given by the fixed $COMP$ relation in the model, with the stipulation that, for fixed time t, each particle can be part of at most one substance. That is, a particle cannot satisfy the P predicate at the same time for two different natural kinds, and the binary relation of being two parts of a substance of a given kind will be an equivalence relation on particles (reflexive, symmetric, and transitive).

The natural kinds will make a real difference by virtue of constraining acceptable models to connect substances of each kind with appropriate branches in the branching-extension of M_{QM} defined by the five conditions in section 3.7 above.

M_{QM+HM} is an **acceptable model of the emergent world** if and only

if there is a branch extension of the base model $M_{QM+branches}$ meeting the five conditions in 3.7 such that, for any kind k in K, for any world w in W, for any particles p_1 and p_2 and time t in $D(w)$, if $M_{QM+HM} \models P(k, p_1, p_2, t)$, then there is a branch b in $M_{QM+branches}$ such that p_1 and p_2 belong to b in w at t, and the tuple $\langle p_1, p_2, t, b\rangle$ satisfies all of the metaphysical conditions associated with natural kind k.

The appeal to natural kinds of substantial forms enable us to overcome the four problems we identified at the end of section 3.7:

1. The account is a general theory of emergence. Every emergent domain depends on an appropriate set of natural kinds (macrophysical, thermodynamic, chemical, biological, etc.).

2. It is the fundamental existence of the emergent natural kinds that lends metaphysical significance to the constraints.

3. The essences of the natural kinds select the privileged basis of operators, by requiring a range of values for the corresponding parameters.

4. The essences of the natural kinds provide the ground for the truth of the multiple-branch structure within the wavefunction, since each essence requires the existence of a branch of an appropriate kind for its actualization.

5.2 Bonus: Restoring the Real World's Unity

Once we have real essences at the macroscopic level, we have the real possibility of diachronic, horizontal causation at that same level. In particular, we can consider positing a dynamic element to the solution of van Inwagen's *Special Composition Question*. We can call the result the *traveling forms* interpretation of quantum mechanics. This interpretation has some similarity to Jeffrey Barrett's *single-mind* interpretation [**?**], except where Barrett has a cohort of conscious minds traveling through the branching structure of the Many Worlds version, the traveling forms interpretation has instead a cohort of composite macroscopic objects. In Barrett's interpretation, all branches but one are occupied by zombies, by living human bodies that lack consciousness. In Barrett's picture, it is the presence of real consciousness that picks out the uniquely actual branch.

On my traveling forms interpretation, in contrast, all branches but one are occupied by pluralities of particles that fail to compose *anything at all*. We might call these

pluralities of fundamental quantum particles "compositional zombies". Although they have, from the microphysical perspective, everything that is needed for the potential existence of macroscopic objects (stars, planets, organisms, macro-molecules), no actual composite entities correspond to these branches. They are occupied wholly be compositional zombies.

Thus, the traveling forms version is not committed to anything like substance dualism: it is consistent with the supervenience of the mental on the physical, so long as the physical includes facts about which particles compose larger physical wholes. It does, however, deny that the *compositional* facts about physical entities supervene on the microphysical or quantum facts alone. Whether a branch corresponds to a domain of *actual* composite physical objects will depend on two factors: (1) Does the branch satisfy decoherence to a sufficient degree of exactitude (a degree determined by the real essences of composite physical kinds)? And, (2) Has the branch been occupied by composite objects in the immediate past, composite objects that are disposed either to persist in time or to generate new composite objects? In addition, there is a third, indeterministic factor: whenever a branch splits into two potential macro-worlds, the substantial forms responsible for composition jointly actualize macroscopic composition in just *one* of the new branches, with a probability normally determined by Born's rule (i.e., a probability proportional to the square of the amplitude of each branch).

I talk of traveling *forms* because I want to relate this interpretation to the Aristotelian theory of *substantial forms*. A substantial form is something (a principle or process) that is responsible for making what would otherwise be a mere cloud or heap of smaller entities into a single composite substance. For Aristotelians, a *substance* is an entity that is primary in the order of existence, an entity that has unity to the highest degree (per se unity) and which is the bearer of fundamental causal powers. It is an axiom of Aristotelian metaphysics that no substance is composed of other substances. Substances stand at the top of the compositional hierarchy. (For more details, see the chapter in this book by Alexander Pruss [**?**] and the appendix below.)

6 Conclusion

Clearly, what's needed to rescue Wallace's picture is some further constraint on the interpretation function. Wallace's intention is surely that this extra constraint should have something to do with decoherence, with a linkage of some kind between

macroscopic and quantum dynamics. However, as we have seen, the functionalist model that Wallace adopts, following the lead of Dennett's "Real Patterns" [**?**] won't deliver what is needed.

Ultimately, the extra constraint must have a top-down salient to it: it must derive somehow from the real essences of macroscopic *substances* (in Aristotle's sense, primary beings). But once we add these new element of constraint, it will be hard to resist the temptation to move still farther away from the Everettian picture. The essences of macroscopic substances can give rise to novel causal powers at that level, causal powers that can determine which branch of the quantum wavefunction is really occupied by composite substances of the appropriate kind, restoring a single, unified world to the picture and thereby avoiding the twin problems of possible trans-world values and of anti-Darwinian branches.

A The Traveling Forms Interpretation

My version of the Traveling Forms interpretation draws heavily upon the decoherence program and the work of the Oxford Everettians. I take decoherence as defining a set of branches, each of which constitutes the *potential* existence of an emergent realm of composite objects. The potential existence of emergence objects is a product of two things: the decoherence of a branch of the wave function, and a fixed inventory of macroscopic essences. However, the combination of these two is not *sufficient* for existence of any composite physical entity. In addition to the material cause (the branch of the wave function) and a formal cause (the macroscopic essences) there must also be an efficient cause at the emergent level: some pre-existing composite substances whose causal powers are responsible for jointly *actualizing* the potential of one of the branches to be the material substrate of further emergent composite substances.

When the actualized branch of the world splits into two or more potential successors, the substances making up the actual branch determine, through the exercise of indeterministic active and passive causal powers, which one of the successor branches shall be actual. Each exercise of such a causal power is a thoroughly local affair, but by actualizing one of the potential successor-states in its environment, each substance contributes to choosing a single branch for the entire cosmos. Hence, all of the substantial forms *travel* together through the branching structure of the decohering quantum world.

A.1 Traveling Forms and Ontic Vagueness

There is an obvious objection to this wedding of the Oxford Everettian QM with Aristotle's hylomorphism. The processes of decoherence produce branches that are only *approximately* classical. The emergent entities occupying each branch are only approximately localized in a spacetime region, and, in fact, the very number of branches is indeterminate, depending on how fine-grained a set of macroscopic descriptions we deploy. How then can it be a metaphysically fundamental fact that only certain macroscopic substances exist and exist in a way that corresponds to just one branch? The emergent entities of the Oxford school's version of Everettian QM are irreducible vague, and vague things cannot be fundamental.

There are two possible responses. First, one could hold that the substantial forms added to the theory by the traveling forms interpretation are able, in and of themselves, to fill in the indeterminacies left by the quantum wavefunction. So long as something sufficiently branch-like exists in the quantum wavefunction, the wavefunction is enabled to play its role as the material cause of the existence of macroscopic objects, with the substantial forms supplying definite locations for composite wholes, locations that are more determinate than the sum of the locations of their parts.

But there is a second solution that I think is preferable: simply assert that vague objects can be fundamental. Many philosophers have defended a thesis of *ontic* vagueness, vagueness in the world that is not merely the by-product of ambiguity or linguistic looseness. We could, for example, model such ontic vagueness by postulating that there can be more than one actual world: see Barnes[**?**] or Koons and Pickavance[**?**, pp. 275-9]. These multiple actual worlds must be "bunched" together pretty tightly (at least, at the macroscopic scale): no cases of cats that are both alive and dead are allowed. Indeed, as Pickavance and I argue, ontic vagueness seems unavoidable, since linguistic or conceptual vagueness entails ontic vagueness (since meanings and thoughts are things).

A.2 Three Bonuses for Traveling forms

First, the traveling forms interpretation ensures that everything a rational agent could care about exists in only one branch. Hence, Savage's Sure Thing Principle applies with exception, given the supervenience of value on being. Consequently, we can explain why rational agents must assign probabilities to the various possible branches in a way that respects classical probability theory. We can then appeal,

quite legitimately, to the sort of physical symmetries noted by Deutsch and Wallace, as grounds for identifying objective chance with the square of the quantum wave amplitude.

Second, the traveling forms interpretation solves the problem of anti-Darwinian branches. Once we return to a one-world interpretation of probability, we can again treat possibilities that have astronomically low probabilities as *close to* impossibility. We don't have to imagine that they are all equally denizens of reality. This feature of one-world probability is equally applicable to retrospective and prospective uses of probabilities. Hence, we can justify our a priori confidence that we do not inhabit a madly anti-Darwinian branch.

Third, the traveling forms interpretation enables us to ground scientific knowledge in the interaction with thoroughly local causal powers. Scientific knowledge is possible only because objects have active causal powers and we (and our instruments) have corresponding passive causal powers. Actually, experimentation depends on causal powers running in both directions, so we can suitably *prepare* the experimental situation by undertaking the appropriate interventions or manipulations. (See Nancy Cartwright, *Nature's Capacities and their Measurement* [**?**], and Brian Ellis, *Scientific Essentialism*[**?**].)

The multi-world Everettian functionalist has no room for localized causal powers. The quantum wave function is essentially non-separable. It can perhaps *simulate* localized causal powers, but simulation is not realization. Simulated experimentation is not experimentation. It can yield only the simulacrum of knowledge, not real knowledge. If simulated experiments were as good as real experiments, we could save a lot of money by abandoning our laboratories with their expensive equipment and just run all our experiments in CGI!