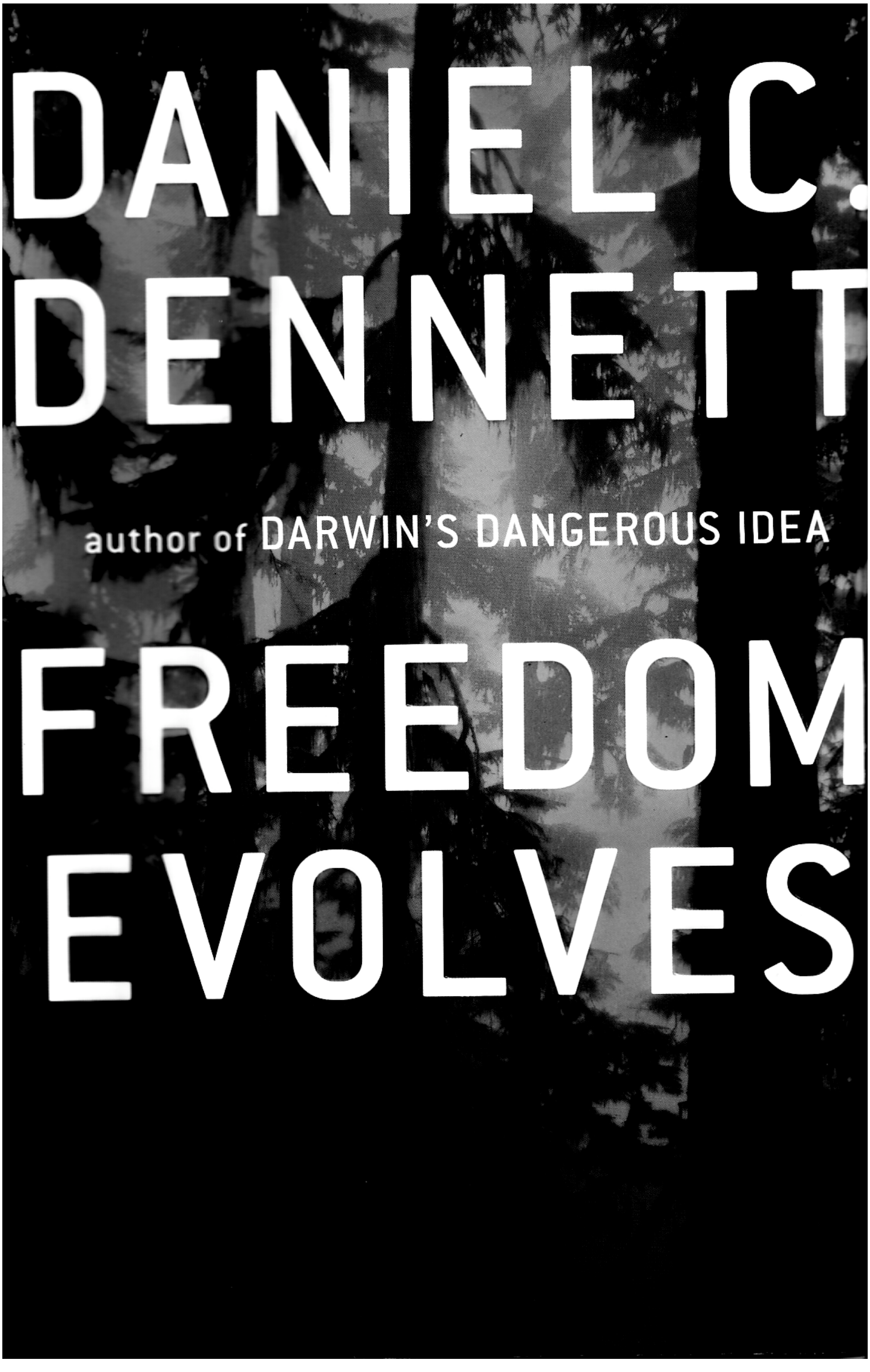


DANIEL C. DENNETT

author of DARWIN'S DANGEROUS IDEA

FREEDOM EVOLVES

Chapter 2

A TOOL FOR THINKING ABOUT DETERMINISM

Determinism is the thesis that "there is at any instant exactly one physically possible future" (Van Inwagen 1983, p. 3). This is not a particularly difficult idea, one would think, but it's amazing how often even very thoughtful writers get it flat wrong. First, many thinkers assume that determinism implies inevitability. It doesn't. Second, many think it is obvious that *in*determinism—the denial of determinism—would give us agents some freedom, some maneuverability, some elbow room, that we just couldn't have in a deterministic universe. It wouldn't. Third, it is commonly supposed that in a deterministic world, there are no *real* options, only apparent options. This is false. *Really?* I have just contradicted three themes so central to discussions of free will, and so seldom challenged, that many readers must suppose I am kidding, or using these words in some esoteric sense. No, I am claiming that the complacency with which these theses are commonly granted without argument is itself a large mistake.

Some Useful Oversimplifications

These errors lie at the heart of the misconceptions about free will and freedom more generally, so before we can make any progress on understanding how freedom could evolve (in a universe that may well be deterministic), we need to equip ourselves with some corrective devices, some tools for thinking that will make us less vulnerable to the siren songs of these powerful illusions. (If you have an aversion to philosophical argumentation about determinism, causation, possibility, necessity, and the indeterminism of quantum physics, you may skip ahead to Chapter 5, but you must then forswear all reliance on these three "obvious" propositions, no matter how intuitive they strike you, and take it on faith when I assure you that they are the false friends of a thousand misguided discussions. I almost guarantee that you cannot keep that resolution, however, so a better choice is to plunge into my demonstrations of these errors, which have their rewards and surprises, and presuppose no background expertise.)

In Thomas Pynchon's novel *Gravity's Rainbow,* a character makes the following portentous speech:

> But you had taken on a greater, and more harmful, illusion. The illusion of control. That A could do B. But that was false. Completely. No one can *do.* Things only happen. (Pynchon 1973, p. 34)

Pynchon's speaker has concluded that since atoms can't *do* anything, and people are made of atoms, people can't *do* anything either, not really. He is right that there is a difference between doing and mere happening, and he is right that there is a harmful illusion lurking in our attempts to understand this difference, but he gets the illusion backward. It is not the mistake of treating people as if they weren't composed of lots of *happening* atoms (they are), but *almost* the reverse: treating atoms as if they were little people *doing* things (they aren't). It arises when we overextend the categories appropriate to evolved *agents* onto the wider world of physics. The world of *action* is the world we live in, and when we try to impose the perspective of that world back down onto the world of "inanimate" physics, we create a deeply misleading problem for ourselves.

Getting clear about this aspect of the complex relationship between fundamental physics and biology sounds terrifying, but fortunately, there is a *toy* version of that relationship that is just what we need. The difference between a toy and a tool can evaporate if the toy can help us understand things that are otherwise too complex for us to keep track of. Science often uses toy models to great advantage. Nobody has seen an atom, but we all know what an atom "looks like": a tiny solar system, with a nucleus like a tight bunch of grapes surrounded by electrons orbiting every which way in their little halos. This familiar friend, the Bohr model (Figure 2.1), is of course hugely oversimplified and distorted, but for many purposes it's a great way to think about the basic structure of matter.

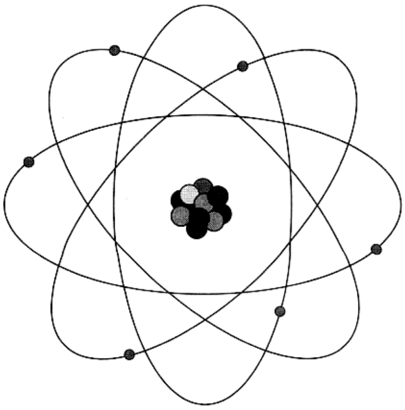


Figure 2.1 Bohr Atom

Becoming just as familiar in our common imagination is the gigantic Tinkertoy construction of a double helix with lots of rungs, the Crick-Watson model of the DNA molecule (Figure 2.2). It, too, is a useful oversimplification.

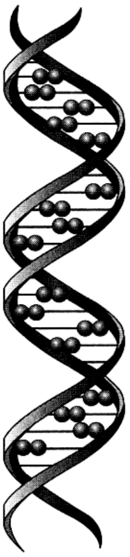


Figure 2.2 DNA Double Helix

The French physicist and mathematician Pierre-Simon Laplace gave us a usefully simple and vivid image of determinism almost two centuries ago, and it has structured our imaginations, and hence our theories and debates, ever since.

> An intellect which at any given moment knew all the forces that animate Nature and the mutual positions of the beings that comprise it, if this intellect were vast enough to submit its data to analysis, could condense into a single formula the movement of the greatest bodies of the universe and that of the lightest atom: for such an intellect nothing could be uncertain; and the future just like the past would be present before its eyes. (Laplace 1814)

Give this all-knowing intellect, often known as *Laplace's demon,* a complete snapshot of "the state of the universe," showing the exact location (and trajectory and mass and velocity) of every particle at that instant, and the demon, using the laws of physics, will be able to plot every collision, every rebound, every near miss in the next instant, updating the snapshot to yield a new state description of the universe, and so on, for eternity.

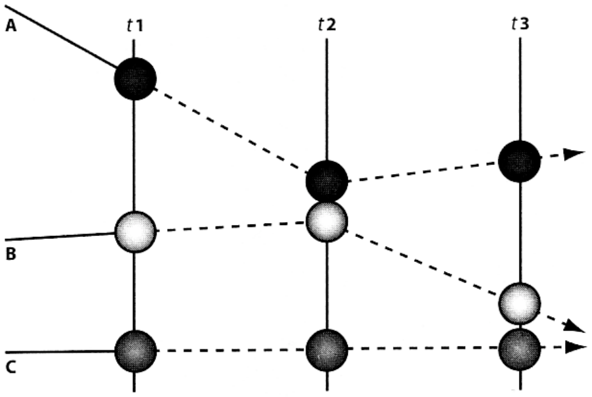


Figure 2.3 Laplacean Snapshot

In Figure 2.3, this snapshot zooms in at time *t1* on just three of the atoms in the world, on their various trajectories, and the demon uses this information to predict the collision and rebound of two of them at *t2,* leading to the positions at *t3* and so on. A universe is *determinis-*

tic if there are transition rules (the laws of physics) that *determine exactly* which state description follows any particular state description. If there is any slack or uncertainty, the universe is indeterministic.

There are too many fudge factors in this simple vision as it stands: How exact must a state description be? Must we plot every subatomic particle, and just which properties of the particles need to be included in the description? We can anchor these slippery factors arbitrarily by adopting another simplifying idea, W.V.O. Quine's (1969) proposal that we restrict our attention to simple imaginary universes, which he calls "Democritean" universes, in honor of Democritus, the most inventive of the ancient Greek atomists. A Democritean universe consists of some "atoms" moving about in "space." That's all. The atoms in a Democritean universe are not modern atoms full of quantum complexities but truly *a-tomic* (unsplittable, unsliceable) atoms, tiny uniform points of matter with no parts at all, rather like those postulated by Democritus. The space they inhabit must be made ultra-simple, too, by *digitizing* it. Your computer screen is a good example of a digitized *plane,* a two-dimensional array of hundreds of rows and columns of tiny *pixels,* little squares, each of which has, at each instant, one of a finite set of different colors. We want to digitize a space, a three-dimensional volume, so we need cubes—*voxels,* in the language of computer graphics. Imagine a universe composed of an infinite latticework of tiny cubical voxels, each one either utterly empty or utterly full (containing exactly one atom). Each voxel has a unique location or address in the latticework, given by its three spatial coordinates, $\{x, y, z\}$. Just as every computer color graphics system has a certain range of values—different shades of color—that each pixel can take on, in a Democritean universe, every voxel that isn't empty (value 0) contains one of a limited number of different types of atoms. It may help to think of them as different colors, such as gold, silver, black (carbon), yellow (sulfur). Just as we can define the set of all possible computer-screen images (for any particular pixel-color system) as the set of all permutations of fillings of the pixels with the defined colors, we can define the set of all Democritean-universe moments as the set of all permutations of fillings of all the voxels in space with the various sorts of atoms.

Now when we want to confront Laplace's demon with a "complete" snapshot from which to work, we can say exactly what we need to provide: a *state description* of a *Democritean universe,* which lists the values of every voxel at some instant. So part of state description S_k might read:

> at time *t:*
>
> > voxel {2,6,7} = silver,
> > voxel {2,6,8} = gold,
> > voxel {2,6,9} = 0,
> > . . . and so forth.

We don't have to worry about how "fine-grained" to make our description, since a Democritean universe has a defined limit, a smallest difference, and we can compare any two state descriptions of the universe and discover any corresponding voxels that are differently occupied. As long as there are a finite number of different elements (gold, silver, carbon, sulfur . . .) we can put all the state descriptions in order—alphabetical order, in effect—by voxel and the element occupying it. State description 1 is the empty universe at time *t;* state description 2 is just like 1 except for having a single aluminum atom occupying voxel {0,0,0}; state description 3 moves that lone aluminum atom to voxel {0,0,1}; and so forth, all the way to the last state description (in alphabetical order), in which the universe is filled—every voxel—with zinc! Now add time, the fourth dimension. Suppose that at the next "instant," the gold atom at {2,6,8} in S_k moves east one voxel. Then in S_{k+1},

> at time $t + 1$:
>
> > . . .
> > voxel {3,6,8} = gold.

Think of each "instant" of time as like a frame of computer animation, specifying the color or value of each voxel at that instant. This digitizing of space and time permits us to count differences and similarities, and to say when two universes, or regions or periods of universes, are exactly alike. A series of state descriptions, one for each successive "instant," yields the history of a whole Democritean universe, for however long that universe lasts—from its Big Bang to its Heat Death (or whatever replaces these openings and closings in these imaginary worlds). *In other words, a Democritean universe is like a 3-D digital video of some length or other.* We can cut time as fine as we like; thirty frames a second (like a movie) or thirty trillion frames a second, depending on our purposes. The size of the voxels is minimal: one indivisible atom per voxel, max. Quine proposed a further simplification: Imagine that the atoms are all alike (rather like electrons), so we can treat each voxel as either empty (value = 0) or full (value = 1). This option is just like replacing a color screen with a black-and-white screen, a simplification good for some purposes, as we shall see, but not necessary.

How many different ways are there of filling voxels with colors (or just with 0 and 1)? Even when we keep the size of a universe not just finite but tiny, the number of possibilities gets huge in a hurry. A universe consisting of just eight voxels (making a two-by-two cube) and one kind of atom (empty or full, 0 or 1), and lasting only 3 "instants," has already more than 16 million different variations ($2^8 =$ 256 different state descriptions, which can be put together in 256^3 different series of three). A second's-worth of the universe contained in a single sugar cube (at the *slow* rate of 30 frames a second and taking the cube to be *only* a million atoms wide) would be a number of states beyond imagining.

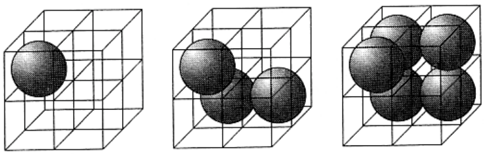


Figure 2.4 Three of the 256 different states of an 8-voxel Democritean universe.

In *Darwin's Dangerous Idea,* I introduced the term "Vast" as a name for numbers that, though finite, are Very much larger than ASTronomical quantities. I used it to characterize the not-really-infinite number of books in Jorge Luis Borges's imaginary Library of Babel, the set of all possible books, and by extension, the number of possible genomes in the Library of Mendel, the set of all possible genomes. I also coined a reciprocal term, "Vanishing," to characterize, for instance, the subset of *readable* books, nearly invisible within the Library of Babel. Let's call the set of all possible Democritean universes, all the logically possible combinations of atoms in space and time, the Library of Democritus. The Library of Democritus is mind-bogglingly large, no matter how tightly we restrict it to a particular finite set of parameters (types of atoms, durations, etc.). Things get interesting when we look at particular subsets of the Library. Some universes in the Library of Democritus are practically empty, and others are full of stuff; some have lots of change over time and others are static—the same state description, repeated forever. In some the change is utterly random—one instant of atomic confetti after another, with individual atoms flicking in and out of existence—and others show patterns of regularity and hence predictability. Why do some universes show patterns? Just because the Library of Democritus contains all the logically possible universes, so *every possible pattern* whatsoever is to be found somewhere in it; the only rule is that each state description should be complete and self-consistent (only one atom to a voxel).

Once we start imposing additional rules about what can be adjacent to what, and about how different state descriptions should succeed each other in time, we can get to more interesting subsets of the Library. For instance, we could prohibit the "annihilation of matter" by a rule that says that every atom that exists at time *t* has to exist somewhere at time *t* + 1, though it can move to a new voxel if that voxel is unoccupied. This guarantees that the universe never loses an atom as time passes. (More precisely, we "prohibit" this by just ignoring the Vastly many universes that don't obey this rule and restricting our attention to the Vast but Vanishing subset of those that do obey it: "Consider the set S of universes in which the following rule always holds. . . .") We could set up a speed limit (rather like the speed of light) by adding that an atom can move only to a neighboring voxel in the next instant, or we could permit longer leaps. We could say that matter *can* be annihilated—or created—under such-and-such conditions: For instance, we could have the rule that whenever two gold atoms are stacked one on top of the other, in the next instant they disappear, and in the lower voxel an atom of silver comes into existence. Such transition rules are tantamount to the fundamental laws of physics that hold in each imaginary universe, and we can usefully look at sets of universes in which these regularities are the same, whatever other differences there might be. Suppose, for instance, that we want to "hold physics constant" but vary the "initial conditions"—the state of the universe at its debut instant. We then consider the set of universes in which a particular transition rule or set of rules always holds but the starting-state descriptions are as varied as we like. This is rather like restricting our attention, in the Library of Babel, to those books written in (grammatical) English; there are regularities in the transition from character to character ("*i*" *before* "*e*" *except after* "*c*" . . . and *Every question begins with a capital letter and ends with a question mark. . . .*), but the topics covered are as varied as can be.

A better analogy between Borges's Library of Babel and our Library of Democritus would be the existence, in the Library of Babel, of Vastly many books that start out just fine—as novels or histories or chemistry textbooks—but then suddenly degenerate into nonsensical word salad, typographical gibberish. For every book that can be read cover to cover for enjoyment and profit, there are Vast numbers of volumes that start out well, with the regularities of grammar, vocabulary, story line, character development, and so forth that are prerequisite for *making sense,* but then degenerate into patternlessness. There is no *logical* guarantee that a book that starts well will continue well. The same is true of the Library of Democritus. This was David Hume's point, back in the eighteenth century, when he observed that even though the sun has risen every day so far, *there is no contradiction* in the supposition that tomorrow will be different, that the sun will not rise. To translate his observation into Library of Democritus talk, note that there is a set of universes, A, in which the sun *always* rises, and there is a set of universes, B, in which the sun always rises *until* [*say*] *September 17, 2004, at which point something else happens.* There's nothing contradictory about those worlds—they just don't turn out to "obey" the physics that always holds in universes in set A. Hume's point can be put this way: No matter how many facts you gather about the past of the universe you find yourself in, you can never prove, logically, that you're in a universe in set A, since for each universe in set A, there are Vastly many universes in set B that are identical to it at every voxel/time up to September 17, 2004, and then diverge in all manner of surprising or fatal directions!

Durchlegt

As Hume noted, we *expect* the physics that has held so far in our world to hold in the future, but we cannot prove by pure logic that it will oblige us. We've had conspicuous success discovering regularities that have held in the past in our universe, and we've even learned how to make real-time predictions, about seasons and tides and falling objects and what you'll find if you dig here, or dissect there, or heat this or mix that with water, and so forth. These transitions are so regular, so unexceptioned in our experience, that we have been able to codify them and project them imaginatively into the future. So far so good; it has worked like a charm, but there are no logical guarantees it will continue to work. Still, we have some reason to believe that we inhabit a universe in which this process of discovery can go on *more or less* indefinitely, yielding ever more specific, reliable, detailed, accurate predictions based on the regularities we have observed. In other words, we may take ourselves to be finite, imperfect approximations of Laplace's demon, but we can't prove, logically, that our success will continue, without presupposing the very regularities whose universality and eternity we would like to establish. And there are some reasons, as we shall see, to conclude that there are absolute limits on our capacity to predict the future. Whether these limits have any implications about our self-image as agents making "free" decisions and choices, for which we might properly be held responsible, is one of the treacherous questions we need to address, and we are approaching it gingerly, getting clear about simpler issues first. We're gradually approaching our target, *determinism,* by closing in on a Vast but Vanishing neighborhood in the still Vaster space of logically possible universes.

Some sets of Democritean universes have transition rules that are deterministic, and some don't. Consider the set of universes in which we specify that whenever an atom is surrounded by empty voxels it has a one-in-thirty-six chance of simply vanishing—otherwise, it stays put in the next instant. In such universes it is as if Nature rolled some dice whenever such an atom got itself isolated in this way; if the dice come up snake eyes, the atom "dies"; otherwise, it lives another instant and Nature rolls the dice again, unless that atom has just acquired a neighbor. This would be an *indeterministic* physics, which does not specify what happens next in all regards but leaves some of the transitions to mere probability. Laplace's demon would have to wait to see how the dice came up before continuing to predict the future. Other sets of universes obey transition rules that leave nothing to chance, that specify exactly what voxels are occupied by what atoms in the next moment. These are the deterministic universes. There are, of course, kazillions of different ways the transition rules for Democritean universes could be deterministic or indeterministic.

How do we *tell* what transition rules govern a particular Democritean universe? We can *stipulate* a rule and then consider what we must or might find to be true in all possible members of the set obeying the rule, but if we are somehow given a particular Democritean universe to study, the only thing we can do is examine the entire history of all its voxels and see what regularities—if any—hold. We can break the job into natural parts by looking for regularities that hold in the early going and seeing if they continue to hold all the way forward. Bearing in mind Hume's ominous discovery that we can never prove that the future will be like the past, we can nevertheless set out to find what regularities we can and make the huge but tempting wager—what do we have to lose?—that the future *will* be like the past, that we are not in one of those bizarre universes that leads us down the garden path only to disappoint us by going haywire after a longish period of regularity.

We now have a way of sorting Democritean universes into the deterministic, the indeterministic, and then all the junk—we might call these the *nihilistic* universes in which there is no permanent regularity of transition at all. Notice that on this construal, *all there is* to being deterministic or indeterministic is always exhibiting one sort of regularity or another—either a regularity with ineliminable probabilities less than one, or a regularity in which all such probability is absent. There is no room, in other words, for the claim that two Democritean universes are exactly alike at each voxel/time, but one of them is deterministic and the other is indeterministic.[1]

The difference between deterministic and indeterministic Democritean universes is now clear, but the best way of understanding just what it means (and what it *doesn't* mean!) is to pamper our overwhelmed imaginations even more and consider a still simpler toy image of determinism. First, let's drop from three dimensions to two (from voxels down to pixels), and let's also avail ourselves of Quine's black-and-white-only option, so that each pixel is either ON or OFF at

[1] Indeed, by definition, no *two* Democritean universes are exactly alike at each voxel/time. One of the virtues of Quine's simplification is that it lets us count universes the same way we count *editions* of books: If all the same elements are in the same places at the same times, that establishes *identity*. Quine's proposed taming of possible worlds also eschews the dubious idea that we need to know the *identity* of the individual atoms—not just their type, carbon or gold—to identify voxel contents from one universe to another. (Maven alert: This is not standard possible worlds lore; it avoids familiar problems of transworld identity.)

any instant. We have now landed on the plane where Conway's Game of Life spins out its astonishing patterns. This audaciously oversimplified toy model of determinism was developed in the 1960s by the British mathematician John Horton Conway. Conway's Life vividly illustrates just the ideas we need in a way that requires no technical knowledge of either biology or physics, and no math beyond the simplest arithmetic.

From Physics to Design in Conway's Life World

The complexity of a living individual minus its ability to anticipate (in respect of its environment) equals the uncertainty of the environment minus its sensibility (in respect of that particular living individual).

—Jorge Wagensberg, "Complexity versus Uncertainty"

Consider, then, a two-dimensional grid of pixels, each of which can be ON or OFF (full or empty, black or white).[2] Each pixel has eight neighbors: the four adjacent cells: north, south, east, and west, and the four diagonals: northeast, southeast, southwest, and northwest. The state of the world changes between each tick of the clock according to the following rule:

> *Life Physics:* For each cell in the grid, count how many of its eight neighbors is ON at the present instant. If the answer is exactly two, the cell stays in its present state (ON or OFF) in the next instant. If the answer is exactly three, the cell is ON in the next instant whatever its current state. Under all other conditions the cell is OFF.

That's all. This one simple transition rule expresses the entire physics of the Life world. You may find it a useful mnemonic crutch to think of this curious physics in biological terms: Think of cells going ON as births, cells going OFF as deaths, and succeeding instants as generations. Either overcrowding (more than three inhabited neighbors) or isolation (less than two inhabited neighbors) leads to death. But remember, this is just a crutch for the imagination: the two-three rule

[2] This introduction to Life is drawn, with revisions, from Dennett 1991A and Dennett 1995.

is the basic *physics* of the Life world. Consider how a few simple starting configurations play themselves out.

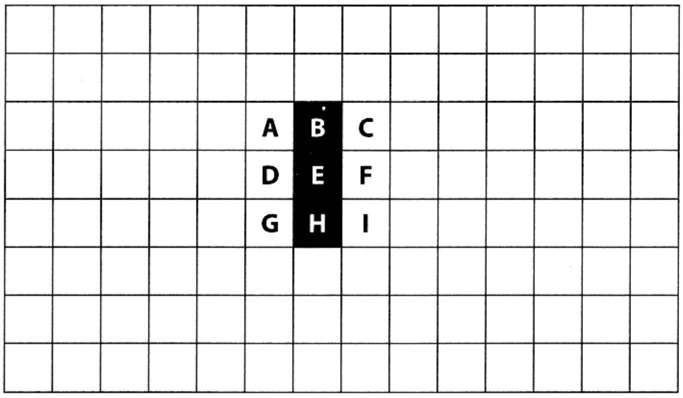


Figure 2.5 Vertical Flasher

Calculate birth cells first. In the configuration shown in Figure 2.5, only cells *d* and *f* have exactly three neighbors ON (dark cells), so they will be the only birth cells in the next generation. Cells *b* and *h* each have only one neighbor ON, so they die in the next generation. Cell *e* has two neighbors ON, so it stays on. So the next instant will look like this:

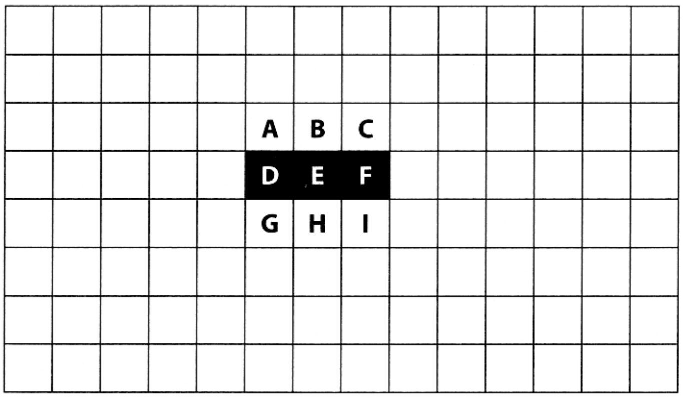


Figure 2.6 Horizontal Flasher

Obviously, the configuration shown in Figure 2.6 will revert back in the next instant, and this little pattern will flip-flop back and forth indefinitely, unless some new ON cells are brought into the picture somehow. It is called a *flasher* or traffic light.

What will happen to the configuration in Figure 2.7?

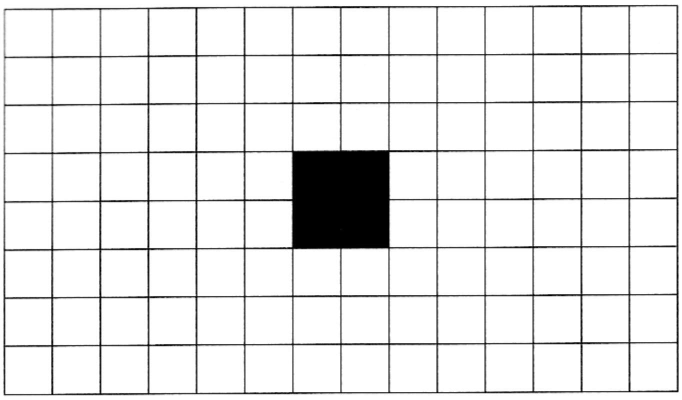


Figure 2.7 Square Still Life

Nothing. Each ON cell has three neighbors ON, so it is reborn just as it is. No OFF cell has three neighbors ON, so no other births happen. This configuration is called a *still life;* there are many different still life configurations that do not change at all over time.

By the scrupulous application of our single law, one can predict with perfect accuracy the next instant of any configuration of ON and OFF cells, and the instant after that, and so forth, so *each Life world is a deterministic two-dimensional Democritean universe.* And to first appearances, it fits our stereotype of determinism perfectly: mechanical, repetitive, ON, OFF, ON, OFF for eternity, with never a surprise, never an opportunity, never an innovation. If you "rewind the tape" and play out the sequel to any configuration again and again, it will always come out exactly the same. Boring! Thank goodness we don't live in a universe like that!

But first appearances can be deceiving, especially when you're standing too close to the novelty. When we step back and consider larger patterns of Life configurations, we are in for some surprises. The flasher has a two-generation period that continues *ad infinitum,* unless some other configuration encroaches. *Encroachment is what makes Life interesting.* Among the periodic configurations are some that swim, amoeba-like, across the plane. The simplest is the *glider,* the five-pixel configuration (Figure 2.8) shown here taking a single stroke to the southeast:

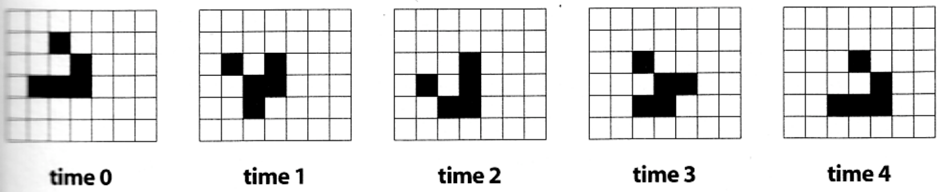


Figure 2.8 Glider

Then there are the eaters, the puffer trains, and space rakes, and a host of other aptly named denizens of the Life world that emerge as recognizable objects at a new level. In one sense, this new level is simply a bird's-eye view of the basic level, looking at large clumps of pixels instead of individual pixels. But, wonderful to say, when we ascend to this level, we arrive at an instance of what I have called the *design level;* it has its own language, a transparent foreshortening of the tedious descriptions one could give at the *physical level.* For instance:

> An eater can eat a glider in four generations. Whatever is being consumed, the basic process is the same. A bridge forms between the eater and its prey. In the next generation, the bridge region dies from overpopulation, taking a bite out of both eater and prey. The eater then repairs itself. The prey usually cannot. If the remainder of the prey dies out as with the glider, the prey is consumed. (Poundstone 1985, p. 38)

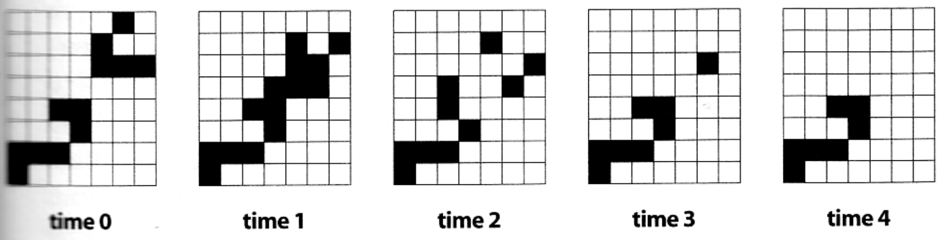


Figure 2.9 Eater Eating a Glider

Door dit [illegible] krijg je een veel rijkere [illegible]
– [illegible] volgende tekst!

Design ~~[illegible]~~ level

Notice that something curious happens to our "ontology"—our catalog of what exists—as we move between levels. At the physical level there is no motion, only ON and OFF, and the only individual things that exist, pixels, are defined by their fixed spatial location, $\{x, y\}$. At the design level we suddenly have the motion of persisting objects; it is one and the same glider (though composed each generation of different pixels) that has moved southeast in Figure 2.8, changing shape as it moves; and there is one less glider in the world after the eater has eaten it in Figure 2.9.

Notice too that whereas at the physical level, there are absolutely no exceptions to the general law, at the design level our generalizations have to be hedged: They require "usually" clauses ("the prey usually cannot" repair itself) or "provided nothing encroaches" clauses. Stray bits of debris from earlier events can "break" or "kill" one of the objects in the ontology at this level. Their *salience as real things* is considerable, but not guaranteed. An element of mortality has been introduced. Whereas the individual atoms—the pixels—flash in and out of existence, ON and OFF, without any possibility of accumulating any changes, any history that could affect their later history, larger constructions can suffer damage, a revision of structure, a loss or gain of material that can make a difference in the future. Larger constructions might also happen to be improved, made *less* vulnerable to later dissolution, by something that happened to them. This historicity is the key. The existence in the Life world of structures that can grow, shrink, twist, break, move . . . and in general *persist over time* opens the floodgates to design opportunities.

Rushing in to explore those opportunities is a worldwide fraternity of Life hackers, hobbyists who delight in testing their ingenuity by devising ever more elaborate arrangements on the Life plane that do interesting things. (If you want to explore the Life world, you can download free a fine, user-friendly implementation Life 32 at the Web site http://psoup.math.wisc.edu/Life32.html. It has a library of interesting configurations, and links to other sites. I require my students to explore the Life world, because I have learned that it renders vivid and robust a set of intuitions that are otherwise absent, and helps them think about these issues. In fact—wonder of wonders—it sometimes leads them to *change their minds* about their philosophical positions. So be careful; it can be addictive fun—and it may lead you to abandon your life-defining hatred of determinism!) To become a Life hacker, you simply ascend to the design level, adopt its ontology, and proceed to predict—sketchily and riskily—the behavior of larger configurations or systems of configurations, *without bothering to compute the physical level.* You can set yourself the task of designing some interesting supersystem out of the "parts" that the design level makes available. It takes only a few minutes to get the hang of it, and who knows what you will be able to concoct. What would you get if you lined up a bunch of still life eaters, and then sprayed them with gliders, for instance? After you've dreamed up your design, you can readily test it; Life 32 will swiftly inform you of any overlooked problems in your design stance predictions. You can get a glimpse of the richness of this design level from a few quotes I once pulled off an excellent Life Web site, http://www.cs.jhu.edu/~callahan/lifepage.html#newresults. The Web site is now defunct, sad to say, and don't bother trying to figure out these comments; they are just meant to illustrate the way Life hackers think and talk.

> The loaf reacts with all the junk the R-pentomino produces as it naturally transforms into a Herschel, and miraculously reappears some time later leaving no debris at all. It is necessary to prevent the first Herschel glider from hitting the fading remnants of the reaction, and there is no room for an ordinary eater. But luckily a tub with tail and a block can be used instead.

> Dave Buckingham found a faster stable reflector that does not use Paul Callahan's special reaction. Instead, the incoming glider hits a boat to make a B-heptomino, which is converted into a Herschel and moved round to restore the boat. A compact form of the 119-step Herschel conduit is needed here, as is a non-standard still life to cope with the 64 64 77 conduit sequence.

These Life hackers are playing God in their simplified two-dimensional universe, trying to design ever more amazing patterns that will propagate themselves, transform themselves, protect themselves, move themselves around on the Life plane—in short, *do things* in the world, instead of merely flashing back and forth or, worse, just persisting unchanged for eternity (unless something encroaches). As the quotations reveal, the problem that confronts anyone who plays God in this world is that no matter how nice your initial pattern is, it always runs the risk of annihilation, of turning into debris, of being eaten by an eater, of vanishing without a trace.

If you want your creations to persist, they have to be protected. Keeping the physics constant (not changing the basic rule of the Life), the only thing you can play with is the initial state description, but you have so many to choose from! A set of Life worlds only 1 million pixels by 1 million pixels already gives you 2 to the trillionth power of different possible universes to explore—the Library of Conway, a Vast but Vanishing branch of the much, much Vaster Library of Democritus. Some of these Life worlds are very, very interesting, but finding them is harder than hunting for a needle in a haystack. The only way to do it, since random search is practically hopeless, is to think of the search as a design problem: How can I *construct* a Life-form that will *do x* or *do y* or *do z?* And once I've designed something that can *do x,* how can I protect my fine *x-er* from harm once I've constructed it? After all, a lot of precious R&D (research and development) went into designing my *x-er.* It would be a shame if it got smashed before it could do its thing.

How can you make things that will last in the sometimes toxic world of Life? This is an objective, non-anthropomorphic problem. The underlying physics is the same for all Life configurations, but some of them, in virtue of nothing but their *shape,* have *powers* that other configurations lack. This is the fundamental fact of the design level. Let the configurations be as un-human, as un-cognitive, as un-agent-like as you can muster. If they last, what is it about them that explains this? A still life is fine until it gets plowed into. Then what happens? Can it restore itself somehow? Something that can nimbly move out of the way might be better, but how can it get any advance warning of incoming missiles? Something that can eat the incoming debris and profit from it might be better yet. But the rule is: Anything that works is fine. Under that rule, what emerges is sometimes strikingly agent-like, but this may be more a function of a bias in our imagination—like seeing animals in the clouds just because we have lots of animal "templates" in our visual memory—than because it is necessary. In any case, we know a set of tricks that work: a set of tricks that is strongly reminiscent of our own biology. The physicist Jorge Wagensberg has recently argued that this resemblance to life as we know it is no accident. In an essay that does not mention Conway Life, he develops definitions of information, uncertainty, and complexity from which he can derive measures of "independence with respect to the uncertainty of the environment" and use these to show that *persistence,* or what he calls "keeping an identity," in a complex environment depends (probabilistically) on various ways of maintaining "independence"—and these ways include such "passive" measures as "simplification" (like seeds and spores), hibernation, isolation (behind shields and shelters), and sheer size, and above all, the "active" measures that require anticipation. "A biota progresses in a particular environment if the new state of the biota is more independent in respect of the uncertainty of that environment" (Wagensberg 2000, p. 504).

A wall is sometimes a good bargain, if it is strong enough so that nothing can smash it. (Nothing? Well, nothing smaller than *G,* the most gigantic projectile we've thrown at it yet.) A wall just sits there and takes a beating, not *doing* anything. A mobile protector, on the other hand, must either move in a fixed trajectory, like a sentry marching around the perimeter of a camp; or in a random trajectory, like the swimming-pool vacuum-sweepers that prowl at random, cleaning the walls; or in a guided trajectory that depends on its obtaining some information about the environment through which it moves. A wall that can repair itself is another interesting possibility, but much harder to design than a static wall. These fancier designs, the designs that can take steps to improve their chances, can get quite expensive, since they depend on reacting to information about their circumstances. Their *immediate* surroundings (the eight neighbors around each pixel) are more than informative—they are utterly determining; it is "too late to do anything" about a collision that has begun. If you want your creation to be able to *avoid* some impending harm, it is going to have to be designed either to do the right thing "automatically" (the thing it always does) or to have some way of anticipating it, so that it can be (designed to be) guided by some harbinger or other down a better path.

This is the birth of avoidance; this is the birth of prevention, protection, guidance, enhancement, and all the other fancier, more expensive sorts of *action.* And right at the moment of birth, we can discern the key distinction we will need later on: Some kinds of harms can, in principle, be avoided, and some kinds of harms are unavoidable, or *inevitable,* as we say. Advance warning is the key to avoidance, and this is strictly limited in the Life world by the "speed of light," which is (for all practical purposes) the speed at which simple gliders can swim [illegible], [illegible] [illegible]

diagonally across the plane. Gliders, in other words, could be the *photons,* the light particles, in the set of Life universes, and *reacting-to-a-glider* could be a way of turning a mere collision or encroachment into an *informing,* a simplest case of noticing or discriminating. We can see why it is that calamities that arrive at the speed of light must "blindside" any creations they encounter; they are truly inevitable. Slower-moving problems can, in principle, be predicted by any Life-form that can extract guidance from the incoming rain of gliders (or other, slower sources of information) and adjust itself appropriately. It may pick up information about what to expect from other things it encounters, but only if *there is* information in those patterns that is predictive of patterns elsewhere, or at other times. In a totally chaotic, unpredictable environment, there is no hope of avoidance except sheer blind luck.

Notice that I have been intermixing two distinct information-gathering processes in this discussion, which now need to be more clearly separated. First, there is the activity of our hacker Gods, who are free to cast their eyes and minds over huge manifolds of possible Life worlds, trying to figure out what will tend to work, what will be robust and what will be fragile. For the time being, we are supposing that they are truly God-like in their "miraculous" interactions with the Life world—they are not bound by the slow speed of glider-light; they can intervene, reaching in and tweaking the design of a creation whenever they like, stopping the Life world in mid-collision, undoing the harm and going back to the drawing board to create a new design. Wherever *they* can foresee a source of difficulty they can set themselves the task of designing a way of countering it. Their creations will be the unwitting, foresightless beneficiaries of the foresight of the hacker Gods, who have designed them to thrive in just such circumstances. Hacker Gods have their limitations, however, and will economize whenever they can. For instance, they might interest themselves in such questions as: What is the *smallest* Life-form that can protect itself from harm x or harm y, under conditions z (but not under conditions w)? After all, gathering information and putting it to use is a costly, time-consuming process, even for a hacker God. The second possibility is the prospect of the hacker Gods designing configurations that *do their own* information gathering, locally, bound by the physics of the world they inhabit. Expect that any finite creation that uses information will be thrifty, keeping only what it (probably) needs or (probably) can use, given the vicissitudes in its neighborhood. After all, the hacker God who designs it wants to make it robust enough to fend for itself not in all possible Life worlds but only in any of the set of Life worlds it has some probability of encountering. Such a creation will, at best, be in a position to *act as if it knew* it was living in a particular *sort* of neighborhood, fending off a particular *sort* of harm or securing a particular *sort* of benefit, instead of acting as if it knew exactly which Life universe it inhabited.

Speaking of these smallest avoiders as if they "knew" anything at all involves a large dose of poetic license, since they would be about as close to clueless as you can imagine—they are much simpler than a real-world bacterium, for instance—but it is still a useful way of keeping track of the design work that has gone into them, giving them capabilities to *do things* that any randomly assembled clumps of pixels of about the same size would lack. (Of course, "in principle"—as philosophers love to say—a Cosmic Accident could produce exactly the same constellation of pixels with exactly the same capabilities, but this is an utterly negligible possibility, beyond improbability. Only expensively designed things can do things in the interesting sense.)

Enriching the design stance by speaking of configurations as if they "know" or "believe" something, and "want" to accomplish some end ~~or other~~ is moving up from the simple *design stance* to what I call the *intentional stance.* Our simplest doers have been reconceptualized as *rational agents* or ~~*intentional*~~ *systems,* and this permits us to think about them at a still higher level of abstraction, ignoring the details of just how they manage to store the information they "believe" and how they manage to "figure out" what to do, based on what they "believe" and "want." We just assume that however they do it, they do it rationally—they draw the right conclusions about what to do next from the information they have, given what they want. It makes life blessedly easier for the high-level designer, just the way it makes life easier for us all to conceptualize our friends and neighbors (and enemies) as intentional systems.

And subtle lies.

We can move back and forth between the hacker God perspective and the "perspective" of the hacker God's creations. Hacker Gods have their reasons, good or bad, for designing their creations the way they did. The creations themselves can be clueless about these reasons, but they *are* the reasons those features exist, and if the creations persist, it will be thanks to those features. If, beyond that, the creations have been designed to gather information to use in action guidance, the situation becomes more complicated. The simplest possibility is that a hacker God has designed a repertoire of reaction-tricks that tend to work well in the environments encountered, analogous to the IRMs (Innate Releasing Mechanisms) and FAPs (Fixed Action Patterns) that ethologists have identified in many animals. Gary Drescher (1991) calls this architecture a *situation-action machine* and contrasts it with the more expensive, more complex *choice machine,* in which the individual creation generates its *own* reasons for doing *x* or *y,* by anticipating probable outcomes of various candidate actions and evaluating them in terms of the goals it also represents (since these goals can change over time, in response to new information gathered). If we ask "at what point" the designer's reasons become the designed agent's reasons, we may find that there is a seamless blend of intermediate steps, with more and more of the design work off-loaded from designer to designed agent. One of the beauties of the intentional stance is that it allows us to see clearly this shift in the distribution of "cognitive labor" between the originating design process and the efforts of the thing designed.

All this fanciful talk about configurations of Life-pixels as rational agents may strike you as outrageous overstatement, a blatant attempt by me to pull the wool over your eyes. It's time for a sanity check: Just how much, in principle, can a designed constellation of Life-pixels *do,* given glider-discrimination and its kin as the "molecules" of the design level, the fundamental building blocks of higher-level Life-forms? This is the question that inspired Conway to create the Game of Life in the first place, and the answer he and his students came up with is staggering. They were able to prove that there are Life worlds—they sketched one of them—within which there is a Universal Turing Machine, a two-dimensional computer that in principle can compute any computable function. It was far from easy, but they showed how they could "build" a working computer out of simpler Life-forms. Glider streams can provide the input-output "tape," for instance, and the tape-reader can be some huge assembly of eaters, gliders, and other bits and pieces. What this means is mind-boggling: Any program that can run on any computer could, in principle, run in the Life world on one of these Universal Turing Machines. A version of Lotus 1-2-3 could exist in the Life world; so could Tetris or any other video game. So the information-handling ability of gigantic Life-forms is equivalent to the information-handling ability of our real three-dimensional computers. Any competence you can "put on a chip" and embed in a 3-D contraption can be perfectly mimicked by a similarly embedded Life constellation in a still larger Life-form in two dimensions. We know it exists in principle. All you have to do is find it—that is to say, all you have to do is design it.

Can We Get the *Deus ex Machina?*

Now it is time to ask whether we might eliminate the miracle-working hacker Gods from our picture, replacing their ingenious design efforts with evolution *within the Life world itself.* Is there any Life world, of any size, in which the sorts of human R&D just described are carried on by natural selection? More precisely, are there configurations of the Life world such that, if you started the world in one of them, it would eventually *do all the work* of the hacker Gods, gradually discovering and propagating better and better avoiders? This move, to an evolutionary perspective, carries with it a family of ideas that can *seem* paradoxical or self-contradictory from our everyday perspective, and it takes some strenuous exercise of thought to get comfortable with the transitions between the two perspectives. One of Darwin's earliest critics saw what was coming and could scarcely contain his outrage:

> In the theory with which we have to deal, Absolute Ignorance is the artificer; so that we may enunciate as the fundamental principle of the whole system, that, IN ORDER TO MAKE A PERFECT AND BEAUTIFUL MACHINE, IT IS NOT REQUISITE TO KNOW HOW TO MAKE IT. This proposition will be found, on careful examination, to express, in condensed form, the essential purport of the Theory, and to express in a few words all Mr. Darwin's meaning; who, by a strange inversion of reasoning, seems to think Absolute Ignorance fully qualified to take the place of Absolute Wisdom in all the achievements of creative skill. (MacKenzie 1868, p. 217)

MacKenzie identifies what he calls a "strange inversion of reasoning," and he is right on all counts. The Darwinian revolution is indeed an inversion of everyday reasoning in several regards, and it is, for that reason, strange: a *foreign* language, full of traps for the unwary, even after considerable practice, all the more so because there are so many terms that are what linguists call *false friends*—terms that seem to be cognates or synonyms of terms from your mother tongue but differ in treacherous ways. One man's *Gift* is another man's poison; one man's *chair* is another man's flesh. (Hint: Look in German–English and French–English dictionaries.) In the case of the Darwinian perspective, the problem of false friends is exacerbated because the terms that invite confusion are, in fact, closely related and relevant to each other—but just not quite the same. When we invert the top-down perspective of tradition and look at creation from the bottom up, we see intelligence arising from "intelligence," sight being created by a "blind watchmaker," choice emerging from "choice," deliberate voting from mindless "voting," and so on. There will be lots of scare-quotes in the explanations to come. We will see—talk about paradox!—how a whole can be more *free* than its parts.

So the straightforward technical question of whether an evolutionary process could replace the effort of the hacker Gods in the Life world has some far-reaching implications. Moreover, the answer has some curious twists in it. In such a Life world, there would have to be self-reproducing entities, and we do know that they can exist, since Conway and his students embedded their Universal Turing Machine in just such a contraption. They devised the Game of Life, in fact, in order to explore John von Neumann's pioneering thought-experiments about self-reproducing automata, and they succeeded in designing a self-reproducing structure that would populate the empty plane with ever more copies of itself, rather like bacteria in a petri dish, each one containing a Universal Turing Machine. What does this machine look like? Poundstone calculates that the whole construction would be on the order of 10^{13} pixels.

> Displaying a 10^{13}-pixel pattern would require a video screen about 3 million pixels across at least. Assume the pixels are 1 millimeter square (which is very high resolution[3] by the standards of home computers). Then the screen would have to be 3 kilometers (about two miles) across. It would have an area about six times that of Monaco.

[3] When Poundstone was writing (1985) this was very high, but today it would be low. The pixels on my laptop are almost four times smaller, so the whole screen at that resolution would be somewhat less than 1 kilometer across. Still a big screen.

> Perspective would shrink the pixels of a self-reproducing pattern to invisibility. If you got far enough away from the screen so that the entire pattern was comfortably in view, the pixels (and even the gliders, eaters and guns) would be too tiny to make out. A self-reproducing pattern would be a hazy glow, like a galaxy. (Poundstone 1985, pp. 227–28)

In other words, by the time you have built up enough pieces into something that can reproduce itself (in a two-dimensional world) it is roughly as much larger than its smallest bits as an organism is larger than its atoms. That shouldn't surprise us. You probably can't do it with anything much less complicated, though this has not been strictly proven.

But self-reproduction is not enough by itself. We also need mutation, and adding this is going to be surprisingly expensive. In his book *Le Ton Beau de Marot* (1997), Douglas Hofstadter draws attention to the role of what he calls *spontaneous intrusions* into any creative process, whether it is achieved by the exertions of a human artist or inventor or scientist, or by natural selection. Every increment of design in the universe begins with a moment of serendipity, the undesigned intersection of two trajectories that yield something that turns out, retrospectively, to be more than a mere collision. We have seen how collision-detection is a fundamental capacity that can be made available to Life-forms, and indeed how collision is a major problem facing all Life hackers, but *how much* collision can we afford in our Life worlds? This turns out to be a serious problem when we set out to add mutation to the self-replication powers of Life configurations.

Computer simulations of evolution abound, and show us the power of natural selection to create strikingly effective novelties in remarkably short periods of time in one virtual world or another, but they are always, perforce, orders of magnitude simpler than the real world, because they are always much more *quiet*. What happens in a virtual world is only what the designer specifies to happen. Consider a typical difference between virtual worlds and real worlds: If you set out to make a real hotel, you have to put a lot of time, energy, and materials into arranging matters so that the people in adjacent rooms can't overhear each other; if you set out to make a virtual hotel, you get that insulation for free. In a virtual hotel, if you want the people in adjacent rooms to be able to overhear each other, you have to add that

Our senses are the result of millions of years of evolution

[illegible]

capacity. You have to add *non*-insulation. You also have to add shadows, aromas, vibration, dirt, footprints, and wear-and-tear. All these non-functional features come for free in the real, concrete world—and they play a crucial role in evolution. The open-endedness of evolution by natural selection depends on the extraordinary richness of the real world, which constantly provides new *undesigned* elements that can be serendipitously harnessed, once in a blue moon, into new design elements. To take the simplest case, can there be enough interference in the world to produce an appropriate number of mutations without, in the process, simply breaking the whole reproductive system? The reproductive system of Conway's Universal Turing Machine was noise-free, making perfect copies every time. There was no provision for mutation at all, no matter how many copies of itself it produced. Could a still larger, more ambitious self-reproducing automaton be designed that could allow for the occasional unblocked glider to arrive, like a cosmic ray, and produce a mutation in the genetic code being copied? Can a two-dimensional Life world be *noisy* enough to support open-ended evolution, while still *quiet* enough to permit the designer parts to do their good work unassailed? Nobody knows.

It is an interesting fact that by the time you specify Life worlds that are complex enough to be candidates for such capacities, they are much too complex to run in simulation. Noise and debris can always be added to a model, but it has the effect of squandering the efficiency that makes computers such great tools in the first place. So there is a sort of *homeostasis* or self-limiting equilibrium here. The very simplicity, the *over*simplicity, of our models can prevent them from modeling the things we are most interested in, such as creativity, either by a human artist or by natural selection itself, since in both cases that creativity feeds on the very complexity of the real world. There is nothing mysterious or even puzzling about this, no whiff of strange new complexity-forces or unpredictable-in-principle emergence; it is simply an everyday, practical fact that computer modeling of creativity confronts diminishing returns because in order to make your model more open-ended, you have to make your model more concrete. It has to model more and more of the incidental collisions that impinge on things in the real world. Encroachment is, indeed, what makes life interesting.

So it is unlikely that we can ever prove *by construction* that somewhere in the Vast reaches of the Life plane, there are configurations that mimic the full open-endedness of natural selection. Still, we can construct the parts piecemeal, providing the important existence proofs we need. Yes, there exist such configurations as Universal Turing Machines, and self-protective persisters, and reproducers, and limited evolutionary processes. Formal arguments such as Wagensberg's (and Conway's and Turing's) take us beyond construction to fill in the gaps of impracticality, so we can say with some confidence that our toy deterministic world is one in which all the necessary ingredients exist for the evolution of . . . *avoiders!* This proposition is what we need to break the back of the cognitive illusion that yokes determinism with inevitability. But before turning to this, it will help to return from toyland to reality, to see what we know about the evolution of avoidance on our planet.

From Slow-motion Avoidance to Star Wars

We know that in the early days—the first few billion years—of life on this planet, self-protective designs emerged, thanks to the slow and non-miraculous process of natural selection. It took on the order of 1 billion years of replication for the simplest life-forms to work out the best designs—still susceptible to revision today, of course—for the basic processes of replication. Along the way there was much *avoidance* and *prevention,* but at a pace much too slow to appreciate unless we artificially speed it up in imagination. For instance, the incessantly exploratory process of natural selection occasionally spewed forth counterproductive DNA sequences, parasitic genes or *transposons,* that hitched a free ride on the genomes of early life-forms, contributing nothing to the well-being of those life-forms but just cluttering up their genomes with extra copies (and copies of copies of copies) of themselves. These parasites created a problem; something had to be *done.* And in due course the incessantly exploratory process of natural selection, by a more or less exhaustive search, "found" a solution (or two, or more): designs for structures in the valuable, constructive parts of genomes that *prevented* the excessive flourishing of these parasites, *counteracting* their *actions* with *reactions,* and so forth. The parasitic genes reacted in turn to this new development by a counterthrust of their own, developed over many hundreds or thousands or millions of gen-

In J. Tomberlin, ed., Philosophical Perspectives: Mind, Causation, and World, Vol. 11 (Malden, MA: Blackwell, 1997), pp. 375-399.

Weak Emergence[*]

Mark A. Bedau
Reed College, 3203 SE Woodstock Blvd., Portland OR 97202, USA
Voice: (503) 771-1112, ext. 7337; Fax: (503) 777-7769
Email: mab@reed.edu; Web: www.reed.edu/~mab

An innocent form of emergence—what I call "weak emergence"—is now a commonplace in a thriving interdisciplinary nexus of scientific activity—sometimes called the "sciences of complexity"—that include connectionist modelling, non-linear dynamics (popularly known as "chaos" theory), and artificial life.[1] After defining it, illustrating it in two contexts, and reviewing the available evidence, I conclude that the scientific and philosophical prospects for weak emergence are bright.

Emergence is a tantalizing topic because examples of apparent emergent phenomena abound. Some involve inanimate matter; e.g., a tornado is a self-organizing entity caught up in a global pattern of behavior

[*] Many thanks for comments on earlier drafts and discussion on these topics to Hugo Bedau, John Collier, Seth Crook, Dick Grandy, Mark Hinchliff, Terry Horgan, Stephen Leeds, Norman Packard, Teresa Robertson, Bill Uzgalis, and to audiences at the Third Workshop on Artificial Life (in Santa Fe during summer 1993), at Oregon State University (spring 1994), and at the Northwest Philosophy Conference (fall 1994). This research was supported with the help of an external fellowship from the Oregon Center for the Humanities and a research grant from the Oregon Council for the Humanities.

[1] Accessible introductions to the study of chaos, with references to more technical treatments, include Crutchfield et al. (1986), Gleick (1987), and Kellert (1993). The bible of connectionism is Rumelhart and McClelland (1986); discussions for philosophers, and references to the technical literature, can be found in Bechtel and Abrahamsen (1990), Horgan and Tienson (1991), and Ramsey, Stich, and Rumelhart (1991). The locus of much recent activity in the "sciences of complexity" is the Santa Fe Institute, a private, independent multidisciplinary research center. Semi-popular introductions to some of the research centered at the Santa Fe Institute include Levy (1992), Lewin (1992), and Waldrop (1992). A representative range of technical work can be found in the series Santa Fe Institute Studies in the Sciences of Complexity, published by Addison-Wesley; e.g., Langton (1989) and Langton et al. (1992).

that seems to be autonomous with respect to the massive aggregation of air and water molecules which constitute it. Another source of examples is the mind; our mental life consist of an autonomous, coherent flow of mental states (beliefs, desires, etc.) that presumably somehow ultimately arise out of the swarm of biochemical activity among our brain's neurons. Life is a third rich source of apparent emergence. For example, the hierarchy of life embraces ecosystems composed of organisms, which are composed of organs, which are composed of cells, which are composed of molecules, but each level in this hierarchy exhibits behavior that seems autonomous with respect to the behavior found at the level below.

These examples highlight two admittedly vague but nevertheless useful hallmarks of emergent phenomena:

(1) Emergent phenomena are somehow *constituted by*, and *generated from*, underlying processes.

(2) Emergent phenomena are somehow *autonomous* from underlying processes.

If we place these hallmarks against a backdrop of abundant apparently emergent phenomena, it is clear why emergence is a perennial philosophical puzzle. At worst, the two hallmarks seem to make emergent phenomena flat-out inconsistent. At best, they still raise the specter of illegitimately getting something from nothing.

So, aside from precisely defining what emergence is, any philosophical defense of emergence should aim to explain—ideally, explain away—its apparently illegitimate metaphysics. Another important goal should be to show that emergence is consistent with reasonable forms of materialism. But perhaps the most important goal should be to show that emergent properties are useful in empirical science, especially in accounts of those phenomena like life and mind that have always seemed to involve emergence. A defense of emergence will be secure only if emergence is more than merely a philosophical curiosity; it must be shown to be a central and constructive player in our understanding of the natural world.

I will argue that *weak emergence* (defined below) meets these three goals: it is metaphysically innocent, consistent with materialism, and scientifically useful, especially in the sciences of complexity that deal with life and mind. But first I will briefly illustrate the scientific irrelevance characteristic of stronger, more traditional conceptions of emergence.

Problems with Strong Emergence.

To glimpse the problems with stronger forms of emergence, consider the conception of emergence defended by Timothy O'Conner (1994). O'Conner's clearly articulated and carefully defended account falls squarely within the broad view of emergence that has dominated philosophy this century. His definition[2] is as follows: Property *P* is an emergent property of a (mereologically-complex) object *O* iff *P* supervenes on properties of the parts of *O*, *P* is not had by any of the object's parts, *P* is distinct from any structural property of *O*, and *P* has a direct ("downward") determinative influence on the pattern of behavior involving *O*'s parts.

The pivotal feature of this definition, to my mind, is the strong form of downward causation involved. O'Conner (pp. 97f) explains that he wants

> to capture a very strong sense in which an emergent's causal influence is irreducible to that of the micro-properties on which it supervenes; it bears its influence in a direct 'downward' fashion, in contrast to the operation of a simple structural macro-property, whose causal influence occurs via the activity of the micro-properties which constitute it.

I call O'Conner's notion "strong" emergence to contrast it with the weaker form of emergence, defended below, that involves downward causation only in the weak form created by the activity of the micro-properties that constitute structural macro-properties.

It is worth noting that strong emergence captures the two hallmarks of emergence. Since emergent phenomena supervene on underlying processes, in this sense the underlying processes constitute and generate the emergent phenomena. And emergent phenomena are autonomous from the underlying processes since they exert an irreducible form of downward causal influence. Nevertheless, strong emergence has a number of failings, all of which can be traced to strong downward causation.

Although strong emergence is logically possible, it is uncomfortably like magic. How does an irreducible but supervenient downward causal power arise, since by definition it cannot be due to the aggregation of the micro-level potentialities? Such causal powers would be quite unlike anything within our scientific ken. This not only indicates how they will discomfort reasonable forms of materialism. Their mysteriousness will only heighten the traditional worry that emergence entails illegitimately getting something from nothing.

But the most disappointing aspect of strong emergence is its apparent scientific irrelevance. O'Conner finds evidence that strong emergence is useful in the empirical sciences in "the recent proposals of macro-determinitive influence on lower-level sub-structure by Polanyi and Sperry with respect to embryonic cells and consciousness, respectively" (p. 99). But these references to Polanyi and Sperry provide little evidence of the empirical viability of strong emergence unless they refer to a flourishing scientific research program. Our doubts about this should be raised when we note that in the recent philosophical literature on emergence (including O'Conner) all

[2] O'Conner adapts Kim's notion of "strong supervenience" (Kim 1990) and Armstrong's definition of structural property (Armstrong 1978).

citations are to the *same* Polanyi and Sperry papers, which generally date back twenty five years. This is not the trail left by a thriving research program. Strong emergence is perhaps *compatible* with current scientific knowledge. But if Sperry and Polanyi are the best defense of strong emergence's empirical usefulness, then its scientific credentials are very weak. We should avoid proliferating mysteries beyond necessity. To judge from the available evidence, strong emergence is one mystery which we don't need.

Weak emergence contrasts with strong emergence in this respect; science apparently *does* need weak emergence. Fortunately, there are no mysteries like irreducible downward causation in weak emergence, to which we will now turn.

Definition of Weak Emergence.

Weak emergence applies in contexts in which there is a system, call it *S*, composed out of "micro-level" parts; the number and identity of these parts might change over time. *S* has various "macro-level" states (macrostates) and various "micro-level" states (microstates). *S*'s microstates are the intrinsic states of its parts, and its macrostates are structural properties constituted wholly out of its microstates.[3] Interesting macrostates typically average over microstates and so compresses microstate information. Further, there is a microdynamic, call it *D*, which governs the time evolution of *S*'s microstates. Usually the microstate of a given part of the system at a given time is a result of the microstates of "nearby" parts of the system at preceding times; in this sense, *D* is "local". Given these assumptions, I define weak emergence as follows:

> Macrostate *P* of *S* with microdynamic *D* is *weakly emergent* iff *P* can be derived from *D* and *S*'s external conditions but only by simulation.[4]

[3] The macrostate *P* might fall into a variety of categories. It might be a property of *S*, possibly one involving various other macrostates of *S*; it might be some phenomenon concerning *S*, possibly involving a variety of *S*'s other macrostates; it might be a pattern of *S*'s behavior, possibly including other macrostates of *S*. There are also more complicated cases, in which the macrostate is "supple" or "fluid", and the structural definition of the macrostate might be infinitely long. This latter issue is developed in Bedau (1995*c*).

[4] This definition is explicitly restricted to a given macrostate of a given system with a given microdynamic. This is the *core* or *focal* notion in a family of related notions of weak emergence, all others of which would be defined by reference to the core notion and would crucially invoke underivability without simulation. For example, one can speak of a weak emergent *law* when it is a law that a given macrostate of a given system with a given microdynamic is weakly emergent from a range of initial conditions;

Conditions affecting the system's microstates are "external" if they are "outside" the system. If *D* is deterministic and the system is closed, then there is just one external condition: the system's initial condition. Every subsequent microstate of the system is determined by elements inside the system (the microdynamic *D* and the system's microstates). If the system is open, then another kind of "external" condition is the contingencies of the flux of parts and states through *S*. If the microdynamic is nondeterministic, then each accidental effect is an "external" condition. With external conditions understood in this fashion, it is coherent to speak of macrostates being "derivable" from external conditions even in nondeterministic systems.

Although perhaps unfamiliar, the idea of a macrostate being derived "by simulation" is straightforward and natural. Given a system's initial condition and the sequence of all other external conditions, the system's microdynamic completely determines each successive microstate of the system. To simulate the system one iterates its microdynamic, given a contingent stream of external conditions as input. Since the macrostate P is a structural property constituted out of the system's microstates, the external conditions and the microdynamic completely determine whether *P* materializes at any stage in the simulation. By simulating the system in this way one can derive from the microdynamic plus the external conditions whether P obtains at any given time after the initial condition. What distinguishes a weakly emergent macrostate is that this sort of simulation is required to derive the macrostate's behavior from the system's microdynamic. Crutchfield et al. (1986, p. 49) put the essential point especially clearly: the algorithmic effort for determining the systems behavior is roughly proportional to how far into the future the system's behavior is derived. It is obvious that the algorithmic effort required for a simulation is proportional to how far into the future the simulation goes.

this law is underivable without simulations across many initial conditions. Similarly, one can speak of a weak emergent *pattern* involving a range of suitably related macrostates, microdynamics, or systems, but I will not attempt here to define weak emergence in this whole family of contexts. The guiding strategy behind these definitional extensions is reasonably clear. The range of new contexts for weak emergence is limited only by our imagination.

It is worth at least mentioning that the notion of underivability without simulation provides another dimension along which notions of weak emergence can vary. There is a range of more or less stringent conditions. For example, consider a macrostate that in principle is derivable without simulation, yet the derivation uses vastly more resources (e.g., "steps") than any human could grasp; or consider a macrostate that is derivable (only) by simulation but the simulation is infinitely long. I will not elaborate on this dimension here. The paradigm of weak emergence involves underivability except by finite simulation.

The need for simulations in the study of low-dimensional chaos has been emphasized before (see, e.g., Crutchfield et al. 1986, Stone 1989, Kellert 1993). Weak emergence has a special source in this kind of chaos: exponential divergence of trajectories, also known as sensitive dependence on initial conditions or "the butterfly effect". This particular mechanism does not underlie all forms of weak emergence, though. On the contrary, weak emergence seems to rampant in all complex systems, regardless of whether they have the underlying mechanisms that produce chaos. In fact, some include weak emergence as part of the definition of what it is to be a complex adaptive system in general (Holland 1992). Indeed, it is the ubiquity of weak emergence in complex systems that makes weak emergence especially interesting.

Derivations that depend on simulations have certain characteristic limitations. First, they are massively contingent, awash with accidental information about the system's components and external conditions. The derivations can be too detailed and unstructured for anyone to be able to "survey" or understand how they work. The derivations also can obscure simpler macro-level explanations of the same macrostates that apply across systems with different external conditions and different microdynamics. But none of this detracts from the fact that all of the system's macrostates can be derived from its microdynamic and external conditions with a simulation.

The modal terms in this definition are metaphysical, not epistemological. For P to be weakly emergent, what matters is that there is a derivation of P from D and S's external conditions and any such derivation is a simulation. It does not matter whether anyone has discovered such a derivation or even suspects that it exists. If P is a weakly emergent, it is constituted by, and generated from, the system's underlying microdynamic, whether or not we know anything about this. Our need to use a simulation is due neither to the current contingent state of our knowledge nor to some specifically human limitation or frailty. Although a Laplacian supercalculator would have a decisive advantage over us in simulation speed, she would still need to simulate. Underivability without simulation is a purely formal notion concerning the existence and nonexistence of certain kinds of derivations of macrostates from a system's underlying dynamic.

Weak Emergence in the Game of Life.

A good way to grasp the concept of weak emergence is through examples. One of the simplest source of examples is the Game of Life devised more than a generation ago by the Cambridge mathematician John Conway and popularized by Martin Gardner.[5] This "game" is "played" on a two-dimensional rectangular grid of cells, such as a checker board. Time is

[5] See Berlekamp et al. (1982) and Gardner (1983). An excellent introduction to the intellectual delights of Conway's Game of Life is Poundstone (1985).

discrete. A cell's state at a given time is determined by the states of its eight neighboring cells at the preceding moment, according to the birth-death rule: A dead cell becomes alive iff 3 neighbors were just alive, and a living cell dies iff fewer than 2 or more than 3 neighbors were just alive. (Living cells with fewer than two living neighbors die of "loneliness", those with more than three living neighbors die of "overcrowding", and a dead cell becomes populated by a living cell if it has the three living neighbors needed to "breed" a new living cell.) Although Conway's Game of Life does not represent the state of the art of scientific attempts to understand complex systems, it is a well-known and exquisitely simple illustration of many of the principles of complexity science, including weak emergence, and it illustrates a class of systems—so called "cellular automata"—that are one central paradigm for how to understand complexity in general (see, e.g., Wolfram 1994).

One can easily calculate the time evolution of certain simple Life configurations. Some remain unchanging forever (so-called "still lifes"), others oscillate indefinitely (so-called "blinkers"), still others continue to change and grow indefinitely. Figure 1 shows seven time steps in the history of six small initial configurations of living cells; some are still lifes, others are blinkers. Examining the behavior of these initial configurations allows one to derive their exact behavior indefinitely far into the future. More complex patterns can also be produced by the simple birth-death rule governing individual cells. One simple and striking example—dubbed the "glider", shown as (f) in Figure 1—is a pattern of five living cells that cycles through four phases, in the processes moving one cell diagonally across the Life field every four time steps. Some other notable patterns are "glider guns"—configuration that periodically emit a new glider—and "eaters"—configurations that destroy any gliders that collide with them. Clusters of glider guns and eaters can function in concert just like AND, OR, NOT, and other logic gates, and these gates can be connected into complicated switching circuits. In fact, Conway proved (Berlekamp et al. 1982) that these gates can even be cunningly arranged so that they constitute a universal Turing machine, and hence are able to compute literally every possible algorithm, or, as Poundstone vividly puts it, to "model every precisely definable aspect of the real world" (Poundstone 1985, p. 25).

There is no question that every event and pattern of activity found in Life, no matter how extended in space and time and no matter how complicated, is generated from the system's microdynamic—the simple birth-death rule. Every event and process that happens at any level in a Life world can be deterministically derived from the world's initial configuration of states and the birth-death rule. It follows that a structural macrostate in Life will be weakly emergent if deriving its behavior requires simulation. Life contains a vast number of macrostates that fill this bill. Some are not especially interesting; others are fascinating. Here are two examples.

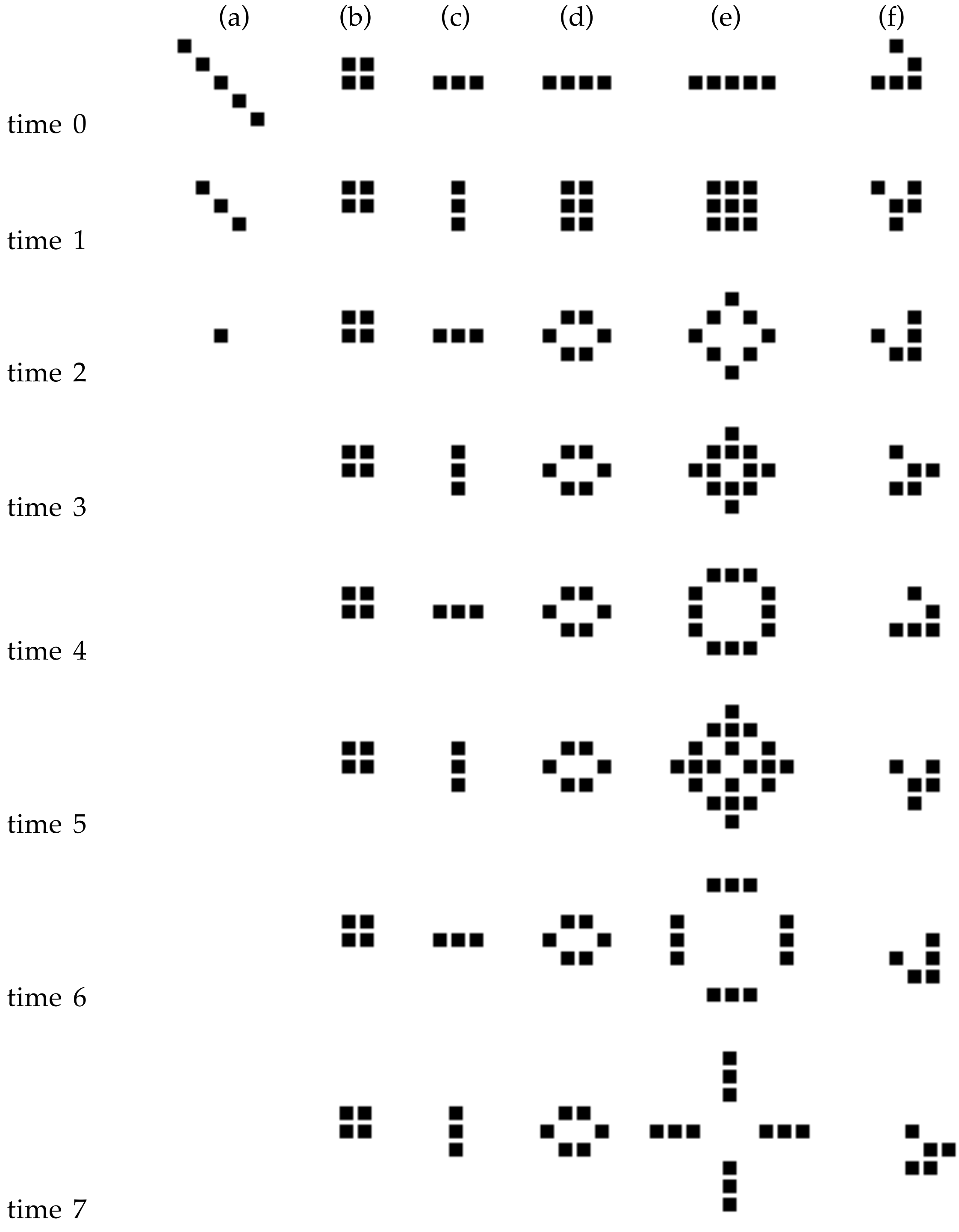


Figure 1. Seven time steps in the evolution of some simple configurations in the Game of Life. Configuration (a) is a "fuse" burning at both ends; after two time steps it is entirely consumed and no life remains. Configuration (b), a still life called the "block", never changes. Configuration (c), a "traffic light", is a blinker with period two. Configuration (d) evolves after two time steps into the "beehive," another still life. Configuration (e) evolves after five time steps into a period two blinker consisting of four traffic lights. Configuration (f) is a glider, a period four pattern that moves diagonally one cell per period.

R pentomino growth. The R pentomino is a wildly unstable five-cell edge-connected pattern. Figure 2 shows the first seven time steps in the evolution of the R pentomino; Figure 3 shows the pattern at time step 100 (above) and time step 150 (below). Listen to part of Poundstone's description (1985, p. 33) of what the R pentomino produces: "One configuration leads to another and another and another, each different from all of its predecessors. On a high-speed computer display, the R pentomino roils furiously. It expands, scattering debris over the Life plane and ejecting gliders."

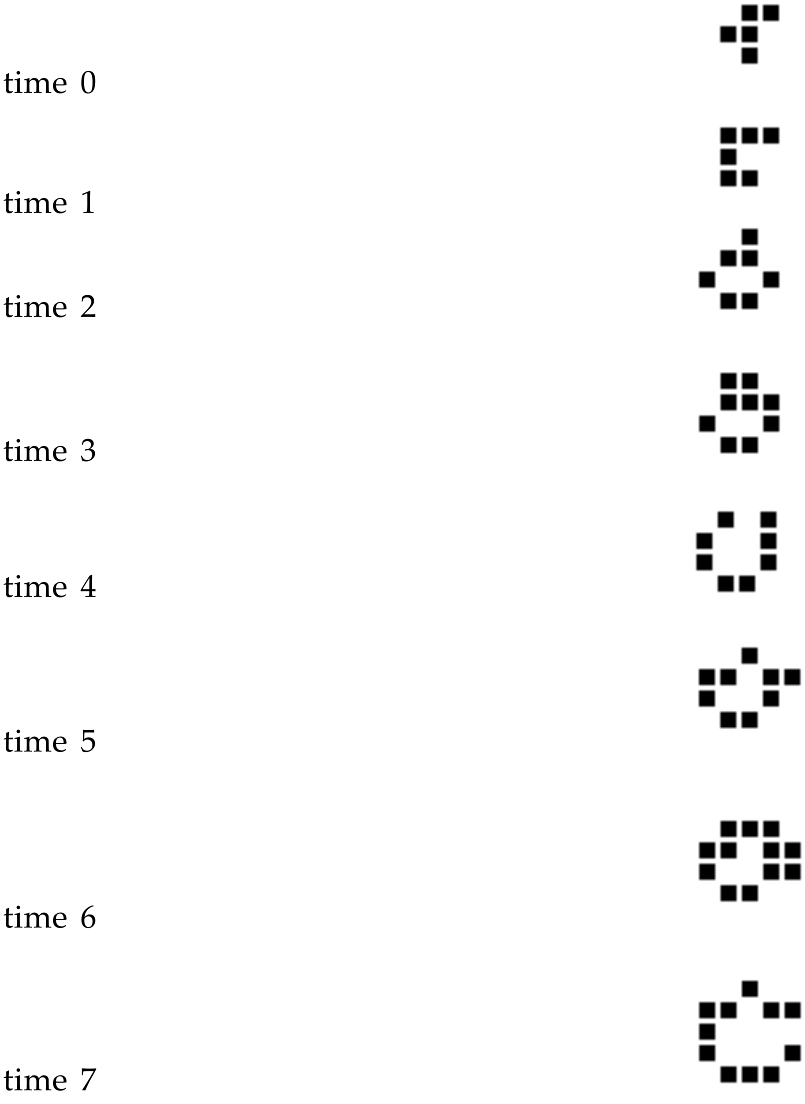


Figure 2. The first seven time steps in the evolution of the R pentomino (the figure at time 0), showing slow and irregular growth.

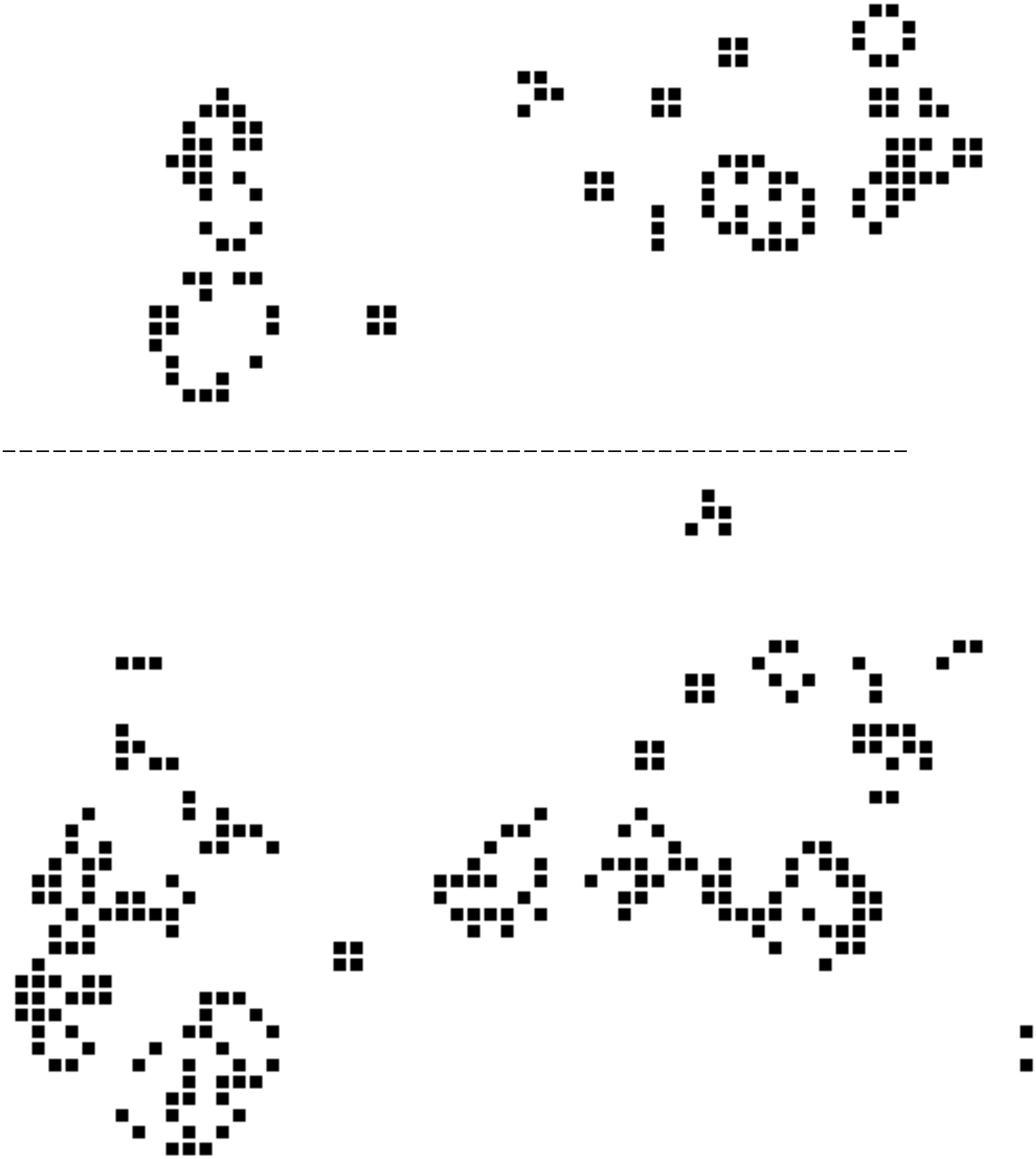


Figure 3. Above: The R pentomino after 100 timesteps. The configuration contains five blocks, a traffic light, a glider, and some unstable clusters of cells. Below: The R pentomino after 150 timesteps. The configuration now includes three blocks, a traffic light, two gliders, and some unstable clusters of cells. The pattern continues to grow steadily but irregularly.

Indefinite growth (i.e., increase in number of living cells) is a structural macrostate constituted by the cells in Life.[6] Does the R pentomino (on an infinite Life grid) grow indefinitely? Some Life configurations do grow forever, such as glider guns, which continually spawn five-cell gliders that glide off into the indefinite distance. So, if the R pentomino continually ejects gliders that remain undisturbed as they travel into the infinite distance, for example, then it would grow forever. But does it? There is no simple way to answer this question. As far as anyone knows, all we can do is let Life "play" itself out when given the R pentomino as initial condition, i.e., observe the R pentomino's behavior. As it happens (Poundstone 1985, p. 35), after 1103 time steps it settles down to a stable state that just fits into a 51-by-109 cell region. Thus, the finite bound of the R pentomino is a weak emergent macrostate of the Game of Life.

The R pentomino is one of the simplest Life configurations that is underivable. What makes Life's underivability so striking is that its microdynamic—the underlying birth-death rule—is so simple.

Glider Spawning. Let G be the structural macrostate of quickly spawning a glider. (To make this property precise, we might define G as, say, the property of exhibiting a glider that survives for at least a three periods, i.e., twelve time steps, within one hundred time steps of evolution from the initial condition.) It is easy to derive that certain Life configurations never spawn a glider and so lack property G. As illustrations, a little a priori reflection allows one to derive that G is absent from each of the five the configurations in Figure 1 (a) - (e), from any configuration consisting of a sparse distribution of those five configurations, from a configuration consisting of all dead cells or all living cells, and from a configuration split straight down the middle into living and dead cells. Similarly, no simulation is necessary to see that some Life configurations have G; for example, consider the configuration consisting of one glider, Figure 1 (f). In general, though, it is impossible to tell whether a given initial Life configuration will quickly spawn a glider, short of observing how the initial condition evolves. Thus, G (or non-G) is weakly emergent in most of the Life configurations that possess (or lack) it, as contemplating a couple of examples makes evident. Figures 4 and 5 show two random initial configurations (above) and their subsequent evolution (below). By timestep 115 the configuration in Figure 4 has spawned no gliders, while by timestep 26 a glider has already emerged from the pattern in Figure 5.

[6] Specifically, indefinite growth is the macrostate defined as the (infinite) disjunction of all those (infinite) sequences s of life states such that, for each positive integer n, there is a time t when s contains more than n living cells.

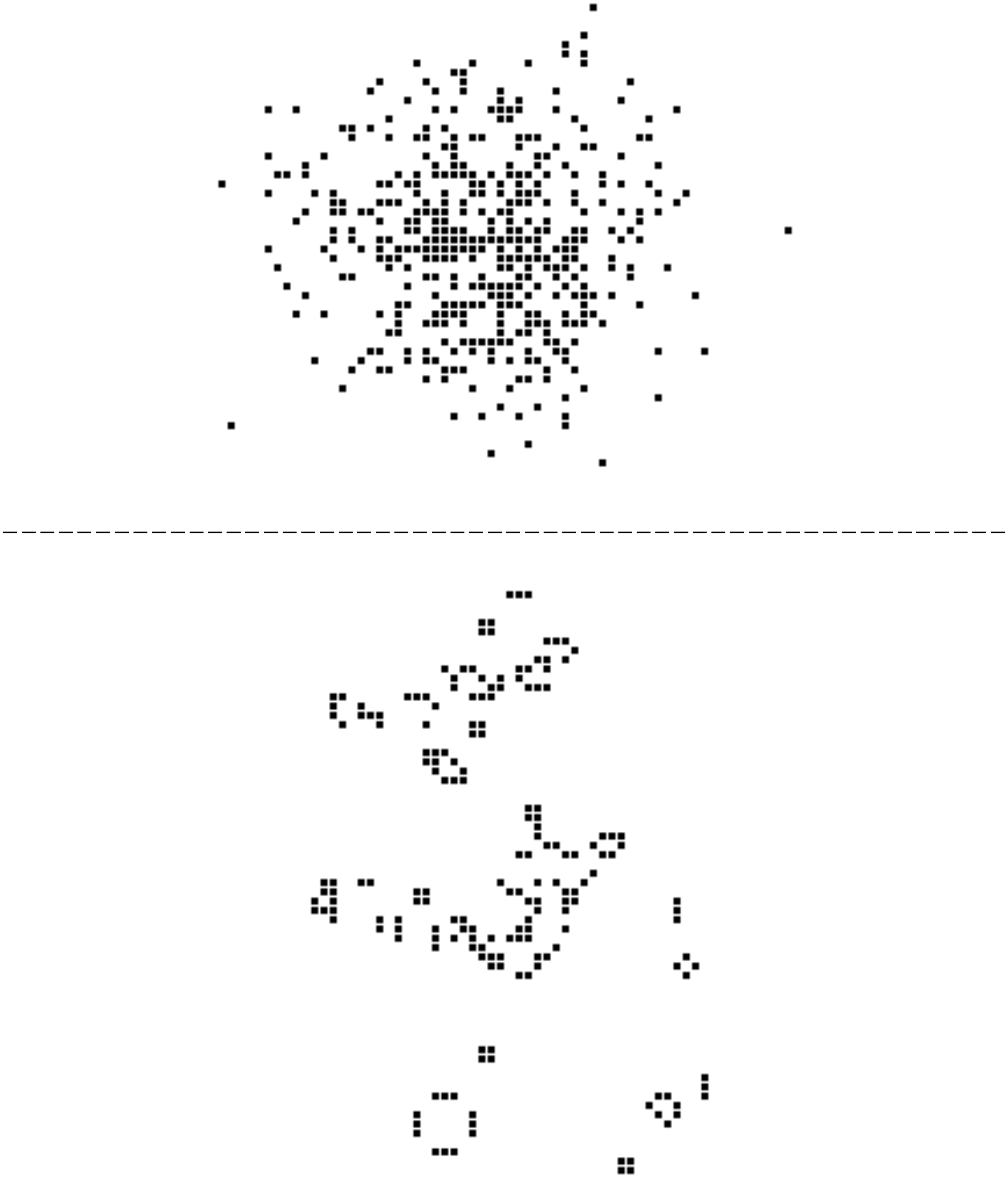


Figure 4. Above: A random distribution of living cells. Below: The distribution after 115 timesteps. No glider has appeared yet.

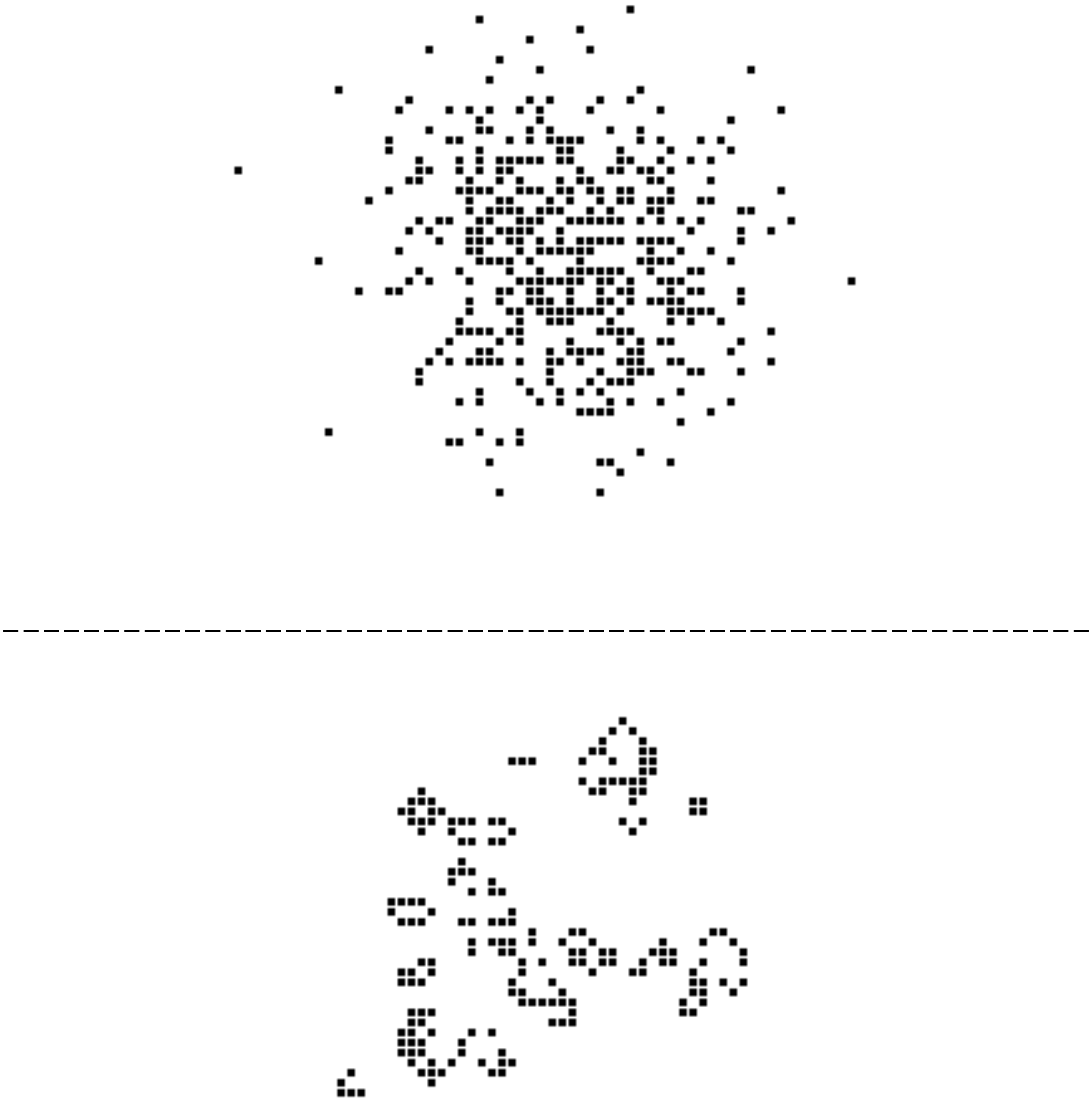


Figure 5. Above: A random distribution of living cells. Below: The distribution after 26 timesteps. A glider is emerging from an unstable cluster of cells at the lower left.

Being weakly emergent does not prevent us from readily discovering various laws involving *G*. If one observes the frequency of occurrence of gliders in lots of random initial configurations, one discovers that usually gliders are quickly spawned; *G* is true of most random Life fields. Extensive enough observation allows one to measure the prevalence of *G* quite accurately, and this information can then be summarized in a little probabilistic law about all random Life fields *X*, of this form: prob(*X* is *G*) = *k*.

Although perhaps not especially fascinating or profound, this little law of the Game of Life nicely illustrates how empirical observation of computer simulations can unearth evidence for laws involving the Game of Life's weakly emergent states.

Empirical observation is generally the *only* way to discover these laws. With few exceptions, it is impossible without simulation to derive the behavior of any macrostate in a Life configuration even given complete knowledge of the configuration. In fact, since a universal Turing machine can be embedded in Life, the undecidability of the halting problem proves that in principle there can be no algorithm for determining whether the behavior exhibited in an arbitrary Life world will ever stabilize. Yet all Life phenomena can be derived from the initial conditions and the birth-death rule. Thus, Conway's Game of Life abounds with weakly emergent properties.

The Game of Life is an exceptionally simple system, simpler than many systems studied in the sciences of complexity. For example, recent artificial life work brims with weak emergence. I will present one illustration involving the emergence of evolvability. Although not as simple as the Game of Life, this next illustration will be more typical of current work in the sciences of complexity.

Weak Emergence in a Model of Evolving Life.

Evolving life forms display various macro-level patterns on an evolutionary time scale. For example, advantageous traits that arise through mutations tend, *ceteris paribus*, to persist and spread through the population. Furthermore, organisms' traits tend, within limits and *ceteris paribus*, to adapt to changing environmental contingencies. These sorts of supple dynamics of adaptation result not from any explicit macro-level control (e.g., God does not adjust allele frequencies so that creatures are well adapted to their environment); rather, they emerge statistically from the micro-level contingencies of natural selection.

Norman Packard devised a simple model of evolving sensorimotor agents which demonstrates how these sorts of supple, macro-level evolutionary dynamic can emerge implicitly from an explicit microdynamical model (Packard 1989, Bedau and Packard, 1992; Bedau, Ronneburg, and Zwick, 1992; Bedau and Bahm, 1993 and 1994; Bedau 1994; Bedau and Seymour, 1994; Bedau 1995*a*). What motivates this model is the view that evolving life is typified by a population of agents whose continued existence depends on their sensorimotor functionality, i.e., their success at using local sensory information to direct their actions in such a way that they can find and process the resources they need to survive and flourish. Thus, information processing and resource processing are the two internal processes that dominate the agents' lives, and their primary goal—whether they know this or not—is to enhance their sensorimotor functionality by coordinating these internal processes. Since the requirements of sensorimotor functionality may well alter as the context of evolution changes, continued viability and vitality requires that sensorimotor functionality can adapt in an open-ended, autonomous fashion. Packard's model attempts to capture an especially simple form of this open-ended, autonomous evolutionary adaptation.

The model consists of a finite two-dimensional world with a resource field and a population of agents. An agent's survival and reproduction is determined by the extent to which it finds enough resources to stay alive and reproduce, and an agent's ability to find resources depends on its sensorimotor functionality—that is, the way in which the agent's perception of its contingent local environment affects its behavior in that environment. An agent's sensorimotor functionality is encoded in a set of genes, and these genes can mutate when an agent reproduces. Thus, on an evolutionary time scale, the process of natural selection implicitly adapts the population's sensorimotor strategies to the environment. Furthermore, the agents' actions change the environment because agents consume resources and collide with each other. This entails that the mixture of sensorimotor strategies in the population at a given moment is a significant component of the environment that affects the subsequent evolution of those strategies. Thus, the "fitness function" in Packard's model—what it takes to survive and reproduce—is constantly buffeted by the contingencies of natural selection and unpredictably changes (Packard 1989).

All macro-level evolutionary dynamics produced by this model ultimately are the result of an explicit micro-level microdynamic acting on external conditions. The model explicitly controls only local micro-level states: resources are locally replenished, an agent's genetically encoded sensorimotor strategy determines its local behavior, an agent's behavior in its local environment determines its internal resource level, an agent's internal resource level determines whether it survives and reproduces, and genes randomly mutate during reproduction. Each agent is autonomous in the sense that its behavior is determined solely by the environmentally-sensitive dictates of its own sensorimotor strategy. On an evolutionary time scale these sensorimotor strategies are continually refashioned by the historical contingencies of natural selection. The aggregate long-term behavior of this microdynamic generates macro-level evolutionary dynamics only as the indirect product of an unpredictably shifting agglomeration of directly controlled micro-level events (individual actions, births, deaths, mutations). Many of these evolutionary dynamics are weakly emergent; although constituted and generated solely by the micro-level dynamic, they can be derived only through simulations. I will illustrate these emergent dynamics with some recent work concerning the evolution of evolvability (Bedau and Seymour 1994).

The ability to adapt successfully depends on the availability of viable evolutionary alternatives. An appropriate quantity of alternatives can make evolution easy; too many or too few can make evolution difficult or even impossible. For example, in Packard's model, the population can evolve better sensorimotor strategies only if it can "test" sufficiently many sufficiently novel strategies; in short, the system needs a capacity for evolutionary "innovation." At the same time, the population's sensorimotor strategies can adapt to a given environment only if strategies that prove beneficial can persist in the gene pool; in short, the system needs a capacity for evolutionary "memory."

Perhaps the simplest mechanism that simultaneously affects both memory and innovation is the mutation rate. The lower the mutation rate, the greater the number of genetic strategies "remembered" from parents. At the same time, the higher the mutation rate, the greater the number of "innovative" genetic strategies introduced with children. Successful adaptability requires that these competing demands for memory and innovation be suitably balanced. Too much mutation (not enough memory) will continually flood the population with new random strategies; too little mutation (not enough innovation) will tend to freeze the population at arbitrary strategies. Successful evolutionary adaptation requires a mutation rate suitably intermediate between these extremes. Furthermore, a suitably balanced mutation rate might not remain fixed, for the balance point could shift as the context of evolution changes.

One would think, then, that any evolutionary process that could continually support evolving life must have the capacity to adapt automatically to this shifting balance of memory and innovation. So, in the context of Packard's model, it is natural to ask whether the mutation rate that governs <u>first-order</u> evolution could adapt appropriately by means of a <u>second-order</u> process of evolution. If the mutation rate can adapt in this way, then this model would yield a simple form of the evolution of evolvability and, thus, might illuminate one of life's fundamental prerequisites.

Previous work (Bedau and Bahm 1993, 1994) with <u>fixed</u> mutation rates in Packard's model revealed two robust effects. The first effect was that the mutation rate governs a phase transition between genetically "ordered" and genetically "disordered" systems. When the mutation rate is too far below the phase transition, the whole gene pool tends to remain "frozen" at a given strategy; when the mutation rate is significantly above the phase transition, the gene pool tends to be a continually changing plethora of randomly related strategies. The phase transition itself occurs over a critical band in the spectrum of mutation rates, μ, roughly in the range $10^{-3} \leq \mu \leq 10^{-2}$. The second effect was that evolution produces maximal population fitness when mutation rates are around values just below this transition. Apparently, evolutionary adaptation happens best when the gene pool tends to be "ordered" but just on the verge of becoming "disordered."

In the light of our earlier suppositions about balancing the demands for memory and innovation, the two fixed-mutation-rate effects suggest the balance hypothesis that the mutation rates around the critical transition between genetic "order" and "disorder" optimally balance the competing evolutionary demands for memory and innovation. We can shed some light on the balance hypothesis by modifying Packard's model so that each agent has an additional gene encoding its personal mutation rate. In this case, two kinds of mutation play a role when an agent reproduces: (i) the child inherits its parent's sensorimotor genes, which mutate at a rate controlled by the parent's personal (genetically encoded) mutation rate; and (ii) the child inherits its parent's mutation rate gene, which mutates at a rate controlled by a population-wide meta-mutation rate. Thus, first-order (sensorimotor) and second-order (mutation rate) evolution happen simultaneously. So, if the balance hypothesis is right and mutation rates at the critical transition produce optimal conditions for sensorimotor evolution because they optimally balance memory and innovation, then we would expect second-order evolution to drive mutation rates into the critical transition. It turns out that this is exactly what happens.

Figure 6 shows four examples of how the distribution of mutation rates in the population change over time under different conditions. As a control, distributions (a) and (b) show what happens when the mutation rate genes are allowed to drift randomly: the bulk of the distribution wanders aimlessly. By contrast, distributions (c) and (d) illustrate what happens when natural selection affects the mutation rate genes: the mutation rates drop dramatically. The meta-mutation rate is lower in (a) than in (b) and so, as would be expected, distribution (a) is narrower and changes more slowly. Similarly, the meta-mutation rate is lower in (c) than in (d), which explains why distribution (c) is narrower and drops more slowly.

If we examine lots of simulations and collect suitable macrostate information, we notice the pattern predicted by the balance hypothesis: Second-order evolution tends to drive mutation rates down to the transition from genetic disorder to genetic order, increasing population fitness in the process. This pattern is illustrated in Figure 7, which shows time series data from a typical simulation. The macrostates depicted in Figure 7 are (from top to bottom): (i) the mutation rate distribution, as in Figure 6; (ii) a blow up distinguishing very small mutation rates in the distribution (bins decrease in size by a factor of ten, e.g., the top bin shows mutation rates between 10^{-0} and 10^{-1}, the next bin down shows mutation rates between 10^{-1} and 10^{-2}, etc.); (iii) the mean mutation rate (note the log scale); (iv) the uningested resources in the environment; (v) three aspects of the genetic diversity in the population's sensorimotor strategies; and (vi) the population level.

Figure 6. Evolutionary dynamics in mutation rate distributions from four simulations of the model of sensorimotor agents. Time is on the X-axis (100,000 timesteps) and mutation rate is on the Y-axis. The gray-scale at a given point (t,m) in this distribution shows the frequency of the mutation rate m in the population at time t. See text.

Figure 7. Time series data from a simulation of the model of sensorimotor agents, showing how the population's resource gathering efficiency increases when the mutation rates evolve downward far enough to change the qualitative character of the population's genetic diversity. From top to bottom, the data are: (i) the mutation rate distribution; (ii) a blow up of very small mutation rates; (iii) the mean mutation rate (note the log scale); (iv) the uningested resource in the environment; (v) three aspects of the diversity of the sensorimotor strategies in the population; (vi) the population level. See text.

The composite picture provided by Figure 7 can be crudely divided into three epochs: an initial period of (relatively) high mutation rates, during the time period 0 – 20,000; a transitional period of falling mutation rates, during the time period 20,000 – 40,000; and a final period of relatively low mutation rates, throughout the rest of the simulation. The top three time series are different perspectives on the falling mutation rates, showing that the mutation rates adapt downwards until they cluster around the critical transition region, $10^{-3} \leq \mu \leq 10^{-2}$. Since resources flow into the model at a constant rate and since survival and reproduction consume resources, the uningested resource inversely reflects the population fitness. We see that the population becomes more fit (i.e., more efficiently gathers resources) at the same time as the mutation rates drop. Although this is not the occasion to review the different ways to measure the diversity of the sensorimotor strategies in the population, we can easily recognize that there is a significant qualitative difference between the diversity dynamics in the initial and final epochs. In fact, these qualitative differences are characteristic of precisely the difference between a "disordered" gene pool of randomly related strategies and a gene pool that is at or slightly below the transition between genetic order and disorder (see Bedau and Bahm 1993, 1994, Bedau 1995).

If the balance hypothesis is the correct explanation of this second-order evolution of mutation rates into the critical transition, then we should be able to change the mean mutation rate by dramatically changing where memory and innovation are balanced. And, in fact, the mutation rate does rise and fall along with the demands for evolutionary innovation. For example, when we randomize the values of all the sensorimotor genes in the entire population so that every agent immediately "forgets" all the genetically stored information learned by its genetic lineage over its entire evolutionary history, the population must restart its evolutionary learning job from scratch. It has no immediate need for memory (the gene pool contains no information of proven value); instead, the need for innovation is paramount. Under these conditions, we regularly observe the striking changes illustrated around timestep 333,333 in Figure 8. The initial segment (timesteps 0 – 100,000) in Figure 8 shows a mutation distribution evolving into the critical mutation region, just as in Figure 7 (but note that the time scale in Figure 8 is compressed by a factor of five). But at timestep 333,333 an "act of God" randomly scrambles all sensorimotor genes of all living organisms. At just this point we can note the following sequence of events: (a) the residual resource in the environment sharply rises, showing that the population has become much less fit; (b) immediately after the fitness drop the mean mutation rate dramatically rises as the mutation rate distribution shifts upwards; (c) by the time that the mean mutation rate has risen to its highest point the population's fitness has substantially improved; (d) the fitness levels and mutation rates eventually return to their previous equilibrium levels.

Figure 8. Time series data from a simulation of the model of sensorimotor agents. From top to bottom, the data are: (i) a blow up of very small mutation rates in the mutation rate distribution; (ii) mean mutation rate (note the log scale); (iii) the level of uningested resources in the world; (iv) population level. At timestep 333,333 all sensorimotor genes of all living organisms were randomly scrambled. See text.

All of these simulations show the dynamics of the mutation rate distribution adjusting up and down as the balance hypothesis would predict. Temporarily perturbing the context for evolution can increase the need for rapid exploration of a wide variety of sensorimotor strategies and thus dramatically shift the balance towards the need for innovation. Then, subsequent sensorimotor evolution can reshape the context for evolution in such a way that the balance shifts back towards the need for memory. This all suggests that, *ceteris paribus*, mutation rates adapt so as to balance appropriately the competing evolutionary demands for memory and innovation, and that, *ceteris paribus*, this balance point is at the genetic transition from order to disorder. An indefinite variety of environmental contingencies can shift the point at which the evolutionary need for memory and innovation are balanced, and the perturbation experiments show how mutation rates can adapt up or down as appropriate.

This sort of supple adaptability in Packard's model can be counted among the hallmarks of life in general (Maynard Smith 1975, Cairns-Smith 1985, Bedau 1995*b*). And, clearly, these evolutionary dynamics are weakly emergent. The model's macro-level dynamic is wholly constituted and generated by its micro-level phenomena, but the micro-level phenomena involve such a kaleidoscopic array of non-additive interactions that the macro-level dynamics cannot be derived from micro-level information except by means of simulations, like those shown above. In a similar fashion, many other characteristic features of living systems can be captured as emergent phenomena in artificial life models; see, e.g., Farmer *et al*. (1986), Langton (1989), Langton *et al*. (1992), Varela and Bourgine (1992), Brooks and Maes (1994), Gaussier and Nicoud (1994), Stonier and Yu (1994), Banzhaf and Eeckman (1995).

Support for Weak Emergence.

Conway's Game of Life and Packard's model of evolving life serve to clarify weak emergence and illustrate its role in the sciences of complexity. But one might still ask whether weak emergence is philosophically interesting and, indeed, whether it deserves the name "emergence" at all. These questions deserve answers, especially since weak emergence differs significantly from traditional twentieth century accounts of emergence.

For example, since weakly emergent properties can be derived (via simulation) from complete knowledge of micro-level information, from that information they can be *predicted*, at least in principle. If we have been observing a simulation of some system *S* and at time *t* we saw that *S* was in state *P*, then we know that there is an appropriate derivation that *S* will be in macrostate *P* at *t*.[7] So, if we are give a system's microdynamic and all relevant

[7] This can be spelled out as follows: Let $\underline{C_i}$ be the set of microstates of all the parts of *S* at time *i*. Apply *D* (possibly with nondeterministic steps) to the *S*'s initial condition $\underline{C_0}$ (and possibly include a property synchronized

external conditions, then in principle we can derive the system's behavior because we can simulate the system and observe its behavior for as long as necessary. And if we can derive how the system will behave, we can predict its future behavior with complete certainty. Thus, on this key issue weak emergence parts company with at least the *letter* of those traditional conceptions of emergence (e.g., Broad 1925, Pepper 1926, Nagel 1961) that focus on in principle unpredictability of macrostates even given complete microstate information.

At the same time, weak emergence does share much of the *spirit* of those traditional views that emphasize unpredictability. For one thing, in the case of open systems, making the prediction would require prior knowledge of all details of the flux of accidental changes introduced by contact with the external world; and in the case of nondeterministic systems, it would require knowledge of all the nondeterministic events affecting the system's behavior. This sort of knowledge is beyond us, except "in principle;" so weak emergent macrostates of such systems are predictable *only* "in principle." Furthermore, even for closed and deterministic systems, weak emergent macrostates can be "predicted" *only* by observing step-by-step how the system's behavior unfolds. For example, one can "predict" whether an R pentomino will grow forever only by observing in time what happens to the configuration. Some might find this so unlike what should be expected of a prediction that they would agree with Stone (1989) that it is no prediction at all.

One might worry that the concept of weak emergence is fairly useless since we generally have no *proof* that a given macrostate of a given system is underivable without simulation.[8] For example, I know no proof that the unlimited growth of the R pentomino and glider-spawning probability can be derived only by simulation; for all I know there *is* no such proof. On these grounds some might conclude that weak emergence "suffers in the course of application in practice", to use Klee's words (1984, p. 49). I would strenuously disagree, however, since unproven weak emergence claims can, and often do, still possess substantial *empirical* support. My earlier weak emergence claims about R pentomino growth and random glider spawning, although unproved, still have more than enough empirical support. Similar weak emergence claims have substantial empirical support. A significant part of the activity in artificial life consists of examining empirical evidence about interesting emergent phenomena in living systems; *mutatis mutandis*, the same holds for the rest of the sciences of complexity.

sequence of external conditions) through successor conditions $\underline{C_i}$ until $\underline{D}$ yields $\underline{C_t}$. From $\underline{C_t}$ and the structural definition of $\underline{P}$, determine whether $\underline{P}$ obtains at $\underline{t}$.

[8] It is a mathematical fact whether a given macrostate of a given system is underivable from the system's microdynamics and external conditions. So, unless it's undecidable, it's provable. Nevertheless, being provable does not entail that it is easy, or even humanly possible, to find and evaluate the proof.

One might object that weak emergence is too weak to be called "emergent", either because it applies so widely or arbitrarily that it does not demark an interesting class of phenomena, or because it applies to certain phenomena that are not emergent. For example, indefinitely many arbitrary, ad hoc Life macrostates are (for all we know) underivability without simulation. Or, to switch to a real world example, even though the potentiality of a certain knife to slice a loaf of bread is "not the sort [of macrostate] emergence theorists typically have in mind" (O'Conner 1994, p. 96), the knife's potentiality might well be weakly emergent with respect to its underlying molecular microdynamic. But this breadth of instances, including those that are arbitrary or uninteresting to "emergence theorists", is not a problem or flaw in weak emergence. Weak emergence explicates an everyday notion in complexity science. It is not a special, intrinsically interesting property; rather, it is widespread, the rule rather than the exception. So not all emergent macrostates are interesting; far from it. A central challenge in complexity science is to identify and study those exceptional, especially interesting weak emergent macrostates that reflect fundamental aspects of complex systems and are amenable to empirical investigation. Simulation gives us a new capacity to identify and study important macrostates that would otherwise beyond the reach of more traditional mathematical or empirical methods.

The micro-level derivability of weak emergent phenomena might be thought to deprive them of the right to be called "emergent"; they might not seem "emergent" enough. The impetus behind this worry might come partly from the history of emergence concepts being ineliminably and unacceptably mysterious—as if no acceptable and non-mysterious concept could deserve to be called "emergence." By contrast, part of my defense of weak emergence is precisely that it avoids the traditional puzzles about emergence.

In any event, there are good reasons for using the word "emergence" in this context. For one thing, complexity scientists themselves routinely use this language and weak emergence is an explication of their language.[9]

[9] Even if we adopt the quite simplistic expedient of restricting our attention to the titles of research reports, we can easily generate a rich range of examples of this language. E.g., rummaging for a few minutes in a handful of books within easy reach produced the following list, all of which speak of emergence in the weak sense defined here in their titles: "Emergent Colonization in an Artificial Ecology" (Assad and Packard 1992), "Concept Formation as Emergent Phenomena" (Patel and Schnepf 1992), "A Behavioral Simulation Model for the Study of Emergent Social Structures" (Drogoul et al. 1992), "Dynamics of Artificial Markets: Speculative Markets and Emerging 'Common Sense' Knowledge" (Nottola, Leroy, and Davalo 1992), Emergent Computation: Self-Organizing, Collective, and Cooperative Phenomena in Natural and Artificial Computing Networks (Forrest 1989), "Emergent Frame Recognition and its Use in Artificial Creatures" (Steels 1991),"The Coreworld: Emergence and Evolution of Cooperative Structures in a Computational

Another compelling reason for allowing the "emergence" language is that weak emergence has the two hallmarks of emergent properties. It is quite straightforward how weak emergent phenomena are constituted by, and generated from, underlying processes. The system's macrostates are constituted by its microstates, and the macrostates are entirely generated solely from the system's microstates and microdynamic. At the same time, there is a clear sense in which the behavior of weak emergent phenomena are autonomous with respect to the underlying processes. The sciences of complexity are discovering simple, general macro-level patterns and laws involving weak emergent phenomena. There is no evident hope of side-stepping a simulation and deriving these patterns and laws of weak emergent phenomena from the underlying microdynamic (and external conditions) alone. In fact, as I emphasized earlier, the micro-level "explanations" of weak emergence are typically so swamped with accidental micro-details that they obscure the macro-level patterns. In general, we can formulate and investigate the basic principles of weak emergent phenomena only by empirically observing them at the macro-level. In this sense, then, weakly emergent phenomena have an autonomous life at the macro-level. Now, there is nothing inconsistent or metaphysically illegitimate about underlying processes constituting and generating phenomena that can be derived only by simulation. In this way, weak emergence explains away the appearance of metaphysical illegitimacy.

It is also clear why weak emergence is consistent with reasonable forms of materialism. By definition, a weak emergent property can be derived from its microdynamic and external conditions. Any emergent phenomenon that a materialist would want to embrace would have materialistic micro-level components with materialist micro-properties governed by a materialistic microdynamic. Thus, the weak emergent phenomena of interest to the materialists would have a completely materialistic explanation.

Conclusion.

Weak emergence is no universal metaphysical solvent. For example, if (hypothetically, and perhaps <u>per impossible</u>) we were to acquire good evidence that human consciousness is weakly emergent, this would not immediately dissolve all of the philosophical puzzles about consciousness. Still, we <u>would</u> learn the answers to some questions: First, a precise notion of emergence <u>is</u> involved in consciousness; second, this notion of emergence is metaphysically benign. Thus, free from special distractions from emergence, we could focus on the remaining puzzles just about consciousness itself.

As Conway's Game of Life and Packard's model of evolving sensorimotor agents illustrate, weak emergence is ubiquitous in the burgeoning, interdisciplinary nexus of scientific research about complex

Chemistry" (Rasmussen, Knudsen, and Feldberg 1991), "Spontaneous Emergence of a Metabolism" (Bagley and Farmer 1992).

systems. The central place of weak emergence in this thriving scientific activity is what provides the most substantial evidence that weak emergence is philosophically and scientifically important. It is striking that weak emergence is so prominent in scientific accounts of exactly those especially puzzling phenomena in the natural world—such as those involving life and mind—that perennially generate sympathy for emergence. Can this be an accident?

References

Armstrong, D. 1978. *Universals and Scientific Realism–Vol. II: A Theory of Universals*. Cambridge: Cambridge University Press.

Assad, A. M., and N. H. Packard. 1992. "Emergent Colonization in an Artificial Ecology." In Varela and Bourgine (1992), pp. 143-152.

Bagley, R. J, and J. D. Farmer. 1992. "Spontaneous Emergence of a Metabolism." In C. Langton *et al*. (1992), pp. 93-140.

Banzhaf, W., and Eeckman, F. eds. 1995. *Evolution and Biocomputation—Computational Models of Evolution*. Berlin: Springer.

Bechtel, W. and A. Abrahamsen. 1990. *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*. Cambridge, Mass.: B. Blackwell, 1990.

Bedau, M. A. 1994. "The Evolution of Sensorimotor Functionality." In Gaussier and Nicoud (1994), pp. 134-145.

Bedau, M. A. 1995*a*. "Three Illustrations of Artificial Life's Working Hypothesis." In Banzhaf and Eeckman (1995), pp. 53-68.

Bedau, M. A. 1995*b*. "The Nature of Life." In M. Boden (forthcoming).

Bedau, M. A. 1995*c*. "Emergent Models of Supple Dynamics in Life and Mind." Unpublished.

Bedau, M. A., and Bahm, A. 1993. "Order and Chaos in the Evolution of Diversity." In *Proceedings of the Second European Conference on Artificial Life (ECAL'93), Brussels, Belgium*, pp. 51-62.

Bedau, M. A., and Bahm, A. 1994. "Bifurcation Structure in Diversity Dynamics." In Brooks and Maes (1994), pp. 258-268.

Bedau, M. A., Bahm, A., and Zwick, M. 1995. "Variance and Uncertainty Measures of Population Diversity Dynamics." *Advances in Systems Science and Applications*, forthcoming.

Bedau, M. A., Giger, M., and Zwick, M. 1995. "Evolving Diversity of Population and Environment in Static Resource Models." *Advances in Systems Science and Applications*, forthcoming.

Bedau, M. A., and Packard, N. 1992. "Measurement of Evolutionary Activity, Teleology, and Life." In C. Langton *et al*. (1992), pp. 431-461.

Bedau, M. A., Ronneburg, F., and Zwick, M. 1992. "Dynamics of Diversity in a Simple Model of Evolution." *Parallel Problem Solving from Nature* 2: 94-104.

Bedau, M. A. and Seymour, R. 1994. "Adaptation of Mutation Rates in a Simple Model of Evolution." In Stonier and Yu (1994), pp. 37-44.

Berlekamp, E.R., J.H. Conway, and R.K. Guy. 1982. *Winning Ways*. Vol. 2. New York: Academic Press.

Boden, M., forthcoming. *The Philosophy of Artificial Life*. New York: Oxford University Press.

Broad, C. D., 1925. *The Mind and Its Place in Nature*. London: Routledge and Kegan Paul.

Brooks, R. and P. Maes, eds. 1994. Artificial Life IV. Cambridge, Mass.: MIT Press.

Cairns-Smith, A. G. 1985. Seven Clues to the Origin of Life. Cambridge: Cambridge University Press.

Crutchfield, J.P., J.D. Farmer, N.H. Packard, and R. S. Shaw. 1986. "Chaos." Scientific American 255 (December): 46-57.

Drogoul, A., J. Ferber, B. Corbara, and D. Fresneau. 1992. "A Behavioral Simulation Model for the Study of Emergent Social Structures." In Varela and Bourgine (1992), pp. 161-170.

Farmer, J. D., Lapedes, A., Packard, N., and Wendroff, B., eds. 1986. Evolution, Games, and Learning: Models for Adaptation for Machines and Nature. Amsterdam: North Holland.

Forrest, S., ed. 1989. Emergent Computation: Self-Organizing, Collective, and Cooperative Phenomena in Natural and Artificial Computing Networks. Amsterdam: North-Holland.

Gardner, M. 1983. Wheels, Life, and Other Mathematical Amusements. New York: Freeman.

Gaussier, P., and Nicoud, J. -D., eds. 1994. From Perception to Action. Los Alamitos, Calif.: IEEE Computer Society Press.

Gleick, James. 1987. Chaos : Making a New Science. New York: Viking.

Holland, J. H. 1992. Adaptation in Natural and Artificial Systems, 2nd edition. Cambridge, Mass: MIT Press.

Horgan, T. and J. Tienson, eds. 1991. Connectionism and the Philosophy of Mind. Dordrecht: Kluwer Academic.

Kellert, S. H. 1993. In the Wake of Chaos: Unpredictable Order in Dynamical Systems. Chicago: The University of Chicago Press.

Kim, J. 1984. "Concepts of Supervenience." Philosophy and Phenomenological Research 45: 153-176.

Klee, R. 1984. "Micro-Determinism and Concepts of Emergence." Philosophy of Science 51: 44-63.

Langton, C., ed. 1989. Artificial Life. SFI Studies in the Sciences of Complexity, Vol. VI. Redwood City, Calif.: Addison-Wesley.

Langton, C., C. E. Taylor, J. D. Farmer, S. Rasmussen, eds. 1992. Artificial Life II. SFI Studies in the Sciences of Complexity, Vol. X. Reading, Calif.: Addison-Wesley.

Levy, S. 1992. Artificial Life: The Quest for a New Creation. New York: Pantheon Books.

Lewin, R. 1992. Complexity: Life at the Edge of Chaos. New York: Macmillan

Maynard Smith, J. 1975. The Theory of Evolution, 3rd edition. New York: Penguin.

Nagel, E., 1961. The Structure of Science. New York: Harcourt, Brace & World.

Nottola, C., F. Leroy, and F. Davalo. 1992. "Dynamics of Artificial Markets: Speculative Markets and Emerging 'Common Sense' Knowledge." In Varela and Bourgine (1992), pp. 185-194.

O'Conner, T. 1994. "Emergent Properties." American Philosophical Quarterly 31: 91-104.

Packard, N. 1989. "Intrinsic Adaptation in a Simple Model for Evolution." In Langton (1989), pp. 141-155.

Patel, M., and U. Schnepf. 1992. "Concept Formation as Emergent Phenomena." In Varela and Bourgine (1992), pp. 11-20.

Pepper, S., (1926), "Emergence." Journal of Philosophy 23: 241-245.

Poundstone, W. 1985. The Recursive Universe. Chicago: Contemporary Books.

Ramsey, W., S. P. Stich, and D. E. Rumelhart, eds. 1991. Philosophy and Connectionist Theory Hillsdale, N.J.: L. Erlbaum Associates.

Rasmussen, S., C. Knudsen, and R. Feldberg. 1991. "The Coreworld: Emergence and Evolution of Cooperative Structures in a Computational Chemistry." Physica D 42: 121-30.

Rumelhart, D. E., J. L. McClelland, and the PDP Research Group. 1986. Parallel Distributed Processing : Explorations in the Microstructure of Cognition. Cambridge, Mass.: MIT Press.

Steels, L. 1991. "Emergent Frame Recognition and its Use in Artificial Creatures." In Proceedings of the 10th IJCAI. San Mateo, Calif.: Morgan Kaufmann.

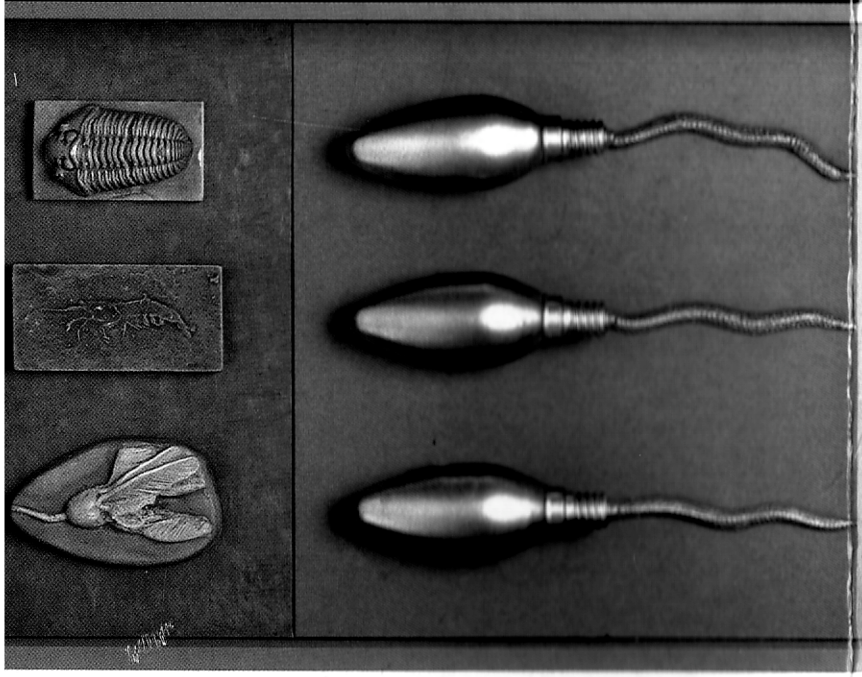
Stone, M. A. 1989. "Chaos, Prediction, and Laplacean Determinism." American Philosophical Quarterly 26: 123-131.

Stonier, R. and X. H. Yu, eds. 1994. Complex Systems--Mechanisms of Adaptation. Amsterdam: IOS Press.

Varela, F., and P. Bourgine. 1992. Towards a Practice of Autonomous Systems. Cambridge, Mass.: MIT Press.

Waldrop, M. M. 1992. Complexity: The Emerging Science at the Edge of Order and Chaos. New York: Simon & Schuster.

Wolfram, S. 1994. Cellular Automata and Complexity. Reading, Mass.: Addison-Wesley.



Professor George C. Williams is recognized internationally as a leading authority in the study of evolutionary biology, and in *Plan & Purpose in Nature* he examines Darwinian evolution in the natural world. He tells the story not only of adaptations which natural selection produces through nature, but also the limitations of evolution for 20th-century human beings, and how the rapid evolution of micro-organisms is likely to pose an alarming threat to human health.

PHOENIX

COVER ILLUSTRATION :
STUART HAYGARTH

NON-FICTION/SCIENCE
£5.99 IN UK ONLY

science masters

PLAN & PURPOSE IN NATURE

GEORGE C. WILLIAMS

CHAPTER 2

FUNCTIONAL DESIGN AND NATURAL SELECTION

In 1859, when Darwin published *The Origin of Species*, the idea of evolution was very much in the air. Scientists generally recognized that fossils are the petrified remains of creatures long dead and often extinct, some strikingly different from anything living today. They also recognized that the plants and animals they knew were totally absent from fossil assemblages in many rock layers. Life forms had apparently changed through the ages, and explanations for why this should be were various. The *catastrophist* school of thought assumed that the strange organisms of earlier ages had all been destroyed in great calamities, with other plants and animals created to replace them after each calamity. The renowned French biologist Lamarck thought that current animals and plants had evolved slowly from earlier forms. He envisioned evolution occurring partly as a kind of predetermined developmental process and partly from the compulsive strivings of the organisms themselves.

Physical scientists studying rock formations were also devising evolutionary theories for what they observed in the Earth's crust. They came to believe that some rocks form from the slow consolidation of sediments that gradually accumulate and that others form in other ways, for instance, from hot volcanic masses forcing their way through overlying rocks or flowing out onto the surface. They understood that the relative ages of adjacent rocks might be inferred. Younger sediments usually lie on top of older ones, and a volcanic intrusion must be younger than the rocks through which it flowed. Absolute ages were more in doubt, but calculating how long it would take known processes to produce observed results suggested that the Earth must be far older than would be allowed by biblical reckonings.

Perhaps the most noteworthy of these pioneers was the Scotsman James Hutton, regarded by some as the founder of historical geology. His writings implied a possibly infinite age for the Earth, which he envisioned to be in a slow but unending state of repetitive upheaval. In 1785, a quarter-century before Darwin's birth, he maintained that an objective and comprehensive examination of crustal rocks revealed "no vestige of a beginning, no prospect of an end." So the idea of an immense amount of time available for evolutionary change was intellectually respectable in Darwin's time, and it implied that even very slow evolutionary processes might bring about great changes.

DARWIN'S THEORY OF EVOLUTION

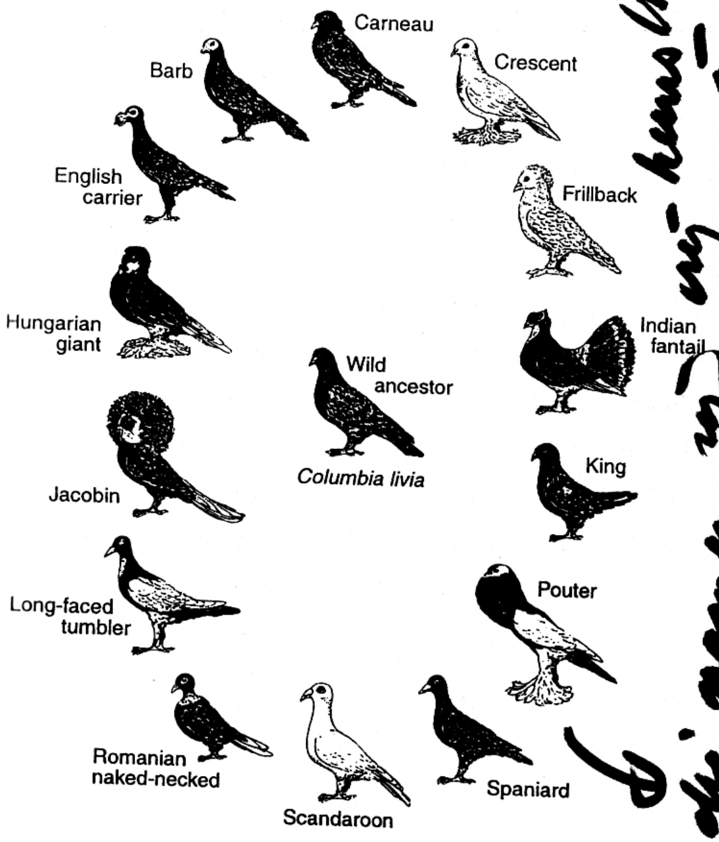
The slow process that Darwin proposed as most important in evolution was *natural selection*, a process deduced from two abundantly supported generalizations. The first is that there is a struggle for existence throughout the living world. In every species of plant or animal, more individuals are produced in every generation than can possibly survive and reproduce. Some will succeed, others fail. The second generalization is that there is such a thing as heredity. Offspring tend to resemble their parents more than they do other adults of the parental generation. Darwin reasoned that such variation could affect characters important in the struggle for existence, and he found many examples of this sort of variation. It follows that each generation will have a biased representation of the variations found in the preceding one. Whatever helped their parents in their struggle for existence will be more abundantly represented in the surviving offspring than traits that handicapped individuals in the parental generation.

He supported this theory by a massive array of evidence from natural history and by analogies with artificial selection. Breeders, in choosing individual plants and animals for breeding stock, usually select those with the features they like best. Those less well endowed are sold or eaten or otherwise eliminated. By this sensible practice they can induce gradual change over many generations, so that domesticated forms often look and act quite different from their wild ancestors. Today, after thousands of years of selective breeding under domestication, we have breeds of dogs and horses and roses and strawberries that are quite different among themselves and from the wild species from which they were derived.

Darwin himself bred pigeons and used the origin of pigeon breeds as a model for the origin of diverse stocks from a single ancestral species. The illustration opposite shows some of the diversity of pigeon types, all produced by rapid evolution under domestication.

Darwin argued that if farmers or hobbyists, by frequently culling the least valuable of their pigeons or pigs or potatoes, can produce varieties that are economically or aesthetically superior to the ancestors, nature can surely do something similar. Competition and adverse conditions of life impose an automatic culling process in every generation. The result should be that the wild animals and plants develop ever greater ability to survive this culling process. The philosopher Herbert Spencer later called this principle *the survival of the fittest*, a handy if somewhat misleading phrase.

Darwin argued that if this process operated through enormous numbers of generations and, especially, if the environmental conditions that caused the culling



A sampling of the diversity of domestic pigeons, all derived from the wild Eurasian rock dove (center), through recent centuries of selective culling by breeders. Compare the diversity here with that of Darwin's finches (p. 37), for which hundreds of thousands, perhaps millions, of years were available.

die namen op wij-kenns (maten)
alsof we naar een film aan het kijke zijn ende
Reeks veranderende scene één naam wille geve

changed from time to time, major evolutionary modifications would be expected. Descendants of a single species of ancestor, if they inhabited different regions subject to different conditions, could diverge to such an extent that they would have to be considered different species from the ancestor and from each other. One of his examples from nature was the finches of the Galápagos Islands, a small archipelago on the equator a thousand kilometers west of South America. The islands rose from the sea as volcanoes a few million years ago, and were never a part of the mainland. They were largely inaccessible for land animals and plants, and most of the common inhabitants of South America did not exist there when the islands were first explored. The absence of things such as frogs and small mammals (other than bats) is easily explained: the great expanse of open ocean is a forbidding barrier, even for birds of the tropical American deserts and forests. Yet it should not be surprising that some limited colonization by land birds did take place as a result of accidental straying from the closest continent.

Darwin, in his visit to the islands as naturalist aboard the research vessel *Beagle,* found the descendants of such colonists, about a dozen species of finch whose closest relatives were in South America. He noted that each major island had one or more of the distinct species. He theorized that sometime after the islands had formed and become habitable, some South American finches, conceivably just one of each sex, reached one of the islands. Perhaps they had been caught in a storm, with an easterly wind blowing faster than they could fly, and lucked upon this isolated land instead of dropping exhausted into the sea. They survived and bred in their newfound home, where competitors were absent and conditions at least minimally met their requirements for food and nest sites. Soon, perhaps in just a few years, they built up a large population, then some of them occasionally reached another of the islands in the archipelago to repeat the process.

But why should there be so many species of Galápagos finch, and not just the one original colonist? Darwin answered this question by noting the differing environmental conditions on different islands. Some are large, comparable in size to Rhodes or Minorca, others smaller than the one that supports the Statue of Liberty. Some have high mountain peaks that catch considerable rain, others are low and dry. The diversity of conditions produces a diversity of vegetation and of the seeds and insects that finches feed upon. These differing environmental circumstances demanded different capabilities in the finches' struggle for existence.

Of special importance is food suitability. Some potential food sources are seeds with formidable shells. If only some of the newly established finches were able to break the shells, powerful selection for these more powerful beaks and jaw muscles would be brought to bear. In perhaps a few thousand generations, the finch populations of different islands would show differences in their feeding adaptations. In hundreds of thousands of generations, the diversity of these features could have reached the level found by Darwin, as illustrated in the figure.

This gradual alteration of a line of descent, with different lines showing different evolutionary changes, was the central theme of Darwin's 1859 book and the inspiration for its title. One of his main sources of inspiration was the differences he often found in comparing members of the same species from different areas. His trip around the world in the *Beagle* provided abundant opportunity for comparing animals and plants across their geographic ranges.

He often found that differences shown by some organism from different parts of its range were only moderately distinct, and he recognized them as *varieties* of the same species. At other times it was not at all clear whether he was dealing with different varieties or fully distinct species. Alfred Russel Wallace, who proposed the theory of natural selection simultaneously with Darwin, entitled his publication "On the Tendency of Varieties to Depart Indefinitely from the Original Type." Today natural selection is part of the standard conceptual equipment of biology, routinely invoked to explain differences among

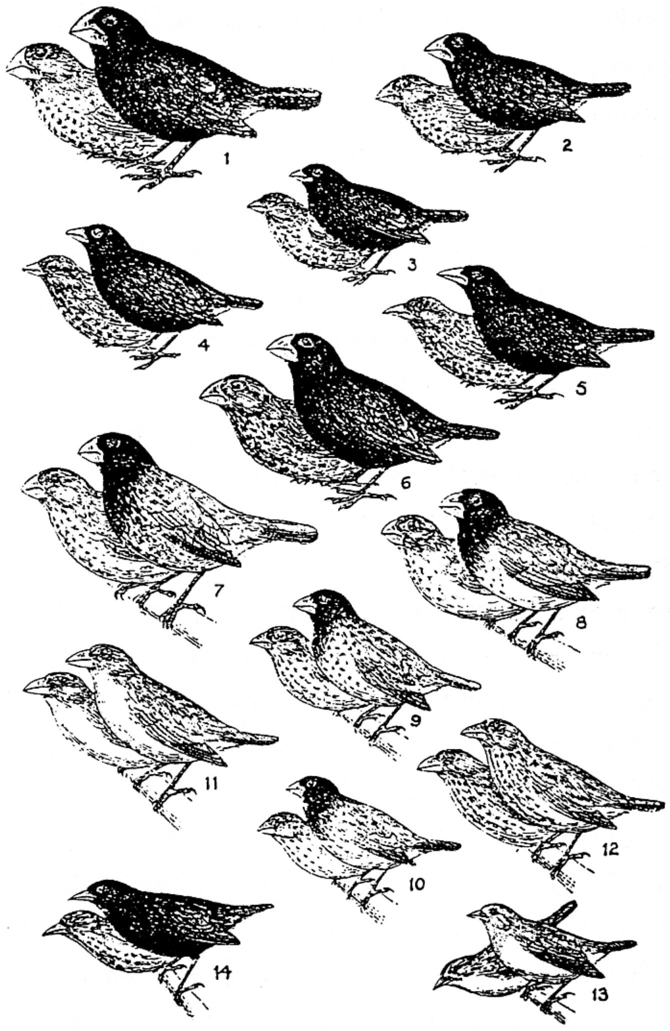


Illustration from David Lack's 1947 book on Darwin's finches. Compare with pigeon diversity shown on page 33.

closely related organisms living under slightly different conditions.

SEXUAL SELECTION

Relentless competition was an essential premise of the theory of natural selection as proposed by Darwin and Wallace. Today biologists recognize two kinds of competitive behavior among animals: *scramble* and *contest* competition. A flock of sparrows picking up grass seeds as fast as a gardener can broadcast them would exemplify scramble competition. Another would be two or more dogs chasing a squirrel, with one catching it and eating the whole carcass. If two dogs seize the same squirrel and a tug of war ensues, we have a contest, the two dogs pitted one on one against each other. It would be even more obviously a contest if they fought over the dead squirrel, directing their attention exclusively to each other, even though the squirrel is the real focus of the dispute. It would also be a contest if the dogs merely threaten each other, without actually fighting. If one dog's threat causes another to back away, it has won the contest.

Tactically similar contests may be waged over food items, roosting sites, mates, and many other resources. In fact, there need not be a currently identifiable resource for a contest to occur. It may be waged merely to establish the winner-loser relationship itself. Later on, when something like a food item is found, the loser will concede it to the winner without further dispute. This phenomenon of stratified social structure is often seen by observers of animal behavior both in the wild and in captivity. Contests are often provoked by the appearance of an item coveted by more than one individual, but the prize being sought may often be an elevated social status, which can later be used to gain access to needed resources.

The general prevalence of competition for social status has been recognized only recently. It was obscured for Darwin, and for generations of his followers, by the fact that the most conspicuous contests are often among males for opportunities to mate with females. This fact led Darwin to propose a special evolutionary factor, *sexual selection*, that operates in addition to natural selection. Sexual selection normally depends on contests between males, with the winner gaining and the loser forgoing sexual access to one or more females. This competition for social status often takes place without any females being present. Many male migratory birds, for example, precede females on the spring migration, and establish their social hierarchy before the females arrive.

In some animal species, one male contends directly with another, the fight between stags in the rutting season being a clear example. In other species, males contend indirectly by displaying to females a train of enormous feathers, a peacock being the classic example. The contest is still between males, the winners being those best at impressing the females. In still other species, the contest is for a territory, and waged with threats or actual combat at tentative territory boundaries. Persistent winners get larger territories in choice breeding habitats; losers settle for smaller territories or inferior habitats. The females may choose males indirectly by seeking the better places to lay their eggs. In other territorial species, a male must actively court a female and try to entice her to his own nesting site. The male threespine stickleback, an object of many classic studies of reproductive behavior and sexual selection, threatens and fights other males to secure a territory in a good nesting habitat. Then he builds a nest and must entice females to it while keeping other males away.

Darwin was led to propose sexual selection by the many conspicuous features of animals that could not be attributed to natural selection as he envisioned it. They were features more likely to hinder than to aid an organism in its struggle for survival. Again, the conspicuous nuptial train of a peacock, which is so burdensome as to make normal flight difficult, is a good example. As a general rule, these conspicuous and burdensome features characterize adult males rather than the females or

eigenlijk geldt hier een marketing idee: een mooie verpakking (veren) is een indicatie voor kwaliteit (good manners)

→ zelfs in de natuur helpt marketing (karakter!)

juveniles of a species, and they often appear only in the breeding season.

Darwin reasoned that the great burden of ornamental plumage may make life difficult and dangerous for a peacock, but he proposed that it may be favorably selected if its display makes it easier for the male to compete for a mate. Sometimes the burdens are clearly weapons. The antlers of a male deer, grown anew each breeding season, are an obvious instance. Other sexually selected features are only for display, playing a role in courting females, threatening rival males, or both.

Sexual selection is an idea routinely invoked in biological research today. It is generally viewed as a pervasive evolutionary factor, more important perhaps than any other kind of selection in relation to contests with other members of the species. The concept has moved beyond an explanation for competitive behavior or the production of weapons or displays to encompass many physiological processes and has been extended to such phenomena as the floral displays and selective use of incoming pollen grains by plants. But during Darwin's lifetime, he was the only prominent biologist to argue the importance of sexual selection, and he used it only to explain special features of adult male animals. Wallace, who in some ways seemed a more extreme advocate of natural selection than Darwin, had little use for Darwin's theory of sexual selection. He ridiculed the idea that females could

dat 't moeilijk eruit zal [illegible] krijgen en zich voortplanten — ongeacht dat ook nog 'ns lelijkerd is — het einde van voedsel maar jammer genoeg moet kwaliteit [illegible] [illegible] [illegible] maar de [illegible]; marketing loont!

[illegible] genoeg in de natuur. Het mannetje [illegible] kwaliteiten. Dat gevolg [illegible] [illegible] heeft — geen [illegible]. Natuur [illegible] dat maar [illegible] [illegible] man be influenced by the displays that males seemed to be directing toward them. Even as late as 1950, biologists studying animal behavior made little mention of sexual selection. In this and other ways, Darwin was far ahead of his contemporaries.

[illegible]

But many modern biologists would concede that Darwin erred in thinking of sexual selection as a process separate from natural selection. They regard it as a special category of selection for social status, which is a kind of natural selection. This idea implies recognizing that members of one's own species are a feature of the environment and that adaptations to this feature are expected. Social status is a kind of resource that can never be in adequate supply. The top dog in a pack has all he needs, no doubt at great cost to himself and others, but the other dogs all need more than they have and will do what they can to get it. Unlike food, an individual's social status is a resource that can never be lost to a member of another species. A ferret may compete with a fox for a rabbit, but never for social status among foxes. (Or, perhaps, almost never. The first-century horse Incititus gained a lofty rank in human society at the expense of multitudes of men of lower rank.)

Darwin achieved a high level of recognition in Victorian society as a scholar and expert on many aspects of natural history and for establishing the acceptability of evolution. Yet in retrospect we can recognize that he failed to convince many people that natural selection is the main force of adaptive change. From Darwin's death in 1882 to the 1920s, evolution, his "descent with modification," was generally accepted by biologists, but not natural selection or sexual selection as causes of the modification. Many leading scientists of this period advocated theories that today seem naive and fanciful, such as orthogenesis, the idea that evolution has some kind of momentum that keeps it going. Some continued to prefer Lamarck's ideas to Darwin's.

NATURAL SELECTION AND THE PREVENTION OF EVOLUTION

Paradoxically, much reference to natural selection today relates not so much to evolution as to its absence. If natural selection is the reason the pony fish keeps its photophore, then it is preventing the loss of this organ by evolutionary change. We know now, from abundant experiments on the evolutionary potential of living organisms, that they are capable of evolving far more rapidly than is normally observed today or found in the fossil record. What natural selection mainly does is to cull departures from the currently optimum development of the features shown by organisms. If some species of bird has wings that average 20 centimeters long, it is assumed that individuals with wings of 19 or 21 centimeters would have a slight disadvantage. They would be less likely to survive to maturity and would have lower survival or fertility rates thereafter. Evidence for exactly this was shown by a classic study of natural selection in the wild. In 1899 the British biologist Herman Bumpus measured the wings of a large number of sparrows that had been killed in a storm. He found that those with markedly longer or shorter wings were more abundantly represented among those killed than in the population at large.

The advantage of having intermediate character development (wing length, insulin production, coloration, and so on) is often called *normalizing selection* or *optimization*. Most of the selection taking place in nature is assumed to be of this sort, rather than any that would cause an observable shift in average values from one generation to the next. Even the weak directional selection that does take place is usually thought to be corrective. The population would evolve to be less well adapted if natural selection did not weed out occasional adverse mutations or locally maladaptive genes introduced by individuals moving in from places where conditions are different. So the process proposed by Darwin as the major cause of evolution is now thought to operate mainly to *prevent* evolution. Aristotle's descriptions of wild animals and plants, written 2,500 years ago, are still accurate for their descendants today, mainly because natural selection has been preventing their evolution. The domesticated animals and plants that Aristotle observed were often strikingly different from what farmers grow today, because artificial selection has been causing their rapid evolution.

The concept of character optimization has been with us ever since people first tried to understand the workings of their own bodies and those of other organisms. Aristotle and Galen used the idea habitually, as noted in chapter 1. In 1779 the British philosopher David Hume, in contemplating the quantitative precision of biological mechanisms, proclaimed:

> All these various machines, and even their most minute parts, are adjusted to each other with an accuracy which ravishes into admiration all men who have ever contemplated them. The curious adapting of means to ends, throughout all nature, resembles exactly, though it much exceeds, the productions of human contrivance.

Note the similarity in sentiment between this statement from Hume, an atheist, and that of the orthodox Christian Paley (see chapter 1). Both were clear-thinking and keen observers of nature.

More recently the concept of optimization has been extended to aspects of biology where its applicability is less obvious. Many recent studies of life histories and animal behavior are good examples. Biologists today speak of the optimization of egg size and number, of mate choice, of the seasonal timing of migration. Optimization is used to understand and predict such things as how long a bee will stay at one clump of flowers, how big a load of pollen or nectar it will pick up before returning to the hive, and at what times of the day it will go foraging.

It is ironic that many prominent biologists, during Darwin's time and for many decades thereafter, tried valiantly to demonstrate some force of evolutionary adaptation that could cause change more rapidly than natural selection. They could not imagine that so weak and misguided a process as Darwin proposed could actually produce the observed complexity and diversity of life, even with liberal estimates of the amount of time available for it. Nowadays it is more fashionable to wonder what makes evolution so slow. Some organisms living today are closely similar to ancestors of more than a hundred million years ago. The orthodox reasoning is, in the words of Roger Milkman, a distinguished geneticist at the University of Iowa, that "[t]he main day-to-day effect of natural selection is the maintenance of the status quo, the stabilization of the phenotype. To a relatively small directional residue, we attribute the great panorama of evolution."

The current trend is not to doubt that natural selection could produce the great panorama but to doubt that it can

account for the stability. It is proposed that selection also acts at higher levels than the loss or survival of genes in populations. The extinction of whole evolving lineages could have persistent biases that cull most newly formed groups of organisms in ways analogous to the weeding out of most mutations within evolving populations. So a natural selection among whole populations or larger groups of organisms would be, like selection within populations, concerned mainly with maintaining the status quo.

GENETICS, MOLECULAR BIOLOGY, AND MODERN DARWINISM

Darwin's vague generalization that "like begets like" is an adequate premise for the basic logic of his theory of natural selection, but it does not permit many quantitative inferences. It gives no hint as to why offspring should show a resemblance to their parents. Today we have the science of genetics, with a detailed theory of heredity that allows much more rigorous thought about evolution than was possible in the nineteenth century. Histories of scientific fields are usually vague about origins, but genetics is an exception. It began decisively in the 1860s with experiments on peas grown in his monastery garden by the Augustinian monk Gregor Mendel. He published his work in 1868, but it was ignored for the rest of the century. In the early 1900s it was discovered by several biologists investigating heredity in a variety of different organisms. They belatedly recognized the profound significance of the work of that lonely scientist.

What Mendel had found, and the later workers confirmed, was that crosses between parents of strains that differ strikingly in some character will often show predictable ratios of the contrasting features in subsequent generations. A parental character, such as short stems, may disappear entirely in the first offspring generation, all of which have long stems (the *dominant* character). Crosses between these individuals will produce offspring of which about 25 percent show the (*recessive*) short stems missing in their hybrid parents. Crosses between a first-generation hybrid and the recessive strain produce a nearly equal number of long and short stems (dominants and recessives) in their offspring.

These regularities (*Mendelian ratios*) can be explained by a precisely controlled and strictly particulate theory of heredity. Today we call the inherited particles *genes*. They are particulate in that they retain their identity in passing through generations. A gene is either inherited or not, passed on or not, with never any sort of partial presence. By about 1930 it was clearly shown that the genes are in a nearly constant linear arrangement on the chromosomes, visible with special techniques in dividing

cells. The chromosomes are present in pairs, with each member of each pair having the same linear arrangement of genes, one in each pair having come from the mother, the other from the father. So the paired chromosomes imply paired genes. If the gene inherited from one parent differs from that from the other, biologists refer to two different *alleles* of that gene.

When an individual forms an egg or sperm, the corresponding (homologous) chromosomes line up, exchange some corresponding parts, and then separate, each chromosome going at random to one or the other of the resulting cells. This exchange of parts and random segregation of chromosomes assures that any two alleles in an ancestor will ultimately go their separate ways in descendants. The genes pass indefinitely through the generations, but gene combinations (genotypes) are unique and fleeting, as long as reproduction is sexual. The implications are well worth bearing in mind. You got half your genes from your mother and the other half from your father, one-eighth from each great-grandparent, and so on. Each of your children got half your genes, each grandchild a quarter, and so on. You are the bearer of a legacy of genes from the past. Each allele at each locus has its own unique history, back to a possibly remote origin by mutation from a contrasting allele. Yet your genotype never existed before you were conceived and will never be produced again.

For the first half of this century, there was great uncertainty about the genes' chemical nature. In retrospect we can say that it was obvious by the 1940s that genes can be identified with deoxyribonucleic acid (DNA). All doubt was laid to rest by the much-lauded work of James Watson and Francis Crick in 1953. They resolved the detailed chemical structure of DNA and showed how it serves as a medium of communication within a cell lineage and between the generations of multicellular organisms. Because of Watson and Crick, we now know that heredity is not only particulate but *digital.*

Other examples of digital information transfer are printed English words with their twenty-six-letter alphabet, Arabic numerals with their ten-letter alphabet, and Morse code and "computerese" with their binary alphabets. The genetic code has a four-letter alphabet of molecular structures with names abbreviated as *A*, *T*, *G*, and *C*. Any sequence of three such groups can specify a particular amino acid, one of the building blocks of proteins. For example, *C-A-G* specifies the amino acid *glycine*. If the code were changed to *C-C-G*, some protein would contain the amino acid *proline* at the position that would have been occupied by the glycine. On the other hand, changing *C-A-G* to *C-A-A* has no effect: the amino acid specified is still glycine. This is one of many examples of redundancy in the genetic code. Some different DNA sequences are functionally synonymous, just as, in English, *gray* and *grey* mean the same thing. An understanding of the DNA code is basic to an understanding of evolution.

Imagine that, in some population of some organism, a gene has, among its thousands of base pairs, the sequence *C-A-G* and has, for many generations, been reliably putting a glycine into some protein, perhaps an enzyme. The mechanism that allows this gene to be so amazingly stable will be discussed in chapter 5, but for now I will merely point out that no mechanism has absolute reliability. Rarely, the *C-A-G* may change to some other sequence, perhaps *C-C-G*, so that the resulting enzyme, in cells containing the new sequence, will have a proline in place of the glycine. This might affect the action of the enzyme to a considerable extent, or maybe only slightly, perhaps scarcely at all. If the change is an improvement, natural selection may cause the allele with the *C-C-G* to replace that with the *C-A-G* at its position on its chromosome throughout the population.

Any new mutation can be lost by chance. This is the most likely event for any new allele, even one that gives a substantial advantage. But mutations occur with finite frequencies. If *C-A-G* → *C-C-G* has a one-in-a-million probability per germ cell (egg or sperm), and there are about a thousand individuals per generation, a mutant individual should appear about once in every thousand generations, and ten times in ten thousand—a trifle, in evolutionary terms, for many organisms. Sooner or later a favorable mutation should catch on and start replacing the ancestral allele at that locus.

The great beauty of Mendelian heredity in its evolutionary application is that it lends itself readily to quantification and precise reasoning. This is the subject matter of population genetics, a field well established by the 1930s. Population geneticists can deal with such quantities as mutation rate; frequency of recombination of genes on the same chromosome; expected rate of replacement of alleles by better-adapted mutant forms; expected levels of chance deviations from expected rates as a function of population size and other variables; differences in these rates between recessive and dominant genes; and many other influences on the evolutionary process. These quantitative variables can be related to one another algebraically and evolutionary conclusions drawn as solutions to algebraic equations.

For instance, it can be shown that natural selection can be far more powerful than we might intuitively expect, and can accomplish major changes in brief periods of evolutionary time. Imagine maintaining a herd of a thousand gray horses, with a modest level of starting variability in shades of gray and rates of new mutations affecting this character. Visit this herd once per century and remove the palest specimen. Simple calculations can show that this procedure could result in a herd of uniformly black horses well within a million years.

Recently some Swedish workers reached an even more startling conclusion. Assuming nothing more than some cells of a primitive animal with some sensitivity to light and modest rates of mutations affecting that sensitivity, the position of the cells in the body, the transparency of overlying tissues, and other relevant variables, they showed that it could take as little as 400,000 years to evolve the vertebrate eye. This is less than a thousandth of the time that has elapsed since multicellular animals first appeared. This is an especially interesting example because Darwin's critics have long cited the eye as an example of an organ that is far too complex and precise for any short-sighted process such as natural selection to produce.

Intuitions about the evolutionary process can be a great source of ideas but not of conclusions. Conclusions must be based on precise quantitative reasoning, such as realistically formulated mathematical equations or carefully designed graphic models. Such reasoning must be focused in a way that leads to testable expectations about the real world, such as what a series of measurements on a group of fossils will reveal, how an experiment on microorganisms grown in specified environments will turn out, and so on. The maintenance of proper scientific rigor is, of course, seldom easy, even for well-trained scientists.

CHAPTER 3

DESIGN FOR WHAT?

> Now, as each of the parts of the body, like every other instrument, is for the sake of some purpose, viz. some action, it is evident that the body as a whole must exist for the sake of some complex action.
>
> —Aristotle

The textbook for a college biology course I took in 1947 gave the following statement of the theory of natural selection:

> **Variations** of all grades are present among individuals . . .
>
> By the **geometric ratio of increase** the numbers of every species tend to become enormously large; yet the population of each remains approximately constant because . . . many individuals are eliminated; this involves:
>
> A **struggle for existence**; individuals having variations unsuited to the particular conditions in nature are eliminated, whereas those whose variations are favorable will continue to exist and reproduce.

[illegible]

A **process of natural selection** is therefore operative, which results in:

The **survival of the fittest**, or "the preservation of favored races."

The quotation in item 5 is from the subtitle to Darwin's *Origin of Species*. Unfortunately, Darwin was never clear about what he meant by "race." Is the appearance of a novel feather pattern in a flock of domestic pigeons the start of a new race? Or is there a new race only when the pattern is bred for and comes to characterize a large stock? Are individual differences in wild animals and plants to be considered racial differences? Or are races in nature always groups of individuals, often inhabiting different regions but recognizable as belonging to the same species? These variants of different geographic regions were often called *varieties* or *subspecies*, rather than *races*, in Darwin's time. *Subspecies* is the preferred term today.

Whatever Darwin may have meant by *races* in the title of his book, the concept for the theory of natural selection as Darwin used it and as taught in 1947 was that implied in items 1, 2, and 3. It is variation among competing *individuals* that provides the raw material for natural selection. This is also clear from the textbook's subsequent detailed discussion of the theory. Other texts in use for the next two decades continued to imply that natural selection operates among competing individuals of the

Dat is echt zo: "soort" verklaart biologisch gezien niets; is een nergens uit af te leiden fenomeen etc. (nog wat woorden aan wijden?)

[illegible] "soort" is een [illegible]

same neighborhood. Since the 1970s, textbooks have been more likely to be explicit on this point, and to insist that natural selection operates strongly among individuals and that selection among races or other collective entities is usually a weak influence on the course of evolution.

Moving ahead a few years, I find myself in a graduate seminar in marine ecology. The subject is the adaptations by which little fish try to avoid being eaten by big fish. One example recognized is toxic flesh. If a 10-kilogram barracuda eats a 1-kilogram poisonous perch, it may die or at least be sickened and thereby deterred from attacking that kind of prey in the future. It was agreed by the discussion leader and most of the students that poisonous flesh was a good example of a protective adaptation.

But there was one skeptic, and his name was Murray A. Newman. He would soon go on to a distinguished career as director of the splendid public aquarium in Vancouver, but that day he was a lone dissenter. "Wait a minute," he muttered. "How can being toxic protect you? It does nothing to the predator until long after you're dead." The immediate vehement reaction from me and several others: "That's stupid, Murray. The toxicity doesn't have to protect the toxic individual; it protects the species in general." There were no further objections and the discussion continued, but I am not sure that Murray was convinced. I think I was at the time, but not firmly and not for long. I would soon be increasingly nagged by the

grappig en niet [illegible]! wat is in hemelsnaam een soort] (veel [illegible] seeming inconsistency between the theory of natural selection as presented in textbooks and the "good of the species" adaptations that were routinely attributed to the process, sometimes glibly by the same texts that presented a strictly selection-among-individuals form of the theory.

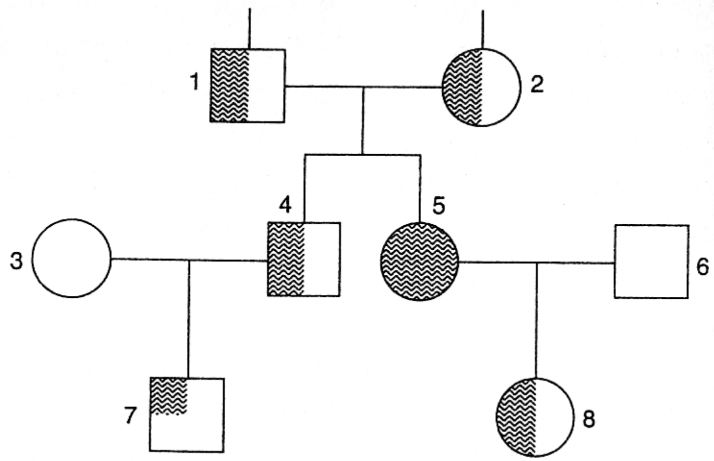
WHAT IS AN ADAPTATION'S ULTIMATE PURPOSE?

The message of chapter 1 was that the parts of organisms are functionally well designed: the eye for vision, the hand for manipulation, and so on. But what are vision and manipulation for? They are important for a long list of vital functions, and without them life would be much more difficult. Vision and manipulation are often called on at the same time, the former helping coordinate the latter. A blind man or a man without hands would take a long time to gather enough wood for a useful fire. What is fire useful for? Cooking or perhaps just warmth. And what good are cooking and warmth?

I could go on with questions beyond questions, but perhaps not much further. Soon I would come to such values as health and useful skills and social status. What are they good for? For Aristotle's "complex action," that's what. Or at least for the modern biological interpretation of the complex action to which all adaptations contribute: reproductive success. Continued physical survival is ordinarily significant in evolution only if it increases the likelihood or extent of reproduction.

But reproduction is a tricky concept for sexual organisms like ourselves. I have children and grandchildren, but I am not really present in either generation. None of those individuals inherited any of my body parts, and none is more than 50 percent similar to me genetically. I happened only once, I will never be duplicated, and when I am gone it will be forever. My descendants' biological heritage from me is limited to a sampling of my genes. Half the genes in each of my children and a quarter of those in each grandchild came from me. As noted in chapter 2, my complete set of genes, my *genotype*, cannot be passed on. So reproduction, for a sexual organism, has a restricted meaning. It means passing on genes, dissociated fragments of the organism's genotype. A mother provides an egg and a father a sperm, and each such cell contains the single set of genes on a single set of chromosomes. The combination of egg and sperm genes provides the genotype of the new individual, which then develops according to its new and unique instructions.

It looks as if reproduction is the ultimate adaptation to which all others are subordinate, but that is too simple, because producing offspring is not the only way to get one's genes into future generations. An organism can also



Genealogical relations in a sexual population. Squares represent males, circles females. (Mates are connected by horizontal lines. Vertical lines indicate the offspring.) The shading indicates genes necessarily shared with individual #5. The process of kin selection would be expected to result in her treating her father (#1) as if his survival and reproduction were half as important as her own, her nephew (#7) a quarter as important (all else being equal), and so on.

transmit its genes by helping with the survival and reproduction of relatives. Consider the diagram, which is meant to show genealogical connections among eight individuals. The overall ability of individual #5 to get her genes (indicated by the shading) into future generations is termed her *inclusive fitness*. The evolutionary process that maximizes the ability to treat others according to their genetic similarity to oneself is termed *kin selection*.

Statements about genetic variation and uniformity in sexually reproducing populations are often confusing

because different measures are being discussed. In a statement such as "We're still 98 percent chimps in our genes," it is obviously base-pair similarity of chimpanzee and human gene pools that is being discussed. If part of some gene in a human cell reads GTTAGCC and exactly the same sequence of the chemical groups (nucleotides) is found in the same place on the same gene in an ape cell, the two are 100 percent similar for this sample. So what? A gene is made up of thousands of base pairs, not just a sequence of seven, as in the example shown. For two genes to be really the same, every one of the thousands of base pairs has to be the same. If proportions of such exactly similar genes are estimated, it may be found that a quarter to a third of the genes may differ in human and ape cells. Human cells from two different individuals may well differ in several percent of their genes.

For a pedigree diagram, statements such as "a child has a 25 percent similarity to a half sibling" would have a still different meaning. Twenty-five percent of the genes are assuredly the same, because they came from the same parent. What of the other 75 percent? We do not know. Many of them might be the same. We know that they came from the same population of which the individuals are members, but that is all we know. We have to consider them a random sample of the genes in that population. The genes indicated by the shading in the diagram are special, and are technically termed genes *identical by descent.* Favorable selection of individual #5 in modern evolutionary theory relates to her success in getting the genes marked by the shading into future generations at a greater rate than competing genes (those shown by the unshaded regions). For this reason, it is better to speak of *genetic success* rather than reproductive success or the survival of the fittest. The important kind of survival cannot just be physical survival; it has to lead to some kind of genetic survival beyond any individual's lifetime.

Suppose we are dealing with an inbred population with very little genetic variability; perhaps 99 percent of the genes in any individual are exactly the same as those in any other. Kin-selection theory still applies, because the shading still represents assuredly identical genes, while only 99 percent of the other genes are the same. How much natural selection can take place in a population in which 99 percent of the genes are identical? Surely some, less than if only 90 percent were the same, but more than if 99.9 percent were. All I am doing here is applying to kin selection the same constraint that applies to any other form of natural selection. The process will not work if there is no genetic variation, and any shortage of such variability will retard the process. Without genetic variation there can be no evolution, from natural selection or any other cause.

The complete absence of genetic variation in a sexually reproducing population that numbers in the thousands or more is so improbable that it is seldom a part of the thinking of evolutionary biologists. Also, as explained in chapter 2, biologists often use natural selection to explain why an organism has the features it has rather than conceivable alternative features. They are not concerned about whether it evolved those features rapidly or slowly. So such biologists are seldom concerned about whether the organisms they study differ at 10 percent of their gene loci or at only 1 percent. Either way, they would expect the same conditions to be produced by selection in the long run, and they assume that the long run has in fact happened. They are interested not in evolutionary change by natural selection but in the evolutionary equilibria already established by this process.

THE FUNDAMENTAL IMPORTANCE OF GENETIC SUCCESS

The distinction between reproductive success and genetic success is clearest in the social insects of the order *Hymenoptera* (ants, bees, and wasps). For the moment I will ignore various complications, such as the great variability of mating systems and social structures, and discuss the classic pattern in this group. A fertile female (*queen*) mates with a male and starts a new colony, with

? Amazing to me!

or without the help of companions from her old colony. She can lay two kinds of eggs: those fertilized by her mate and those from which she withholds fertilization. These unfertilized eggs get a single set of chromosomes and genes from the queen, none from any other source. They all grow up into males. The fertilized eggs develop into individuals with a set of chromosomes and genes from each parent, and they all become females. Most of these females grow up into sterile workers that stay with the queen and do everything for the colony except reproduce. All the workers are daughters of the queen, and none of these workers have offspring that carry their genes into future generations.

But the workers have superb adaptations for their way of life, just the sort of thing that inspired David Hume's admiration (see chapter 2). Can these adaptations be produced and maintained by natural selection? This process is normally thought to depend on the inheritance of those features that aided, directly or indirectly, the reproduction of ancestors. Worker ants or bees do not reproduce, and no worker can inherit an adaptation from any ancestral worker. This logical difficulty so impressed Darwin that he discussed the possibility that the adaptations of sterile workers must "at once annihilate the theory."

Obviously, if natural selection depends on the survival of the fittest for reproduction, it cannot explain the

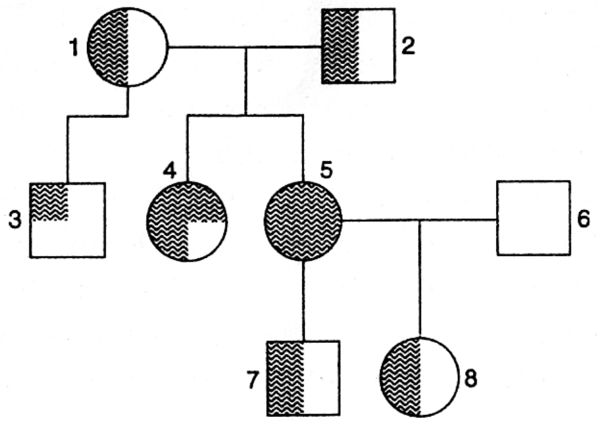
What a word!

!

adaptations of worker ants or bees. But suppose the process can be based on fitness for genetic success rather than reproduction. A worker does not reproduce, but her genes are present in varying proportions in her relatives. If those relatives are successful in reproducing, they are passing her genes to future generations. This argument applies with special force to the kind of life history I am discussing. If a worker were to produce a son or daughter, half her genes would go to each offspring. If her mother produces a daughter (the worker's sister), three-quarters of the worker's genes will go to that new female.

It is important to understand why. The worker's father, who fertilized the queen, had only one set of genes, because he came from an unfertilized egg. All his offspring get the same genes and are like identical twins in these genes from their father. Only the mother, who has two sets of genes from which she puts a randomized single set into each egg, provides any genetic variation. Hymenopteran sisters therefore have a three-quarter relation to one another, but only one-half to each offspring, if they produce any. For a worker, there is more payoff (genetic success) in her mother's reproduction than in her own. So a pedigree diagram for this group of insects looks a bit different from the previous one. The three-quarters genetic similarity between sisters is balanced by the one quarter relation of a brother to a sister. He shares none of

mea!



Pedigree diagram showing degrees of relationship for female bee or ant (#5) with various other individuals. Note especially the difference in relationship of a sister (#4) and a brother (#3).

the genes that the sister got from her father, because the brother has no father.

If this seems confusing, just remember this basic definition of genealogical relationship. Point your finger at some random gene on some chromosome in one individual and ask: What is the probability that this gene went from the same source to some other individual? This is the relationship of the other individual to the first. Such relationships can be asymmetrical in the Hymenoptera. A randomly selected gene in a female has a 25 percent probability of being in her brother, and this makes him 25 percent related to her. A randomly selected gene from the single set in a male has a 50 percent probability of being in his sister, so that she is 50 percent related to him. The especially close relationship between sisters is often used to explain why, in this insect order, females so regularly forgo their own reproduction and devote themselves to their mother's.

It has been pointed out, in confirmation of this theory, that advanced social organizations have evolved in this order of insects about a dozen times independently, and in every case the societies consist solely of females. Males (drones) make no contributions to the economy of the hive. They leave shortly after they emerge from their brood cells. The importance of the three-quarter relationship is confirmed by comparison with termites, another order of insects with advanced social organizations. Termites have normal sexual reproduction, with both males and females arising from fertilized eggs, and without the special relationship between sisters. So this special reason for societies being based entirely on females is absent in the termites. Exactly as expected, both sexes participate about equally in the workings of termite societies.

It would seem that the elaborate organization of a honeybee colony is an incidental consequence of each individual's efforts to maximize its own genetic success. It must also be pointed out that biologists differ widely in their acceptance of this simple picture, because there are many complications and worrisome details. All the Hymenoptera have a 75 percent relationship between sisters, but many do not have complex societies based on sterile workers. So while the special genetic relationships in this order of insects may make the evolution of complex societies more likely, they are neither necessary nor sufficient. The mothers in nonsocial species raise their own young with complex nurturing behaviors and no help from males. If they were to evolve elaborate societies, would you not expect them to be based on females? The exclusion of males from the social life of the colony would simply be an extension of their ancestral exclusion from parental roles. In many social species, workers may only be half sisters, because the queen mated with more than one male. Some of her daughters may have the same father, some different fathers. A complication to this complication is that daughters of multiply mated queens produced at the same time may be mainly from the same father.

One of the great delights of scholarly pursuits such as biology is that we can all form our own opinions on any issue. There is clear consensus about many important questions, and it is wise to follow it when there is no reason to challenge it, but the jury is still out on just how socially important the 75 percent relationship between full sisters is in the Hymenoptera. But there is a strong consensus on the most basic issue: each individual social insect, no matter how large and complexly organized its society may be, can be regarded as maximally devoted to getting its own genes into future generations. It pursues this goal by adaptive choice of options open to it.

The tactics required for a honeybee's genetic success and that of a typical mammal are radically different. A mammal's adaptations mainly serve its individual interests through that individual's reproduction. A honeybee's adaptations are mainly individual features that serve that individual's genetic interests through its mother's reproduction, which depends on the survival and reproduction of the colony. Such quantitative measures as the amount of time a bee forages for nectar would be optimized in relation to the interests of the colony, not the forager. Many observations indicate that this is so, and give the consistent impression that the colony is a tightly organized team with all members devoted exclusively to group interests. The intricacies of the adaptive organization of the social insects have suggested, to many biologists, the term *superorganism* for entities such as honeybee colonies.

THE SCARCITY OF INDIVIDUAL SUBORDINATION TO GROUP INTERESTS

The same kind of analysis should be used to identify the purpose of the adaptations of any other organism or group of organisms. There are natural-history accounts for [illegible] may

general audiences that tell us that a salmon leaves its rich hunting grounds in the ocean and migrates up rivers to a remote mountain stream in order to propagate its kind, even though that surely means its own death. Is that what a spawning salmon is doing, sacrificing its life for the continuation of its species? Its spawning may perpetuate its genes, but surely it also perpetuates its species. Does it matter which we think of as its real purpose?

!

It does indeed, because an examination of the details of its effort supports one conclusion and refutes the other. When a female spawns, she puts her own eggs safely under a layer of gravel in some spot chosen for its special suitability for salmon egg development. In so doing, she may be dislodging previously laid eggs, which then have very little chance of survival. If she were to avoid a previously used area and settle for a slightly less favorable location, the total productivity of young salmon would be increased, even if some other female would have a bit more genetic success than she would. Every action by a spawning female salmon is just what we would expect for an effort to maximize her own genetic success, regardless of its effect on group productivity.

The males are even more obviously competitive. They fight ferociously with each other for the opportunity to fertilize eggs laid by females. One male is adequate to fertilize almost all the eggs of many females, so why the violent and exhausting effort? Because the whole point of the reproductive behavior of a male salmon is to win for itself the greatest attainable proportion of the fertilizations of eggs produced by the local females. Everything it does is obviously a part of this effort to do better than competing males. Nothing it does suggests that it is trying to maximize the number of young salmon produced, or their collective rate of survival, or any other result that might be seen to serve the general welfare of its population.

Another comparison is particularly apt in identifying what Aristotle's "complex action" must really be, even if Aristotle missed it. What happens when (1) an independent individual, plant or animal, is threatened with death, (2) when a honeybee colony is threatened with destruction, and (3) when a population is threatened with extirpation? The answers to the first two questions are the same: the individual or the colony takes emergency measures. An animal fights back, or flees, or hides in a retreat. A plant's response is less obvious, but a plant suddenly attacked by a large number of chewing insects will often change its metabolism so as to devote resources less to growth, more to making defensive toxins.

The bee colony likewise fights back, and may do so in ways that make its functional unity particularly clear. Individual bees may not only risk their lives but actively sacrifice them for the good of the colony. When a bee stings, the sting may come out and take on an existence of interessant: hoe zit dat, met mensen?

its own, actively pumping venom into the stung animal even if the bee itself is brushed away. The breaking loose of the sting always kills the bee, but this in no way reduces the enthusiasm of its attack. It thus clearly shows its priorities: my life is worth very little; it is the survival of my colony that matters, because only thus can my genes survive.

What does a population do when threatened with extirpation, for instance, the reduction of the sockeye salmon stock of the Yukon River to 1 percent of its normal size? Nothing special happens at all. The individual salmon keep on with their normal activities, each trying to reproduce more than its neighbors, with no regard to effects on the stock as a whole. Individual salmon respond to individual threats in adaptive ways, but salmon populations take no concerted action to avoid being wiped out. Their populations show no functional organization like that of a bee colony.

FOR THE HARM OF THE SPECIES

The absence of a functional organization for populations or species actually has worse consequences than might be imagined, because natural selection within the groups can produce results that not only fail to help the group as a whole but may be harmful, or at least systematically

wasteful. One example is the effect of selection on a population-sex ratio. The really basic question of why reproduction is often sexual is explored in chapter 5. Here I will just assume that it is, and is made up of males and females rather than hermaphrodites. What then determines what fraction of the total is male and what fraction female?

The history of thought on this question is rather curious. Darwin worried about it a little, but shrugged it off. It was then totally ignored until 1930, when Ronald A. Fisher, one of the patriarchs of modern Darwinism, tersely provided the essence of the currently accepted idea. This explanation invokes what is now known as frequency-dependent selection, a rather elementary idea in traditional economic reasoning. Suppose you are equally skilled as a smith and a carpenter, and have the resources to set up shop in a village for one trade but not both. You know that the village already has a smith. Should this influence your decision? A well-known example of the same principle, applied in fact to sex ratio, is provided by Shakespeare in *The Taming of the Shrew*. Think of the trouble Baptista would have saved himself if he had fathered a son and a daughter instead of Bianca and Katharina. Minimizing your children's competition for mates is a good idea if you want to maximize your production of grandchildren.

Problems in frequency-dependent selection in biology

		opponent	
		male	female
player	male	0	1
	female	1	0

are often analyzed by what is known in game theory as a payoff matrix, and sex ratio provides the simplest possible example. In the illustration, the individual labeled *player* has a choice of being a male or a female. The next individual it encounters, labeled *opponent*, will be one or the other. Whichever our player chooses to be, if the opponent turns out to be of the opposite sex, it wins in its game of reproduction. The winning is represented by the score of *1* in the matrix. If the opponent is of the same sex, there is no reproduction, no winning, score *0*.

What advice do you give the player in this kind of game? Unless you know something about the sex of the opponent, there can be nothing wiser than flipping a coin. But suppose you know that the males are slightly in the majority in the population. Now you can offer sound advice: be a female; or, if it is too late for that, have a daughter, not a son. A closer-to-home example: you are a lecherous heterosexual man and there are two singles' bars in town. One of them has mostly men in it, the other mostly women. Which will you choose?

Selection on sex ratio operates by favoring any individual who becomes a member of the minority sex or produces mostly that sex among its offspring. The expected immediate result of this selection is to increase the abundance of the minority sex. The long-term result is that the minority sex stops being in the minority and the sex ratio stabilizes at equal numbers. An example is found in the approximately equal numbers of men and women in the world. As long as this condition prevails, men will, collectively and on the average, produce about the same number of babies as women do. Neither sex has an advantage, and selection on sex ratio disappears. It will reappear and act to reestablish the nearly equal numbers if ever this equilibrium is disturbed.

All sorts of questions will crop up at this point: At what age do we expect equal numbers of males and females? What is the effect of sons and daughters having different mortality rates or different costs to the parents? What if males and females mature at different ages? Why is there, in fact, a slight preponderance of boys at birth? These are questions that biologists have discussed at great length, but their answers imply quite minor quantitative modifications of the 50:50 ratio expected from simple frequency-dependent selection. This selection always favors the minority sex. This is true regardless of how it affects the well-being of the group as a whole.

And the effect can be decidedly negative. Richard

Dawkins, in his book *River Out of Eden*, discusses the dramatic example of elephant seals. In the breeding season, the females come ashore on suitable beaches to give birth and nurse their babies and be fertilized for next year's births. The beaches are crowded by adults, a scattering of individual males each with a large harem of females. This is not the adult sex ratio; it is merely the sex ratio of reproducing adults. The true ratio is not far from equal. This means that most of the males are unsuccessful. For every male with a harem, there are many celibate bachelors. They represent a waste of resources, because only a small fraction of them will reproduce. Yet because of frequency-dependent selection, the population goes on, generation after generation, producing about the same number of males as females.

Actually it is worse than would be inferred from just the equality of numbers. There is a gross difference in size, the males being far larger than the females. This is because only the biggest and strongest males have any hope of winning mates against the fierce competition from other males. So frequency-dependent selection keeps the population producing a wastefully large number of males, and sexual selection goes on making each male wastefully large. A successful male, over his lifetime, consumes far more food than does a successful female, who bears the entire physiological burden of reproduction for both her own and her mate's genes. Our

own species is afflicted with the same difficulties, but fortunately not to the same extent. The recent feminist slogan "men are not cost-effective" is entirely correct biologically: there are too many of them; they are too big; they accomplish less than women per unit of resources consumed.

The formal game of prisoners' dilemma, for which traders' dilemma would be more apt a name, has a payoff matrix rather different in form from the one in the previous illustration, but shows another way in which natural selection may have negative effects at the group level. Suppose some evening you are driving a car with a German license plate through a little town in Switzerland to your home in Italy. You notice a store that sells computer supplies, and you remember that you need diskettes. You stop because the store is selling for ten marks what would cost you twenty at home. So you are willing to pay the ten marks and be a winner by ten marks, according to your evaluation of the goods. But wait—you might do even better. You happen to have some worthless counterfeit marks. It is too dark in the store now to notice, and the storekeeper won't see until morning that the money is not real. By then you will be in another country, and it is unlikely that you will ever be back here again. Your situation is that described in the matrix. *H* represents the strategy of being honest, and paying real money, *D* the dishonest use of the counterfeit.

		opponent	
		H	D
player	H	10	-10
	D	20	0

What should you do? Obviously, if self-interest is your only motivation, you should be dishonest. That way you get the twenty marks' worth of goods for nothing. If you paid real money, you would give up ten marks and have a net gain of only ten.

Unfortunately, you are not the only player with a dishonest option. The storekeeper has diskettes that he knows to be grossly defective. No buyer will realize this before getting them home and trying to use them. The storekeeper notices your German car and your Friulano accent and thinks, "I'll never meet this guy again. Why waste good merchandise when I can foist off this worthless box?" What should he do? Again, self-interest provides one clear answer: be dishonest! The net expectation from all this rational decision making should be clear from the payoff matrix. If everyone in such a game were consistently honest, everyone would win ten marks per game. But what happens in such an honest society if a cheater appears? The cheater's winnings are greater, and some honest player is penalized. What happens if all players cheat? No one wins. The storekeeper gets worthless counterfeit money, the traveler some worthless diskettes. Yet dishonesty remains the best policy for everyone, because of a simple rule apparent from the payoffs. *No matter what your opponent does, you do better by cheating.* So natural selection proceeds to establish this rule, and reduces the payoff for everyone to zero from a potential positive gain.

I will mention only one of many possible biological applications of the traders' dilemma. There is often an optimal group size for its members, for instance, the number of fish in a school in a pond. When a predator attacks, it is likely that no more than one fish will die, because, once one is caught, the others can get away. This means that if there are *m* fish in the school, the probability is 1/m that a given fish will be the next victim. Obviously the safest place to be is in the largest school available. Unfortunately, the bigger the school the greater the competition for food and the less there will be for each fish. The optimum school size will be that with the greatest excess of benefit, from predator avoidance, over cost, from decreased nutrition.

Suppose ten is the optimum number, and a fish finds itself in a school of twenty. What should it do, purely from the standpoint of self-interest? If it stays, it and all the others will suffer from a food shortage. If it leaves, it would help all the others by increasing their nutrition, but it would expose itself to much greater risk of death from predation. It could well be that, from the standpoint of its long-term fitness considerations, it is better to keep its risk at 1/20 for the next predator attack and make do with a deficient diet. If so, the twenty fish will continue to swim together, even though, from every individual's perspective, it would be better to break up into two schools of ten.

[illegible]!

A human level of rationality in this situation might well result in the two optimal groups. An individual could assume the lead and say, "Look, fellas, there are too many of us. Let's all us on the left side of the school turn left, and you guys on the right head the other way. Then we will achieve the optimum trade-off between predation hazard and competition for food." Unfortunately, the only decision making the fish can manage is a simple "I should stay in this bad situation" versus "I should move away into a worse situation." This inevitably, in the dilemma described, results in schools that are too big, and at least one study suggests that this happens regularly in nature. An enormous number of other examples could be described that would illustrate the principle that, although some groups, such as honeybee colonies, are functionally organized, most animal groupings are not. They are just mobs of self-seeking individuals. In the next chapter I return to the examination of evolved mechanisms, with emphasis on two related questions: How are

they produced (development), and how do they work (physiology)? I also consider the more fundamental question of how such problems are legitimately resolved.

CENTRAL THEMA !!

Volume 82, No. 4 December 2007

The Quarterly Review
of Biology

RETHINKING THE THEORETICAL FOUNDATION OF SOCIOBIOLOGY

David Sloan Wilson
Departments of Biology and Anthropology, Binghamton University
Binghamton, New York 13902 USA
e-mail: dwilson@binghamton.edu

Edward O. Wilson
Museum of Comparative Zoology, Harvard University
Cambridge, Massachusetts 02138 USA

keywords
altruism, cooperation, eusociality, group selection, human evolution, inclusive fitness theory, kin selection, major transitions, multilevel selection, pluralism, sociobiology

abstract

Current sociobiology is in theoretical disarray, with a diversity of frameworks that are poorly related to each other. Part of the problem is a reluctance to revisit the pivotal events that took place during the 1960s, including the rejection of group selection and the development of alternative theoretical frameworks to explain the evolution of cooperative and altruistic behaviors. In this article, we take a "back to basics" approach, explaining what group selection is, why its rejection was regarded as so important, and how it has been revived based on a more careful formulation and subsequent research. Multilevel selection theory (including group selection) provides an elegant theoretical foundation for sociobiology in the future, once its turbulent past is appropriately understood.

The Quarterly Review of Biology, December 2007, Vol. 82, No. 4

0033-5770/2007/8204-0001$15.00

DARWIN perceived a fundamental problem of social life and its potential solution in the following famous passage from *Descent of Man* (1871:166):

> It must not be forgotten that although a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe . . . an increase in the number of well-endowed men and an advancement in the standard of morality will certainly give an immense advantage to one tribe over another.

The problem is that for a social group to function as an adaptive unit, its members must do things for each other. Yet, these group-advantageous behaviors seldom maximize relative fitness within the social group. The solution, according to Darwin, is that natural selection takes place at more than one level of the biological hierarchy. Selfish individuals might out-compete altruists within groups, but internally altruistic groups out-compete selfish groups. This is the essential logic of what has become known as multilevel selection theory.

Darwin's insight would seem to provide an elegant theoretical foundation for sociobiology, but that is not what happened, as anyone familiar with the subject knows. Instead, group selection was widely rejected in the 1960s and other theoretical frameworks were developed to explain the evolution of altruism and cooperation in more individualistic terms. The following passage from George C Williams's book, *Adaptation and Natural Selection* (1966:92–93), illustrates the tenor of the times, which seemed to make the rejection of group selection a pivotal event in the history of evolutionary thought:

> It is universally conceded by those who have seriously concerned themselves with this problem . . . that such group-related adaptations must be attributed to the natural selection of alternative *groups* of individuals and that the natural selection of alternative alleles within populations will be opposed to this development. I am in entire agreement with the reasoning behind this conclusion. Only by a theory of between-group selection could we achieve a scientific explanation of group-related adaptations. However, I would question one of the premises on which the reasoning is based. Chapters 5 to 8 will be primarily a defense of the thesis that group-related adaptations do not, in fact, exist. A *group* in this discussion should be understood to mean something other than a family and to be composed of individuals that need not be closely related.

Forty years later, this clarity has been lost. In the current sociobiological literature, it is easy to find the following contradictory positions, side by side in the same journals and bookshelves:

- Nothing has changed since the 1960s.
- Multilevel selection theory (including group selection) has been fully revived.
- There is a "new" multilevel selection theory that bears little relationship to the "old" theory.
- Group selection is not mentioned, as if it never existed in the history of evolutionary thought.

Part of this confusion can be explained in terms of the diffusion of knowledge. It takes time for the newest developments in theoretical biology to reach scientists who conduct empirical research, and longer still to reach diverse audiences who receive their information third, fourth, and fifth hand. However, part of the confusion continues to exist at the highest level of scientific discourse, as we will show.

We think that sociobiology's theoretical foundation can be as clear today as it appeared to be in the 1960s, but only if we go back to the beginning and review the basic logic of multilevel selection, what appeared to be at stake in the 1960s, and why the original rejection of group selection must be reevaluated on the basis of subsequent research. Everyone can benefit from this "back to basics" approach, from the most advanced theorists to students learning about sociobiology for the first time.

A WORD ABOUT TAINTED WORDS

It is a natural human tendency to avoid associating oneself with people or ideas that have acquired a bad reputation in the past. Thus, there are evolutionists who study social behavior, but avoid the term "sociobiology," or who study psychology, but avoid the term "evolutionary psychology," because of particular ideas that were associated with these terms in the past, including their supposed political implications. At a broader scale, there are people who avoid the word "evolution" because of past negative associations, even though they are clearly talking about evolutionary processes. We think that this very understandable temptation needs to be resisted in the case of scientific terminology, because the short-term gain for the user (avoiding negative associations) results in long-term confusion for the field as a whole (a proliferation of terms that mean the same thing). The problem has been especially severe for multilevel selection theory because many evolutionists have felt that their very careers would be jeopardized if they invoked group selection. In some cases, their fears were well founded; we could provide numerous examples of colleagues whose articles and grant proposals were rejected when stated in terms of multilevel selection theory, and then accepted when restated using other terms. In this article, we define our terms at face value, regardless of past associations: sociobiology is the study of social behavior from a biological perspective, group selection is the evolution of traits based on the differential survival and reproduction of groups, and so on. Returning to face-value definitions is a first step toward resolving the confusion that plagues the modern sociobiological literature (see also Foster et al. 2007).

From an evolutionary perspective, a behavior can be regarded as social whenever it influences the fitness of other individuals in addition to the actor. Social behaviors need not be *pro*social; aggression fits the definition as does cooperation. Also, the interactions need not be direct; a feeding behavior that reduces the fitness of others by depleting their resources counts as social. Even genetic and developmental interactions within a single individual can be regarded as social, since the organisms of today are now known to be the social groups of past ages, as we will describe in more detail below. Narrower definitions of social behavior might be useful for some purposes, but the important point to keep in mind is that the concepts reviewed in this article apply to any trait that influences the fitness of others in addition to the actor, regardless of how "social" these traits might appear in the intuitive sense.

The History and Basic Logic of Multilevel Selection Theory

During evolution by natural selection, a heritable trait that increases the fitness of others in a group (or the group as a whole) at the expense of the individual possessing the trait will decline in frequency within the group. This is the fundamental problem that Darwin identified for traits associated with human morality, and it applies with equal force to group-advantageous traits in other species. It is simply a fact of social life that individuals must do things for each other to function successfully as a group, and that these actions usually do not maximize their relative fitness within the group.

Why is there usually a tradeoff? Because there is usually a tradeoff between *all* adaptations. Antipredator adaptations usually interfere with harvesting food, adaptations for moving through one medium (such as the air) usually interfere with moving through another medium (such as the water), and so on. The same principle applies to adaptations for functioning as a team player in a well-coordinated group, compared to maximizing one's relative fitness within the group. This does not mean that the tradeoff must necessarily be severe. Benefiting others or one's group as a whole does *not* invariably require extreme self-sacrifice, such as rushing into a burning house to save a child, but it *does* require some set of coordinating mechanisms, such as organizing and paying for a fire department, passing and enforcing fire safety legislation, and so on. It is unlikely that these coordination mechanisms evolve as a coincidental byproduct of organisms that are adapted exclusively to survive and reproduce better than other members of their same group. That is why Darwin felt confident in saying that "a high standard of morality gives but a slight or no advantage to each individual man and his children *over the other men of the same tribe.*" As for human morality, so also for group-level adaptations in all species.

Something more than natural selection

within single groups is required to explain how altruism and other group-advantageous traits evolve by natural selection. For Darwin, in the passage quoted above, that "something" was between-group selection. Group-advantageous traits do increase the fitness of groups, relative to other groups, even if they are selectively neutral or disadvantageous within groups. Total evolutionary change in a population can be regarded as a final vector made up of two component vectors, within- and between-group selection, that often point in different directions.

The basic logic of multilevel selection applies to an enormous range of social behaviors, including the evolution of sexual reproduction and sex ratio, distastefulness in insects, prudent use of resources, warning others about predators, social insect colonies as superorganisms, and more. The relevant groupings are equally diverse, from a social insect colony (as a superorganism) or an ephemeral flock of birds (for warning calls), to multigenerational groups (for prudent use of resources), to entire species and clades (for sexual reproduction). Two related themes give these examples conceptual unity. First, single traits can evolve despite being locally disadvantageous wherever they occur. For this to happen, an advantage at a larger scale (between groups) must exist to counteract the disadvantage at a smaller scale (within groups). Second, a higher-level unit (such as a social insect colony) can become endowed with the same adaptive properties that we associate with single organisms. There can be such a thing as a superorganism. D S Wilson (1997) referred to these themes as "altruism" and "organism." They are closely related but not entirely overlapping, since becoming a superorganism involves more than the evolution of a single trait.

Evolutionary theory was placed on a mathematical foundation by the first population geneticists, in particular Ronald Fisher, Sewall Wright, and J B S Haldane. Each considered the problem of multilevel selection, but only briefly, because it was not the most important issue compared to even more foundational issues such as the consequences of Mendelian genetics (reviewed by Sober and D S Wilson 1998). All three men shared Darwin's perception that group-advantageous traits seldom maximize relative fitness within groups, thereby requiring a process of between-group selection to evolve. Unfortunately, many other biologists did not share this insight and uncritically assumed that adaptations evolve at all levels of the biological hierarchy without requiring a corresponding level of selection. When the need for between-group selection was acknowledged, it was often assumed that between-group selection easily trumped within-group selection. The following passage from the textbook *Principles of Animal Ecology* (Allee et al. 1949:729) illustrates what became known in retrospect as "naïve group selectionism":

> The probability of survival of individual living things, or of populations, increases with the degree with which they harmoniously adjust themselves to each other and to their environment. This principle is basic to the concept of the balance of nature, orders the subject matter of ecology and evolution, underlies organismic and developmental biology, and is the foundation for all sociology.

Another naïve group selectionist was V C Wynne-Edwards, who proposed that organisms evolve to assess and regulate their population size to avoid overexploiting their resources in his book, *Animal Dispersion in Relation to Social Behavior* (Wynne-Edwards 1962, 1986). He was aware that group selection would be required and would often be opposed by selection within groups, but he assumed that group selection would usually prevail and proceeded to interpret a vast array of animal social behaviors according to his thesis without evaluating the levels of selection in any particular case.

These issues began to occupy center stage among evolutionary biologists in the 1960s, especially under the influence of George C Williams's (1966) *Adaptation and Natural Selection.* Williams began by *affirming* the importance of multilevel selection as a theoretical framework, agreeing with Darwin and the population geneticists that group-level adaptations require a process of group-level selection. He then made an additional claim that between-group selection is almost invariably weak compared to within-group selection (both positions are represented in the above-quoted passage). It was this additional claim that turned multilevel selection theory into what became known as "the theory of individual selection." Ever since, students have been taught that group selection is possible in principle, but can be ignored in practice. Seemingly other-oriented behaviors must be explained as forms of self-interest that do not invoke group selection, such as by helping one's own genes in the bodies of others (kin selection), or by helping others in expectation of return benefits (reciprocity). The concept of average effects in population genetics theory, which averages the fitness of alleles across all genotypic, social, and environmental contexts, was elaborated by both Williams and Richard Dawkins (1976) into the "gene's eye view" of evolution, in which everything that evolves is interpreted as a form of "genetic selfishness."

The rejection of group selection in the 1960s was based on three arguments, like the legs of a stool: a) group selection as a *significant* evolutionary force is theoretically implausible; b) there is no solid empirical evidence for group selection as a distinctive, analytically separable process; and c) alternative theories can explain the evolution of apparent altruism without invoking group selection. In the following sections, we will show that all three arguments have failed, based on subsequent research. If this information had been available to Williams and others in the 1960s, the history of sociobiology would have headed in a completely different direction. The component vectors of within- and between-group selection would need to be calculated on a case-by-case basis to determine the final vector of evolutionary change in the total population. Traits could legitimately be regarded as "for the good of the group" whenever they evolve by group selection, in the same sense that an individual-level adaptation (such as the eye) is regarded as "for the good of the individual." Instead, sociobiology proceeded along a seemingly triumphant path based entirely on the calculus of individual and genetic self-interest, under the assumption that group selection can be categorically ignored. It is precisely this branch point that must be revisited to put sociobiology back on a firm theoretical foundation.

The Theoretical Plausibility of Group Selection as a Significant Evolutionary Force

The rejection of group selection was based largely on theoretical plausibility arguments, which made it seem that between-group selection requires a delicate balance of parameter values to prevail against within-group selection. These early models were published at a time when the desktop computing revolution, the study of complex interactions, and appreciation of such things as social control (e.g., Ratnieks and Visscher 1989; Boyd and Richerson 1992) and gene-culture coevolution (Lumsden and E O Wilson 1981; Boyd and Richerson 1985; Richerson and Boyd 2005) were barely on the horizon. It should surprise no one that the initial assessment must be revised on the basis of four decades of subsequent research.

All of the early models assumed that altruistic and selfish behaviors are caused directly by corresponding genes, which means that the only way for groups to vary *behaviorally* is for them to vary *genetically*. Hardly anyone regards such strict genetic determinism as biologically realistic, and this was assumed in the models primarily to simplify the mathematics. Yet, when more complex genotype-phenotype relationships are built into the models, the balance between levels of selection can be easily and dramatically altered. In other words, it is possible for modest amounts of genetic variation among groups to result in substantial amounts of heritable phenotypic variation among groups (D S Wilson 2004).

The early models also assumed that variation among groups is caused primarily by sampling error, which means that it declines precipitously with the number of individuals that independently colonize each group and migration among groups during their existence. This assumption must be revised on the basis of agent-based models. When individual agents interact according to biologically plausible decision rules, a spatial patchiness emerges that has little to do with sampling error (e.g., Johnson and Boerlijst 2002; Pepper and Smuts 2002; Pepper 2007). An example is a recent simulation model on the kind of social signaling and population regulation envisioned by Wynne-Edwards (Werfel and Bar-Yam 2004). Individuals create a local signal when crowded and curtail their reproduction accordingly. Their base reproductive rate and response to the signal are allowed to vary as independent continuous traits, including "cheaters" who reproduce at the maximum rate and ignore the signal altogether. Interactions occur on a two-dimensional lattice in which each cell represents an area occupied by the resource alone, both the resource and consumers, or by neither. Consumers that reproduce at the maximum rate are selectively advantageous within groups, but tend to drive their resource (and, therefore, themselves) extinct, exactly as envisioned by Wynne-Edwards and the early group selection models. More prudent consumers are maintained in the total population by spatial heterogeneity, which emerges spontaneously on the basis of complex interactions among the various traits. The local disadvantage of curtailed reproduction does not entirely determine the outcome of selection in the total population. In general, complex social and ecological interactions, coupled with limited dispersal, result in a kind of spatial heterogeneity that is far outside the envelope conceived by earlier models based on sampling error in the absence of complex interactions (see also Gilpin 1975; Avilés et al. 2002; Aktipis 2004).

Another early conclusion was that group selection is weak for groups that last for multiple generations, because the "generation time" is greater for groups than for individuals. Three examples will show how this conclusion has been overturned by subsequent theoretical models. First, even though altruists decline in frequency within each group and ultimately go extinct after a sufficient number of generations, the differential fitness of groups also increases with each generation, especially when the groups grow exponentially at a rate determined by the frequency of altruists. Simulations show that group selection can remain a significant force even when the groups last 10 or 15 generations between dispersal episodes (D S Wilson 1987; Avilés 1993). Second, Gilpin (1975) showed that when predator/prey dynamics are nonlinear, a small increase in predator consumption rate can have a large effect on extinction rates, causing group selection to be effective in multiple-generation groups. Third, Peck (2004) modeled altruism and selfishness as suites of traits that must occur in the right combination to function correctly, rather than as single traits. In this case, when a selfish individual migrates into an altruistic group, its genes do not spread because they become dissociated by sexual reproduction and no longer occur in the right combination. An altruistic group can persist indefinitely, replacing less altruistic groups when they go extinct. These and other examples do not imply that group selection is *invariably* effective in multigenerational groups, but they do overturn the earlier conclusion that group selection can be categorically ignored.

Acknowledging the theoretical plausibility of group selection as a significant evolutionary force is not a return to the bad old days of naïve group selectionism. It has always been the goal of population genetics theory to provide a complete accounting system for evolutionary change, including selection, mutation, drift, and linkage disequilibrium. The question is whether group selection can be categorically ignored when natural selection is separated into within- and between-group components. Few theoretical biologists would make this claim today, however reasonable it might have appeared in the 1960s. Yet, these developments have not resulted in an appropriately revised theory, even among some of the theorists, nor have they spread to the wider community of scientists interested in the evolution of social behavior. There is a form of naïve selectionism that needs to be corrected, as before the publication of *Adaptation and Natural Selection,* but today it is the naïve assumption that group selection can be consistently ignored.

Empirical Evidence for Group Selection

The rejection of group selection in the 1960s was not based upon a distinguished body of empirical evidence. Instead, Williams (1966) used the theoretical implausibility of group selection as a significant evolutionary force to argue that hypotheses framed in terms of individual selection are more parsimonious and, therefore, preferable to hypotheses that invoke group selection. In this fashion, broad categories of behavior such as dominance and territoriality were interpreted individualistically on the basis of plausibility arguments, without careful measurements of within- versus between-group selection for particular traits in particular species. Parsimony can be a factor in deciding between alternative hypotheses, but it cannot substitute for an evaluation of the data (Sober and D S Wilson 1998; Sober 2008). No population geneticist would argue that drift is more likely than selection and no ecologist would argue that predation is more likely than competition on the basis of parsimony. These alternatives are all plausible and their relative importance must be determined empirically on a case-by-case basis. Similarly, the direction and strength of within- and between-group selection must be determined on a case-by-case basis if both are theoretically plausible.

The closest that Williams came to a rigorous empirical test was for sex ratio, leading him to predict that female-biased sex ratios would provide evidence for group selection. The subsequent discovery of many examples of female-biased sex ratios, as well as evidence of group selection in the evolution of disease organisms, brought him back toward multilevel selection in the 1990s (Williams and Nesse 1991; Williams 1992).

Some of the best recent evidence for group selection comes from microbial organisms, in part because they are such efficient systems for ecological and evolutionary research spanning many generations (Velicer 2003). The "wrinkly spreader (WS)" strain of *Pseudomonas fluorescens* evolves in response to anoxic conditions in unmixed liquid medium, by producing a cellulosic polymer that forms a mat on the surface. The polymer is expensive to produce, which means that nonproducing "cheaters" have the highest relative fitness within the group. As they spread, the mat deteriorates and eventually sinks to the bottom. WS is maintained in the total population by between-group selection, despite its selective disadvantage within groups, exactly as envisioned by multilevel selection theory (Rainey and Rainey 2003).

As another example, Kerr et al. (2006) created a metapopulation of bacteria (the resource) and phage (the consumer) by culturing them in 96-well microtiter plates. Migration between groups was executed by a high-throughput, liquid-handling robot according to a prespecified migration scheme. Biologically plausible migration rates enabled "prudent" phage strains to outcompete more "rapacious" strains, exactly as envisioned by Wynne-Edwards and subsequent theorists such as Gilpin (1975) and Werfel and Bar-Yam (2004). As Kerr et al. put it, "spatially restricted migration reduces the probability that phage reach fresh hosts, rendering rapacious subpopulations more prone to extinction through dilution. Consequently, the tragedy of the commons is circumvented at the metapopulation scale in the Restricted treatment" (2006:77). More generally, the well-established fact that reduced virulence often evolves by group selection in disease organisms (Bull 1994; Frank 1996) provides a confirmation of Wynne-Edwards's hypothesis—not for *all* species, but for at least *some* species.

Multilevel selection experiments in the laboratory have been performed on organisms as diverse as microbes, plants, insects, and vertebrates (Goodnight et al. 1992; Goodnight and Stevens 1997). Phenotypic variation among groups is usually considerable, even when the groups are founded by large numbers of individuals, as expected on the basis of the newer theoretical models. For example, microcosms colonized by millions of microbes from a single well-mixed source nevertheless become variable in their phenotypic properties within a matter of days. When microcosms are selected on the basis of these properties and used to colonize a new "generation" of microcosms, there is a response to selection (Swenson et al. 2000a,b).

Quantitative genetics models separate phenotypic variation into additive and nonadditive components, with only the former leading to a response to selection (narrow-sense heritability). Laboratory selection experiments show that the nonadditive component of variation within groups can contribute to the additive component of variation among groups, causing group-level traits to be more heritable than individual-level traits. For example, selecting plants within a single group on the basis of leaf area did not produce much response to selection, but selecting whole groups on the basis of leaf area produced a strong response to selection. This result makes sense theoretically when phenotypic traits such as leaf area are influenced by interactions among individuals within the group, rather than being directly coded by genes (Goodnight 2000, 2005).

Field studies of social vertebrates are seldom as precise as laboratory experiments but nevertheless provide convincing evidence for group selection. The following description of territorial defense in lions (Packer and Heinsohn 1996:1216; see also Heinsohn and Packer 1995) is virtually identical to Darwin's passage about human morality that began this article: "Female lions share a common resource, the territory; but only a proportion of females pay the full costs of territorial defense. If too few females accept the responsibilities of leadership, the territory will be lost. If enough females cooperate to defend the range, their territory is maintained, but their collective effort is vulnerable to abuse by their companions. Leaders do not gain 'additional benefits' from leading, but they do provide an opportunity for laggards to gain a free ride." In this field study, extensive efforts to find a within-group advantage for territorial defense failed, leaving between-group selection as the most likely—and fully plausible—alternative.

To summarize, four decades of research since the 1960s have provided ample empirical evidence for group selection, in addition to its theoretical plausibility as a significant evolutionary force.

Are There Robust Alternatives to Group Selection?

Inclusive fitness theory (also called kin selection theory), evolutionary game theory (including the concept of reciprocal altruism), and selfish gene theory were all developed explicitly as alternatives to group selection. In addition to these major theoretical frameworks, there are numerous concepts such as indirect reciprocity (Nowak and Sigmund 2005; Nowak 2006), byproduct mutualism (Dugatkin 2002; Sachs et al. 2004), and costly signaling (Lachmann et al. 2001; Cronk 2005) that claim to explain the evolution of cooperation and altruism without invoking group selection. Nevertheless, *all* evolutionary models of social behavior share certain key features, no matter what they are called. Recognizing the similarities can go a long way toward establishing theoretical unity for the field.

First, all models assume the existence of multiple groups. Why? Because social interactions almost invariably take place among sets of individuals that are small compared to the total population. No model can ignore this biological reality. In N-person game theory, N refers to the size of the group within which social interactions occur. In kin selection theory, r specifies that individuals are interacting with a subset of the population with whom they share a certain degree of genealogical, genetic, or phenotypic similarity (depending upon the specific formulation), and so on. The groups need not have discrete boundaries; the important feature is that social interactions are *local,* compared to the size of the total population.

Second, all models must converge on the same definition of groups for any particular trait. Why? Because all models must calculate the fitness of individuals to determine what evolves in the total population. With social behaviors, the fitness of an individual depends upon its own phenotype and the phenotypes of the others with whom it interacts. These other individuals must be appropriately specified or else the model will simply arrive at the wrong answer. If individuals interact in groups of $N = 5$, two-person game theory will not do. Evolutionary theories of social behavior consider many kinds of groups, but that is only because they consider many kinds of traits. For any particular trait, such as intergroup conflict in humans, mat formation in bacteria, or territorial defense in lions, there is an appropriate population structure that must conform to the biology of the situation, regardless of what the theoretical framework is called. That is the concept of the *trait-group* (D S Wilson 1975); the salient group (and other aspects of population structure) for any particular trait.

Third, in virtually all cases, traits labeled cooperative and altruistic are selectively disadvantageous within the groups and require between-group selection to evolve. W D Hamilton (1975) realized this property of inclusive fitness theory when he encountered the work of George Price in the early 1970s (Price 1970, 1972). Price had derived an equation that partitions total gene frequency change into within- and between-group components. When Hamilton reformulated his theory in terms of the Price equation, he realized that altruistic traits are selectively disadvantageous within kin-groups and evolve only because kin-groups with more altruists differentially contribute to the total gene pool. Hamilton's key insight about the importance of genetic relatedness remained valid, but his previous interpretation of inclusive fitness theory *as an alternative to group selection* was wrong, as he freely acknowledged (Hamilton 1996:173–174; Schwartz 2000). The importance of genetic relatedness can be explained in terms of the parameters of multilevel selection, rather than requiring additional parameters (Michod 1982). For example, genetic relatedness might be an important factor in the evolution of territorial defense in lions, but only because it increases genetic variation among groups, thereby increasing the importance of between-group selection compared to within-group selection. Much the same conclusion has been drawn from social insects (e.g., Queller 1992; Bourke and Franks 1995; Wenseleers et al. 2003), as we will describe in more detail below.

For two-person game theory, the cooperative tit-for-tat strategy never beats its social partner; it only loses or draws. The only reason that tit-for-tat and other cooperative strategies evolve in a game theory model is because groups of cooperators contribute more to the total gene pool than groups of noncooperators, as Anatol Rapoport (1991) clearly recognized when he submitted the tit-for-tat strategy to Robert Axelrod's famous computer simulation tournament. The pairs of socially interacting individuals in two-person game theory might seem too small or ephemeral to call a group (Maynard Smith 2002), but the same dynamic applies to N-person game theory as a whole, including large and persistent groups that are described in terms of evolutionary game theory, but which overlap with traditional group selection models. All of these models obey the following simple rule, regardless of the value of N, the duration of the groups, or other aspects of population structure: *Selfishness beats altruism within single groups. Altruistic groups beat selfish groups.* The main exception to this rule involves models that result in multiple local equilibria, which are internally stable by definition. In this case, group selection can favor the local equilibria that function best at the group level, a phenomenon sometimes called "equilibrium selection" (Boyd and Richerson 1992; Samuelson 1997; Gintis 2000; the model by Peck 2004 described earlier provides an example).

Dawkins (1976, 1982) envisioned selfish gene theory and the concept of extended phenotypes as arguments against group selection but, in retrospect, they are nothing of the sort. The concept of extended phenotypes notes that genes can have effects that extend beyond the body of the individual, such as a beaver dam. Genes that cause beavers to build dams are still at a local disadvantage compared to genes in beavers in the same pond that do not build dams, so the concept of extended phenotypes does nothing to prevent the fundamental problem of social life or to provide a solution other than that provided by between-group selection. The concept of genes as "replicators" and "the fundamental unit of selection" is identical to the concept of average effects in population genetics, which averages the fitness of alleles across all genotypic, environmental, and social contexts. The average effect provides the bottom line of what evolves in the total population, the final vector that reflects the summation of all the component vectors. The whole point of multilevel selection theory is, however, to examine the *component vectors* of evolutionary change, based on the targets of selection at each biological level and, in particular, to ask whether genes can evolve on the strength of between-group selection, despite a selective disadvantage within groups. Multilevel selection models calculate the average effects of genes, just like any other population genetics model, but the final vector includes both levels of selection and, by itself, cannot possibly be used as an argument against group selection. Both Williams (1985:8) and Dawkins (1982:292–298) eventually acknowledged their error (reviewed in D S Wilson and Sober 1998; see also Okasha 2005, 2006), but it is still common to read in articles and textbooks that group selection is wrong because "the gene is the fundamental unit of selection."

A similar problem exists with evolutionary models that are not explicitly genetic, such as game theory models, which assume that the various individual strategies "breed true" in some general sense (Maynard Smith 1982; Gintis 2000). The procedure in this case is to average the fitness of the individual strategies across all of the social groupings, yielding an average fitness that is equivalent to the average effect of genes in a population genetics model. Once again, it is the final vector that is interpreted as "individual fitness" and regarded as an argument against group selection, even though the groups are clearly defined and the component vectors are there for all to see, once it is clear what to look for.

To summarize, all of the theories that were developed as alternatives to group selection assume the basic logic of multilevel selection within their own frameworks.

Pluralism

The developments outlined above have led to a situation that participants of the controversy in the 1960s would have difficulty recognizing. The theories that were originally regarded as alternatives, such that one might be right and another wrong, are now seen as equivalent in the sense that they all correctly predict what evolves in the total population. They differ, however, in how they partition selection into component vectors along the way. The frameworks are largely intertranslatable and broadly overlap in the kinds of traits and population structures that they consider. To make matters more confusing, each major framework comes in a number of varieties (e.g., Fletcher and Zwick 2006; Okasha 2006; West et al. 2007; D S Wilson 2007a). Considerable sophistication is required to interpret the meanings of terms such as "altruism," "selfishness," "relatedness," and "individual selection," depending upon the specific model being employed.

This kind of pluralism is a mixed blessing. On the positive side, multiple perspectives are helpful for studying any complex problem, so long as they are properly related to each other (Sober and D S Wilson 2002; Foster 2006). On the negative side, it is easy to lose sight of the fundamental issues that made the rejection of group selection appear so important in the first place. The central issue addressed by Williams in *Adaptation and Natural Selection* was whether adaptations can evolve at the level of social groups and other higher-level units. The problem, as recognized by Darwin and affirmed by Williams, was that traits that are "for the good of the group" are usually not favored by selection within groups—what we have called the fundamental problem of social life. When Williams and others rejected group selection, they were rejecting the possibility that adaptations evolve above the level of individual organisms. This is not a matter of perspective, but a fundamental biological claim. If true, it is every bit as momentous as it appeared to be in the 1960s. If false, then its retraction is equally momentous.

A sample of issues debated by contemporary theorists and philosophers of biology will show that, whatever the merits of pluralism, they do not deny the fundamental problem of social life or provide a solution other than between-group selection. Let us begin with inclusive fitness theory. Hamilton (1963, 1964) originally interpreted the coefficient of relatedness (r), as a measure of genealogical relatedness, based on genes that are identical by descent. When he reformulated his theory in terms of the Price equation, he realized not only that kin selection is a kind of group selection, but also that r can be interpreted more broadly as any positive correlation among altruistic genes—not just based on identity by descent (Hamilton 1975). Subsequent theorists have broadened the interpretation of r still further. For example, altruistic genes can evolve as long as they associate positively with altruistic *phenotypes,* coded by the same or different altruistic genes (Queller 1985; Fletcher and Doebeli 2006). When individuals benefit their entire group (including themselves) at their own expense, r can be positive even in randomly formed groups (Pepper 2000; Fletcher and Zwick 2004). Models that were originally conceptualized as examples of group selection, in contrast to kin selection, such as Maynard Smith's (1964) haystack model, can be reconceptualized as models of kin selection by noting that members of groups are more genetically similar to each other than to members of the total population. Generality is a virtue, so it is understandable that theorists might want to push the boundaries of inclusive fitness theory as far as possible. Nevertheless, when everything that was ever called group selection can now be described in terms of inclusive fitness theory, it is time to take stock of the original empirical issues at stake. Is the fundamental problem of social life present in the broadened form of inclusive fitness theory? Absolutely. Altruistic traits are locally disadvantageous, just as they always were. Are the ingredients of between-group selection required to solve the fundamental problem of social life? Absolutely. Altruistic traits still must be favored at a larger scale to counteract their local disadvantage. Does altruism evolve only among immediate genealogical relatives? Absolutely not. In the passage quoted at the beginning of this article, Williams (1966) rejected group-level adaptations for any groups "other than a family" or "composed of individuals that need not be closely related," by which he meant genealogical relatedness. Inclusive fitness theory refuted this claim as soon as r became generalized beyond immediate genealogical relatedness (e.g., Avilés 2002).

To pick a second example of pluralism, Kerr and Godfrey-Smith (2002a) outline two equivalent frameworks that they call *collective* and *contextual* (similar to Dugatkin and Reeve's 1994 distinction between multilevel selection and broad-sense individualism). In the collective framework, groups are assigned fitnesses and individuals are assigned different shares of their group's fitness. In the contextual framework, individuals are assigned fitnesses that are functions of the composition of their group. The distinction between the two frameworks is similar to thinking of genotypes as individuals, as in standard population genetics theory, as opposed to environments of genes, as in selfish gene theory. Kerr and Godfrey-Smith stress that the two frameworks are fully equivalent, which means that any statement in one can be translated into a statement in the other. Equivalence also means that neither is more "correct" in any causal sense, although one might provide more insight than the other in any particular case. Fair enough, but this kind of pluralism by itself does not address any particular empirical issue. When we begin to ask the empirical questions that endow the group selection controversy with such significance, we discover that the contextual approach does not avoid the fundamental problem of social life or provide a solution other than between-group selection. It merely describes these processes in different terms. In this sense "broad-based individualism" (= the contextual approach) is nothing like "the theory of individual selection" that claimed to be a genuine alternative to group selection, such that one could be right and the other wrong (for more detailed discussion of this issue, see Kerr and Godfrey Smith 2002b; Sober and Wilson 2002).

As a third example of pluralism, even though the Price equation elegantly partitions selection into within- and between-group components, it misclassifies certain cases. In particular, when individuals that differ in their individual fitness (without behaving socially at all) are separated into groups, the between-group component of the Price equation is positive, even though there is no group selection (Sober 1984). Another statistical method called contextual analysis avoids this problem, but it misclassifies other cases. Thus, there is no single statistical method that captures all aspects of multilevel selection theory (van Veelen 2005; Okasha 2006). This is interesting and important, but does not cast doubt on the basic empirical issues. In fact, the reason that we can spot classification errors in statistical methods such as the Price equation is because we have such a strong sense of what multilevel selection means in the absence of formal statistical methods.

In general, the issues discussed under the rubric of pluralism are important but also highly derived, to the point of becoming detached from the issues that endowed multilevel selection with such importance in the first place. There is a need for all perspectives to converge upon a core set of empirical claims, including the following:

1) There *is* a fundamental problem that requires a solution in order to explain the evolution of altruism and other group-level adaptations. Traits that are "for the good of the group" are seldom selectively advantageous within groups. At worst, they are highly self-sacrificial. At best, they provide public goods at little cost to the actor, making them close to selectively neutral, or they constitute a stable local equilibrium. Notice that the only way to evaluate this claim is by making a local relative fitness comparison. It is not enough to show that an individual increases its absolute fitness because it might increase the fitness of others in its own group even more (D S Wilson 2004).

2) If a trait is locally disadvantageous wherever it occurs, then the only way for it to evolve in the total population is for it to be advantageous at a larger scale. Groups whose members act "for the good of the group" must contribute more to the total gene pool than groups whose members act otherwise. This is the *only* solution to the problem from an accounting standpoint, which is why the basic logic of multilevel selection is present in all theoretical frameworks, as we showed in the previous section. In biological hierarchies that include more than two levels, the general rule is "adaptation at any level requires a process of natural selection at the same level and tends to be undermined by natural selection at lower levels." All students of evolution need to learn this rule to avoid the errors of naïve group selectionism. Notice that, so far, we are *affirming* key elements of the consensus that formed in the 1960s.

3) Higher-level selection cannot be categorically ignored as a significant evolutionary force. Instead, it must be evaluated separately and on a case-by-case basis. Furthermore, all of the generalities about the likelihood of group selection that became accepted in the 1960s need to be reexamined. Wynne-Edwards's hypothesis has merit for at least some species, group selection can be significant in groups that last for multiple generations, and so on. One of the biggest problems with the current literature is that the early generalities remain unquestioned, as if there is an "old" group selection that deserves to be rejected and a "new" form that bears little relationship with its own past (e.g., Keller 1999; West et al. 2006, 2007). This is a false portrayal and cannot be justified on the basis of pluralism. Going back to basics requires acknowledgment that Williams and others were right to criticize naïve group selection, but just plain wrong in their own assessment of the likelihood of group selection. New generalities need to be formed on the basis of ongoing research.

4) The fact that a given trait evolves in the total population is not an argument against group selection. Evaluating levels of selection requires a nested series of relative fitness comparisons; between genes *within* individuals, between individuals *within* groups, between groups *within* a population of groups, and so on, each presenting traits that are separate targets for selection. All theoretical frameworks include the information for making these comparisons, as we have seen. In this sense, they are not pluralistic. They merely differ in the degree to which they focus on the comparisons on their way toward calculating evolutionary change in the total population. If we are merely interested in whether a given trait evolves, then it is not necessary to examine levels of selection, and multiple perspectives can be useful. If we want to address the particular biological issues associated with multilevel selection, then we are required to examine the appropriate information and the perspectives converge with each other.

To summarize, it is possible to acknowledge the usefulness of multiple perspectives without obscuring the fundamental biological issues that seemed so clear in the 1960s. We think that items 1–4 above can become the basis for a new consensus about when adaptations evolve at any given level of the biological hierarchy, restoring clarity and unity to sociobiological theory. We will now examine three cases where higher-level selection has been exceptionally important: the evolution of individual organisms, the evolution of eusociality in insects and other taxa, and human evolution.

Individuals as Groups

An important advance in evolutionary biology began with Margulis's (1970) theory of the eukaryotic cell. She proposed that eukaryotic (nucleated) cells did not evolve by small mutational steps from prokaryotic (bacterial) cells, but by symbiotic associations of bacteria becoming so integrated that the associations qualified as single organisms in their own right. The concept of groups *of* organisms turning into groups *as* organisms was then extended to other major transitions during the history of life, including the origin of life itself as groups of cooperating molecular reactions, the first cells, and multicellular organisms (e.g., Maynard Smith and Szathmáry 1995, 1999; Michod 1999; Jablonka and Lamb 2006; Michod and Herron 2006).

Despite multilevel selection theory's turbulent history for the traditional study of social behavior, it is an accepted theoretical framework for the study of major transitions. There is widespread agreement that selection occurs within and among groups, that the balance between levels of selection can itself evolve, and that a major transition occurs when selection within groups is suppressed, enabling selection among groups to dominate the final vector of evolutionary change. Genetic and developmental phenomena such as chromosomes, the rules of meiosis, a single cell stage in the life cycle, the early sequestration of the germ line, and programmed death of cell lineages are interpreted as mechanisms for stabilizing the organism and preventing it from becoming a mere group of evolving elements. At the same time, within-group selection is never completely suppressed. There are many examples of intragenomic conflict that prevent the higher-level units from functioning as organisms in the full and truest sense of the word (Burt and Trivers 2006).

The concept of major transitions decisively refutes the notion that higher-level selection is invariably weaker than lower-level selection. The domain of multilevel selection theory has been expanded to include the internal organization of individuals in addition to the social organization of groups. Ironically, the rejection of group selection made it heresy to think about groups as like organisms, and now it has emerged that organisms are literally the groups of past ages. Okasha (2005:1008) eloquently summarizes the implications of these developments for sociobiological theory as a whole:

> Since cells and multi-celled creatures obviously have evolved, and function well as adaptive units, the efficacy of group selection cannot be denied. Just as the blanket assumption that the individual organism is the sole unit of selection is untenable from a diachronic perspective, so too is the assumption that group selection is a negligible force. For by 'frameshifting' our perspective downwards, it becomes apparent that individual organisms *are* co-operative groups, so are the *product* of group selection!

Eusociality as a Major Transition

Eusociality, found primarily in social insects but now known in other organisms such as mammals (Sherman et al. 1991) and crustacea (Macdonald et al. 2006), has always played a pivotal role in the history of sociobiology. The term "eusocial" is applied to colonies whose members are multigenerational, cooperate in brood care, and are separated into reproductive and nonreproductive castes. For the first half of the 20th century, following W M Wheeler's classic paper of 1911, eusocial colonies were treated as "superorganisms" that evolved by between-colony selection. Hamilton's (1964) inclusive fitness theory appeared to offer a very different explanation based on genetic relatedness, especially the extra-high relatedness among sisters in ants, bees, and wasps based on their haplodiploid genetic system. The focus on genetic relatedness thereafter made it appear as if social insect evolution could be explained without invoking group selection, along with other examples of apparent altruism. The following passage from West-Eberhard (1981:12; parenthetical comments are hers) illustrates the degree to which between-colony selection was rejected as an explanation of eusociality in insects: "Despite the logical force of arguments against group (or colony) selection (e.g., Williams 1966) and the invention of tidy explanations for collaboration in individual terms . . . the supraorganism (colony-level selection) still haunts evolutionary discussions of insect sociality."

Four decades later, there is an urgent need to establish some fundamental biological claims that have been obscured rather than clarified by multiple perspectives. Beginning with Wheeler's original claim that eusocial colonies are superorganisms, the evolution of eusociality falls squarely within the paradigm of major transitions. Most traits associated with eusociality do not evolve by increasing in frequency within colonies, but by increasing the colony's contribution to the larger gene pool. Inclusive fitness theory is not a denial of this fact, although that is how it was originally interpreted. Hamilton's rule calculates the conditions under which an altruistic act increases the proportion of altruistic genes in the total population, not a single colony. Showing that a trait evolves in the total population is not an argument against group selection, as we have already stressed. The Price equation demonstrated to Hamilton that altruism is selectively disadvantageous within kin groups, just as in any other kind of group. The importance of kinship is that it increases genetic variation among groups, therefore the importance of between-group selection compared to within-group selection. There are traits that evolve by within-colony selection, but they are forms of cheating that tend to impair the performance of the colony, similar to intragenomic conflict within individual organisms (Ratnieks et al. 2006). All social insect biologists should be able to agree upon these facts, regardless of the theoretical framework that they employ.

Another substantive biological question is the role of genealogical relatedness in the evolution of eusociality. Hamilton's original theory was that the extra-high sociality of insect colonies can be explained by the extra-high relatedness among workers, at least in haplodiploid species, when groups are founded by single queens who have mated with a single male. More generally, Hamilton's rule ($br > c$, where $b =$ benefit to the recipient, $r =$ coefficient of relatedness, and $c =$ cost to the altruist) easily gives the impression that the degree of altruism should be proportional to r. This perception was in fact a principal reason for the erroneous early acceptance of collateral (indirect) kin selection as a critical force in the origin of eusociality (E O Wilson 1971,1975).

Decades of research have led to a more complicated story in which genealogical relatedness plays at best a supporting rather than a pivotal role. The haplodiploidy hypothesis has failed on empirical grounds. In addition to termites, numerous other diploid eusocial clades in insects and other taxa have been discovered since the 1960s, enough to render the association of haplodiploidy and eusociality statistically insignificant (E O Wilson and Hölldobler 2005). Moreover, many haplodiploid colonies are founded by multiple females and/or females that mated with multiple males, lowering genetic relatedness to unexceptional levels. Further, following colony foundation in primitively eusocial wasp species, the degree of relatedness tends to fall, not rise or hold steady, at least in cases where it has been measured (e.g., Landi et al. 2003; Fanelli et al. 2004). These facts are widely acknowledged by social insect biologists, but it is still common to read in the wider literature that genetic relatedness is the primary explanation for insect eusociality. In fact, extra-high relatedness within colonies may be better explained as a consequence rather than a cause of eusociality (E O Wilson and Hölldobler 2005).

From a multilevel evolutionary perspective, traits that cause an insect colony to function as an adaptive unit seldom increase in frequency within the colony and evolve only by causing the colony to out-compete other colonies and conspecific solitaires, either directly or through the differential production of reproductives. If colonies are initiated by small numbers of individuals, minimally a single female mated with a single male, then there is ample genetic variation among groups and only modest genetic variation within groups. However, this is only one of many factors that can influence the balance between levels of selection. Consider genetic variation for traits such as nest construction, nest defense, provisioning the colony with food, or raiding other colonies. All of these activities provide public goods at private expense. All entail emergent properties based on cooperation among the colony members. Slackers are more fit than solid citizens within any single colony, but colonies with more solid citizens have the advantage at the group level. The balance between levels of selection will be influenced by the magnitude of the group-level benefits and individual-level costs, in addition to the partitioning of genetic variation within and among groups. For example, ecological constraints are more important than genetic relatedness in the evolution of eusociality in mole-rats (Burland et al. 2002). The same is true of the eusocial invertebrates (Choe and Crespi 1997; E O Wilson and Hölldobler 2005). The ancestors of most eusocial insects probably built nests and remained to feed and protect their brood throughout larval development. Such a “progressive provisioning” was evidently the key preadaptation for the origin of eusociality in the Hymenoptera. It is the multigroup population structure provided by this ecological niche and the magnitude of shared benefits that brought these species up to and over the threshold of eusociality, more than exceptional degrees of genetic relatedness.

It might seem that reproductive division of labor must be a form of high-cost altruism that requires a high degree of genetic variation among groups (represented by high r values) to evolve. This is only true, however, if heritable phenotypic variation exists for worker reproduction and if reproductive workers are not suppressed by the queen or other workers. Reproductive suppression is common in eusocial species, and to understand its evolution we need to study the policing and reproduction traits in conjunction with each other (Ratnieks et al. 2006). Suppressing the reproduction of others can be favored by within-group selection, but it can take many forms that vary in their consequences for the reproductive output of the colony, compared to other colonies. Between-group selection is required to evolve forms of reproductive suppression that function well at the colony level, but the amount of genetic variation among colonies need not be exceptional. That need is diminished further when the trait favored by group selection is a form of phenotypic plasticity that enables single genotypes to be reproductive or nonreproductive—which, in fact, is universal in the social insects (E O Wilson 1975; Hölldobler and E O Wilson 1990).

In eusocial insects, it appears that the evolution of anatomically distinct worker castes represents a “point of no return” beyond which species never revert to a more primitively eusocial, presocial, or solitary condition (E O Wilson 1971; Maynard Smith and Szathmáry 1995; E O Wilson and Hölldobler 2005). At this point, the colony has become a stable developmental unit and its persistence depends on its ability to survive and reproduce, relative to other colonies and solitary organisms. The hypothetical mutant reproductive worker that would be favored by within-colony selection simply does not occur at significant levels or at all, although, in some species, “cheating” by workers occurs and is suppressed through policing by fellow workers. This is similar to the evolution of sexual lineages that do not give rise to asexual mutants (Nunney 1999) and the evolution of mechanisms that prevent intragenomic conflict in individual organisms (Maynard Smith and Szathmáry 1995, 1999).

A common assumption of theoretical models is that genes have additive effects on phenotypes, so that phenotypic variation among groups corresponds directly to genetic variation among groups, as we have already stressed. More complex genotype-phenotype relationships enable small genetic differences to result in large phenotypic differences, at the level of groups no less than individual organisms (D S Wilson 2004). Even a single mutant gene in a colony founded by unrelated individuals can have powerful effects on phenotypic traits such as caste development or allocation of workers to various tasks, which might provide a strong advantage to the group, compared to other groups.

Single eusocial insect colonies often have a population structure of their own, which can be spatial or based on kin recognition. There is a multiple-tiered population structure in which selection can occur between individuals within immediate families (such as matrilines or patrilines), between immediate families within a single colony, and between colonies within the larger population. In keeping with the dictum "adaptation at any level requires a process of natural selection at the same level and tends to be undermined by natural selection at lower levels," kin selection becomes part of the problem as far as colony-level selection is concerned. Numerous examples of nepotism as a disruptive force have been documented, along with mechanisms that have evolved to suppress nepotism along with individual selfishness, enabling the multifamily colony to be the primary unit of selection (Ratnieks et al. 2006; Wenseleers and Ratnieks 2006).

Social insect biologists spend much of their time studying the mechanisms that enable a colony to function as an adaptive unit. The title of one book, *The Wisdom of the Hive* (Seeley 1995), alludes effectively to Walter Cannon's (1932) *The Wisdom of the Body,* which famously described the complex physiological mechanisms of single organisms. The social interactions that enable an insect colony to make complex decisions are even directly comparable to the neuronal interactions that enable individual organisms to make decisions (Seeley and Buhrman 1999). These interactions did not evolve by within-colony selection, but by colonies with the most functional interactions out-competing other colonies. A high degree of relatedness was not required and little insight is gained by noting that individuals benefit as members of successful groups. The challenge is to understand the complex mechanisms that enable a colony to function *as a single organism,* exactly as imagined by Wheeler so long ago.

Almost all of the spectacular evolutionary efflorescence of the more than 12,000 known ant species, hence almost all the progressive advance of their communication and caste systems, life cycles, algorithms of colonial self-organization and caste-specific adaptive demographies, are manifestly the product of group selection acting on the emergent, colony-level traits, which are produced in turn by the interaction of the colony members.

We will conclude this section by discussing the extent to which pluralism has facilitated or retarded the study of the landscape of eusociality during the last four decades. The question is not whether everything that we have recounted above can be stated within the rubric of inclusive fitness theory; it can. Moreover, we certainly do not deny the advances in knowledge about social insects in recent decades, some of which has been stimulated by inclusive fitness theory as the dominant paradigm. Nevertheless, we also think that inclusive fitness theory has retarded understanding in a number of other important respects. First, it initially gave the impression that eusociality can be explained as an individual-level adaptation, without distinguishing and invoking group (=between-colony) selection; this turned out to be a monumental mistake. Second, it misleadingly suggested that genetic relatedness is the primary factor that explains the evolution of eusociality, distracting attention from other factors of greater importance. Third, the coefficient of relatedness was originally interpreted in terms of genealogical relatedness, whereas today it is interpreted more broadly in terms of any genetic or even phenotypic correlation among group members (Fletcher et al. 2006; Fletcher and Zwick 2006; Foster et al. 2006a,b). Inclusive fitness theory now completely overlaps with multilevel selection theory, as we have already stressed. Multiple perspectives *are* useful, as long as they are properly related to each other, and we are sure that inclusive fitness theory will be used to study eusociality in the future. However, we also think that multilevel selection theory will prove to be both correct and more heuristic, because it more clearly identifies the colony as the unit of selection that has driven the evolution of social complexity.

Human Evolution as a Major Transition

Anyone who studies humans must acknowledge our groupish nature and the importance of between-group interactions throughout human history. Ever since the 1960s, sociobiologists and evolutionary psychologists have been burdened with the task of explaining these obvious facts without invoking group selection. In retrospect, these explanations appear needlessly contorted. Instead, human evolution falls squarely within the paradigm of major transitions (Lumsden and E O Wilson 1981; Boehm 1999; Richerson and Boyd 1999; D S Wilson 2002, 2006, 2007a,b; Hammerstein 2003; Foster and Ratnieks 2005; Bowles 2006).

A key event in early human evolution was a form of guarded egalitarianism that made it difficult for some individuals to dominate others within their own group (Bingham 1999; Boehm 1999). Suppressing fitness differences within groups made it possible for between-group selection to become a powerful evolutionary force. The psychological traits associated with human moral systems are comparable to the mechanisms that suppress selection within groups for other major transitions, such as chromosomes and the rules of meiosis within multicellular organisms and policing mechanisms within eusocial insect colonies (D S Wilson 2002; Avilés et al. 2004; Haidt 2007). The human major transition was a rare event, but once accomplished, our ability to function as team players in coordinated groups enabled our species to achieve worldwide dominance, replacing other hominids and many other species along the way. The parallels with the other major transitions are intriguing and highly instructive (E O Wilson and Hölldobler 2005).

A common scenario for human evolution begins with the evolution of sophisticated cognitive abilities, such as a "theory of mind," which in turn enabled widespread cooperation (Tomasello 1999). Now it appears more reasonable for the sequence to be reversed (Tomasello et al. 2005). Our capacities for symbolic thought and the social transmission of information are fundamentally communal activities that probably required a shift in the balance between levels of selection before they could evolve. Only when we could trust our social partners to work toward shared goals could we rely upon them to share meaningful information. The shift in the balance between levels of selection is reflected in anatomical features, such as the human eye as an organ of communication (Kobayashi and Kohshima 2001), and basic cognitive abilities, such as the ability to point things out to others (Tomasello et al. 2005) and to laugh in a group context (Gervais and D S Wilson 2005), in addition to more advanced cognitive and cultural activities associated with our species.

Group selection is an important force in human evolution in part because cultural processes have a way of creating phenotypic variation among groups, even when they are composed of large numbers of unrelated individuals. If a new behavior arises by a genetic mutation, it remains at a low frequency within its group in the absence of clustering mechanisms such as associations among kin. If a new behavior arises by a cultural mutation, it can quickly become the most common behavior within the group and provide the decisive edge in between-group competition (Richerson and Boyd 2005). The importance of genetic and cultural group selection in human evolution enables our groupish nature to be explained at face value. Of course, within-group selection has only been suppressed, not entirely eliminated. Thus *multi*-level selection, not group selection alone, provides a comprehensive framework for understanding human sociality.

These ideas can potentially explain the broad sweep of recorded history in addition to the remote past. According to Turchin (2003, 2005), virtually all empires arose in geographical areas where major ethnic groups came into contact with each other. Intense between-group conflict acted as a crucible for the cultural evolution of extremely cooperative societies, which then expanded at the expense of less cooperative societies to become major empires. Their very success was their undoing, however, as cultural evolution within the empire led to myriad forms of exploitation, free riding, and factionalism. That is why the center of the former Roman empire (for example) is today a cultural "black hole" as far as the capacity for cooperation is concerned. Turchin, a theoretical biologist who specializes in nonlinear population dynamics, has marshaled an impressive amount of empirical evidence to support his thesis about the rise and fall of empires as a process of multilevel cultural evolution, with profound implications for interactions among modern cultures and their consequences for human welfare in the future.

A New Consensus and Theoretical Foundation for Sociobiology

Making a decision typically involves encouraging diversity at the beginning to evaluate alternatives, but then discouraging diversity toward the end to achieve closure and to act upon the final decision. It can be very difficult to revisit an important decision that has been made and acted upon, but that is precisely what needs to be done in the case of the 1960s consensus about group selection. Historians of science have made a start, including a recent article appropriately titled "The Rise, Fall, and Resurrection of Group Selection" (Borrello 2005; see also Okasha 2006), but the real need is for practicing sociobiologists to arrive at a new consensus based on the many developments that have taken place during the last four decades.

In concluding this article, it is interesting to revisit the contradictory positions that exist in the current sociobiological literature:

- Nothing has changed since the 1960s. An example is Alcock's (2005) influential textbook *Animal Behavior: An Evolutionary Approach,* in which group selection is described as non-Darwinian and a near impossibility because of the insuperable problem of selection within groups. There is no excuse for this kind of treatment, given the developments over the last four decades that we have reviewed in this article.
- Multilevel selection theory (including group selection) has been fully revived. It is important to stress once again that this is *not* a return to naïve group selectionism. On the contrary, going "back to basics" means *affirming* key elements of the consensus that formed in the 1960s, which insisted that higher-level adaptations require a process of higher-level selection and cannot be expected to evolve otherwise. The revival of multilevel selection is based solely on rejecting the empirical claim that higher-level selection can be categorically ignored as an important evolutionary force. It is notable that key figures such as Williams (for sex ratio and disease virulence), Hamilton (in terms of the Price equation), and Maynard Smith (for major transitions of evolution) easily reverted back to multilevel selection when they became convinced that group selection might be a significant evolutionary force after all. It is time for everyone to follow suit, for sociobiology as a whole rather than specific subject areas.
- There is a "new" multilevel selection theory that bears little relationship to the "old" theory. According to Richard Dawkins (quoted in Dicks 2000:35) "[e]normous credit would accrue to anyone who could pull off the seemingly impossible and rehabilitate group selection . . . [b]ut actually, such rehabilitation can't be achieved, because the great heresy really is wrong." Yet, theoretical biologists widely agree that modern multilevel selection is a legitimate theory for accounting for evolutionary change. The only way to maintain these two positions is by claiming that modern multilevel selection theory bears no relationship to its own past (e.g., Keller 1999; West et al. 2006, 2007). We hope that our "back to basics" approach has established the continuity of ideas, from Darwin to the present. Moreover, other than avoiding naïve group selection, *all* of the major conclusions about group selection that seemed to emerge during the 1960s, such as the rejection of Wynne-Edwards's hypothesis, need to be reconsidered on the basis of ongoing research.
- Avoiding the topic of group selection, as if it never existed in the history of evolutionary thought. We could cite dozens of theoretical and empirical articles from the current literature that describe selection within and among groups without mentioning the term "group selection" or anything else about the group selection controversy. As one example, the microbial experiment by Kerr et al. (2006) elegantly establishes the plausibility of

Wynne-Edwards's hypothesis and describes the process matter-of-factly in terms of selection within and among groups, without citing Wynne-Edwards or the term group selection. This polite silence enables authors such as West et al. (2006) to publish tutorials on social evolution for microbiologists that portray Wynne-Edwards's hypothesis as a theoretical impossibility. This kind of pluralism is not helpful (D S Wilson 2007a). We hope that our article will help to refocus attention on the problem that has always been at the center of multilevel selection theory: the fact that group-level adaptations are seldom locally advantageous and, therefore, must be favored at a larger scale to evolve. The fact that all theoretical frameworks reflect this problem and its (partial) solution is a major simplification that should be welcomed rather than resisted.

When Rabbi Hillel was asked to explain the Torah in the time that he could stand on one foot, he famously replied: "Do not do unto others that which is repugnant to you. Everything else is commentary." Darwin's original insight and the developments reviewed in this article enable us to offer the following one-foot summary of sociobiology's new theoretical foundation: "Selfishness beats altruism within groups. Altruistic groups beat selfish groups. Everything else is commentary."

ACKNOWLEDGMENTS

We thank Mark Borrello, Samuel Bowles, Anne Clark, Michael Doebeli, Jeff Fletcher, Bert Hölldobler, Kevin Foster, Benjamin Kerr, Samir Okasha, David Queller, Francis Ratnieks, Elliott Sober, Tom Wenseleers, and Martin Zwick for helpful discussion.

REFERENCES

Aktipis C A. 2004. Know when to walk away: contingent movement and the evolution of cooperation. *Journal of Theoretical Biology* 231(2):249–260.

Alcock J. 2005. *Animal Behavior: An Evolutionary Approach.* Eighth Edition. Sunderland (MA): Sinauer.

Allee W C, Emerson A E, Park O, Park T, Schmidt K P. 1949. *Principles of Animal Ecology.* Philadelphia (PA): W. B. Saunders.

Avilés L. 1993. Interdemic selection and the sex ratio: a social spider perspective. *American Naturalist* 142(2):320–345.

Avilés L. 2002. Solving the freeloaders paradox: genetic associations and frequency-dependent selection in the evolution of cooperation among nonrelatives. *Proceedings of the National Academy of Sciences of the United States of America* 99(22):14268–14273.

Avilés L, Abbot P, Cutter A D. 2002. Population ecology, nonlinear dynamics, and social evolution. I. Associations among nonrelatives. *American Naturalist* 159(2):115–127.

Avilés L, Fletcher J A, Cutter A C. 2004. The kin composition of social groups: trading group size for degree of altruism. *American Naturalist* 164(2): 132–144.

Bingham P M. 1999. Human uniqueness: a general theory. *Quarterly Review of Biology* 74(2):133–169.

Boehm C. 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior.* Cambridge (MA): Harvard University Press.

Borrello M E. 2005. The rise, fall, and resurrection of group selection. *Endeavour* 29:43–47.

Bourke A F G, Franks N R. 1995. *Social Evolution in Ants.* Princeton (NJ): Princeton University Press.

Bowles S. 2006. Group competition, reproductive leveling, and the evolution of human altruism. *Science* 314:1569–1572.

Boyd R, Richerson P J. 1985. *Culture and the Evolutionary Process.* Chicago (IL): University of Chicago Press.

Boyd R, Richerson P J. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* 13(3):171–195.

Bull J J. 1994. Virulence. *Evolution* 48(5):1423–1437.

Burland T M, Bennett N C, Jarvis J U M, Faulkes C G. 2002. Eusociality in African mole-rats: new insights from patterns of genetic relatedness in the Damaraland mole-rat. *Proceedings of the Royal Society of London Series B* 269:1025–1030.

Burt A, Trivers R. 2006. *Genes in Conflict: The Biology of Selfish Genetic Elements.* Cambridge (MA): Belknap Press of Harvard University Press.

Cannon W B. 1932. *The Wisdom of the Body.* First Edition. New York: W. W. Norton.

Choe J C, Crespi B J, editors. 1997. *The Evolution of Social Behavior in Insects and Arachnids.* Cambridge (UK): Cambridge University Press.

Cronk L. 2005. The application of animal signaling theory to human phenomena: some thoughts and clarifications. *Social Science Information/Information sur les Sciences Sociales* 44:603–620.

Darwin C. 1871. *The Descent of Man, and Selection in Relation to Sex,* Volumes 1 and 2. New York: Appleton.

Dawkins R. 1976. *The Selfish Gene.* New York: Oxford University Press.

Dawkins R. 1982. *The Extended Phenotype: The Gene as the Unit of Selection.* San Francisco (CA): Freeman.

Dicks L. 2000. All for one! *New Scientist* 167(2246):30.

Dugatkin L A. 2002. Cooperation in animals: an evolutionary overview. *Biology and Philosophy* 17(4): 459–476.

Dugatkin L A, Reeve H K. 1994. Behavioral ecology and levels of selection: dissolving the group selection controversy. *Advances in the Study of Behavior* 23:101–133.

Fanelli D, Turillazzi S, Boomsma J J. 2004. Genetic relationships between females, males, and older brood in the tropical hover wasp *Parischnogaster mellyi* (Saussure). *Insect Social Life* 5:35–40.

Fletcher J A, Doebeli M. 2006. How altruism evolves: assortment and synergy. *Journal of Evolutionary Biology* 19(5):1389–1393.

Fletcher J A, Zwick M. 2004. Strong altruism can evolve in randomly formed groups. *Journal of Theoretical Biology* 228(3):303–313.

Fletcher J A, Zwick M. 2006. Unifying the theories of inclusive fitness and reciprocal altruism. *American Naturalist* 168(2):252–262.

Fletcher J A, Zwick M, Doebeli M, Wilson D S. 2006. What's wrong with inclusive fitness? *Trends in Ecology & Evolution* 21(11):597–598.

Foster K R. 2006. Balancing synthesis with pluralism in sociobiology. *Journal of Evolutionary Biology* 19(5): 1394–1396.

Foster K R, Parkinson K, Thompson C R L. 2007. What can microbial genetics teach sociobiology? *Trends in Genetics* 23(2):74–80.

Foster K R, Ratnieks F L W. 2005. A new eusocial vertebrate? *Trends in Ecology & Evolution* 20(7):363–364.

Foster K R, Wenseleers T, Ratnieks F L W. 2006a. Kin selection is the key to altruism. *Trends in Ecology & Evolution* 21(2):57–60.

Foster K R, Wenseleers T, Ratnieks F L W, Queller D C. 2006b. There is nothing wrong with inclusive fitness. *Trends in Ecology & Evolution* 21(11):599–600.

Frank S A. 1996. Models of parasite virulence. *Quarterly Review of Biology* 71(1):37–78.

Gervais M, Wilson D S. 2005. The evolution and functions of laughter and humor: a synthetic approach. *Quarterly Review of Biology* 80(4):395–430.

Gilpin M E. 1975. *Group Selection in Predator-Prey Communities.* Princeton (NJ): Princeton University Press.

Gintis H. 2000. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Behavior.* Princeton (NJ): Princeton University Press.

Goodnight C J. 2000. Heritability at the ecosystem level. *Proceedings of the National Academy of Sciences of the United States of America* 97(17):9365–9366.

Goodnight C J. 2005. Multilevel selection: the evolution of cooperation in non-kin groups. *Population Ecology* 47(1):3–12.

Goodnight C J, Schwartz J M, and Stevens L. 1992. Contextual analysis of models of group selection, soft selection, hard selection, and the evolution of altruism. *American Naturalist* 140 (5):743–761.

Goodnight C J, Stevens L. 1997. Experimental studies of group selection: what do they tell us about group selection in nature? *American Naturalist* 150(Supplement):S59–S79.

Haidt J. 2007. The new synthesis in moral psychology. *Science* 316:998–1002.

Hamilton W D. 1963. The evolution of altruistic behavior. *American Naturalist* 97(896):354–356.

Hamilton W D. 1964. The genetical evolution of social behavior, I and II. *Journal of Theoretical Biology* 7(1): 1–52.

Hamilton W D. 1975. Innate social aptitudes in man, an approach from evolutionary genetics. Pages 133–155 in *Biosocial Anthropology,* edited by R Fox. London (UK): Malaby Press.

Hamilton W D. 1996. *Narrow Roads of Gene Land: The Collected Papers of W. D. Hamilton.* Oxford (UK): W. H. Freeman/Spektrum.

Hammerstein P, editor. 2003. *Genetic and Cultural Evolution of Cooperation.* Cambridge (MA): MIT Press.

Heinsohn R, Packer C. 1995. Complex cooperative strategies in group-territorial African lions. *Science* 269:1260–1262.

Hölldobler B, Wilson E O. 1990. *The Ants.* Cambridge (MA): Belknap Press of Harvard University Press.

Jablonka E, Lamb M J. 2006. The evolution of information in the major transitions. *Journal of Theoretical Biology* 239(2):236–246.

Johnson C R, Boerlijst M C. 2002. Selection at the level of the community: the importance of spatial structure. *Trends in Ecology & Evolution* 17(2):83–90.

Keller L, editor. 1999. *Levels of Selection in Evolution.* Princeton (NJ): Princeton University Press.

Kerr B, Godfrey-Smith P. 2002a. Individualist and multi-level perspectives on selection in structured populations. *Biology and Philosophy* 17(4):477–517.

Kerr B, Godfrey-Smith P. 2002b. Group fitness and multi-level selection: replies to commentaries. *Biology and Philosophy* 17(4):539–549.

Kerr B, Neuhauser C, Bohannan B J M, Dean A M. 2006. Local migration promotes competitive restraint in a host-pathogen "tragedy of the commons." *Nature* 442:75–78.

Kobayashi H, Kohshima S. 2001. Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. *Journal of Human Evolution* 40(5):419–435.

Lachmann M, Számadó S, Bergstrom C T. 2001. Cost and conflict in animal signals and human language. *Proceedings of the National Academy of Sciences of the United States of America* 98(23):13189–13194.

Landi M, Queller D C, Turillazzi S, Strassmann J E. 2003. Low relatedness and frequent queen turnover in the stenogastrine wasp *Eustenogaster fraterna* favor the life insurance over the haplodiploid hypothesis for the origin of eusociality. *Insectes Sociaux* 50(3):262–267.

Lumsden C J, Wilson E O. 1981. *Genes, Mind, and Culture: The Coevolutionary Process*. Cambridge (MA): Harvard University Press.

Macdonald K S, Ríos R, Duffy J E. 2006. Biodiversity, host specificity, and dominance by eusocial species among sponge-dwelling alpheid shrimp on the Belize Barrier Reef. *Diversity and Distributions* 12(2): 165–178.

Margulis L. 1970. *Origin of Eukaryotic Cells: Evidence and Research Implications for a Theory of the Origin and Evolution of Microbial, Plant, and Animal Cells on the Precambrian Earth*. New Haven (CT): Yale University Press.

Maynard Smith J. 1964. Group selection and kin selection. *Nature* 201:1145–1146.

Maynard Smith J. 1982. *Evolution and the Theory of Games*. Cambridge (UK): Cambridge University Press.

Maynard Smith J. 2002. Commentary on Kerr and Godfrey-Smith. *Biology and Philosophy* 17(4):523–527.

Maynard Smith J, Szathmáry E. 1995. *The Major Transitions in Evolution*. New York: W. H. Freeman/Spektrum.

Maynard Smith J, Szathmáry E. 1999. *The Origins of Life: From the Birth of Life to the Origin of Language*. New York: Oxford University Press.

Michod R E. 1982. The theory of kin selection. *Annual Review of Ecology and Systematics* 13:23–55.

Michod R E. 1999. *Darwinian Dynamics: Evolutionary Transitions in Fitness and Individuality*. Princeton (NJ): Princeton University Press.

Michod R E, Herron M D. 2006. Cooperation and conflict during evolutionary transitions in individuality. *Journal of Evolutionary Biology* 19(5):1406–1409.

Nowak M A. 2006. Five rules for the evolution of cooperation. *Science* 314:1560–1563.

Nowak M A, Sigmund K. 2005. Evolution of indirect reciprocity. *Nature* 437:1291–1298.

Nunney L. 1999. Lineage selection: natural selection for long-term benefit. Pages 238–252 in *Levels of Selection in Evolution*, edited by L Keller. Princeton (NJ): Princeton University Press.

Okasha S. 2005. Maynard Smith on the levels of selection question. *Biology and Philosophy* 20(5):989–1010.

Okasha S. 2006. *Evolution and the Levels of Selection*. New York: Oxford University Press.

Packer C, Heinsohn R. 1996. Lioness leadership. *Science* 271:1215–1216.

Peck J. 2004. Sex causes altruism. Altruism causes sex. Maybe. *Proceedings of the Royal Society Series B* 271: 993–1000.

Pepper J W. 2000. Relatedness in trait group models of social evolution. *Journal of Theoretical Biology* 206(3):355–368.

Pepper J W. 2007. Simple models of assortment through environmental feedback. *Artificial Life* 13: 1–9.

Pepper J W, Smuts B B. 2002. A mechanism for the evolution of altruism among nonkin: positive assortment through environmental feedback. *American Naturalist* 160(2):205–213.

Price G R. 1970. Selection and covariance. *Nature* 227:520–521.

Price G R. 1972. Extension of covariance selection mathematics. *Annals of Human Genetics* 35(4):485–490.

Queller D C. 1985. Kinship, reciprocity, and synergism in the evolution of social behavior. *Nature* 318: 366–367.

Queller D C. 1992. Does population viscosity promote kin selection? *Trends in Ecology & Evolution* 7(10): 322–324.

Rainey P B, Rainey K. 2003. Evolution of cooperation and conflict in experimental bacterial populations. *Nature* 425:72–74.

Rapoport A. 1991. Ideological commitments in evolutionary theories. *Journal of Social Issues* 47(3):83–99.

Ratnieks F L W, Foster K R, Wenseleers T. 2006. Conflict resolution in insect societies. *Annual Review of Entomology* 51:581–608.

Ratnieks F L W, Visscher P K. 1989. Worker policing in the honeybee. *Nature* 342:796–797.

Richerson P J, Boyd R. 1999. Complex societies: the evolutionary origins of a crude superorganism. *Human Nature* 10(3):253–289.

Richerson P J, Boyd R. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago (IL): University of Chicago Press.

Sachs J L, Mueller U G, Wilcox T P, Bull J J. 2004. The evolution of cooperation. *Quarterly Review of Biology* 79(2):135–160.

Samuelson L. 1997. *Evolutionary Games and Equilibrium Selection*. Cambridge (MA): MIT Press.

Schwartz J. 2000. Death of an altruist. *Lingua Franca: The Review of Academic Life* 10(5):51–61.

Seeley T D. 1995. *The Wisdom of the Hive: The Social Physiology of Honey Bee Colonies*. Cambridge (MA): Harvard University Press.

Seeley T D, Buhrman S C. 1999. Group decision making in swarms of honey bees. *Behavioral Ecology and Sociobiology* 45(1):19–31.

Sherman P W, Jarvis J U M, Alexander R D, editors. 1991. *The Biology of the Naked Mole-Rat.* Princeton (NJ): Princeton University Press.

Sober E. 1984. *The Nature of Selection: Evolutionary Theory in Philosophical Focus.* Cambridge (MA): MIT Press.

Sober E. 2008. *Evidence and Evolution: The Logic Behind the Science.* Cambridge (UK): Cambridge University Press.

Sober E, Wilson D S. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior.* Cambridge (MA): Harvard University Press.

Sober E, Wilson D S. 2002. Perspectives and parameterizations: commentary on Benjamin Kerr and Peter Godfrey-Smith's "Individualist and multilevel perspectives on selection in structured populations." *Biology and Philosophy* 17(4):529–537.

Swenson W, Arendt J, Wilson D S. 2000a. Artificial selection of microbial ecosystems for 3-chloroaniline biodegradation. *Environmental Microbiology* 2(5): 564–571.

Swenson W, Wilson D S, Elias R. 2000b. Artificial ecosystem selection. *Proceedings of the National Academy of Sciences of the United States of America* 97(16): 9110–9114.

Tomasello M. 1999. *The Cultural Origins of Human Cognition.* Cambridge (MA): Harvard University Press.

Tomasello M, Carpenter M, Call J, Behne T, Moll H. 2005. Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences* 28(5):675–735.

Turchin P. 2003. *Historical Dynamics: Why States Rise and Fall.* Princeton (NJ): Princeton University Press.

Turchin P. 2005. *War and Peace and War: The Rise and Fall of Empires.* Upper Saddle River (NJ): Pi Press.

van Veelen M. 2005. On the use of the Price equation. *Journal of Theoretical Biology* 237(4):412–426.

Velicer G J. 2003. Social strife in the microbial world. *Trends in Microbiology* 11(7):330–337.

Wenseleers T, Ratnieks F L W. 2006. Enforced altruism in insect societies. *Nature* 444:50.

Wenseleers T, Ratnieks F L W, Billen J. 2003. Caste fate conflict in swarm-founding social Hymenoptera: an inclusive fitness analysis. *Journal of Evolutionary Biology* 16(4):647–658.

Werfel J, Bar-Yam Y. 2004. The evolution of reproductive restraint through social communication. *Proceedings of the National Academy of Sciences of the United States of America* 101(30):11019–11024.

West S A, Griffin A S, Gardner A, Diggle S P. 2006. Social evolution theory for microorganisms. *Nature Reviews Microbiology* 4(8):597–607.

West S A, Griffin A S, Gardner A. 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity, and group selection. *Journal of Evolutionary Biology* 20(2):415–432

West-Eberhard M J. 1981. Intragroup selection and the evolution of insect societies. Pages 3–17 in *Natural Selection and Social Behavior: Recent Research and New Theory,* edited by R D Alexander and D W Tinkle. New York: Chiron Press.

Wheeler W M. 1911. The ant colony as an organism. *Journal of Morphology* 22:307–325.

Williams G C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought.* Princeton (NJ): Princeton University Press.

Williams G C. 1985. A defense of reductionism in evolutionary biology. *Oxford Surveys in Evolutionary Biology* 2:1–27.

Williams G C. 1992. *Natural Selection: Domains, Levels, and Challenges.* New York: Oxford University Press.

Williams G C, Nesse R M. 1991. The dawn of Darwinian medicine. *Quarterly Review of Biology* 66(1):1–22

Wilson D S. 1975. A theory of group selection. *Proceedings of the National Academy of Sciences of the United States of America* 72(1):143–146.

Wilson D S. 1987. Altruism in Mendellian populations derived from sibling groups: the haystack model revisited. *Evolution* 41(5):1059–1070.

Wilson D S. 1997. Altruism and organism: disentangling the themes of multilevel selection theory. *American Naturalist* 150(supplement):S122–S134.

Wilson D S. 2002. *Darwin's Cathedral: Evolution, Religion, and the Nature of Society.* Chicago (IL): University of Chicago Press.

Wilson D S. 2004. What is wrong with absolute individual fitness? *Trends in Ecology & Evolution* 19(5): 245–248.

Wilson D S. 2006. Human groups as adaptive units: toward a permanent consensus. Pages 78–90 in *The Innate Mind,* Volume 2: Culture and Cognition, edited by P Carruthers, S Laurence, and S Stich. Oxford (UK): Oxford University Press.

Wilson D S. 2007a. Social semantics: Toward a genuine pluralism in the study of social behaviour. *Journal of Evolutionary Biology.* Available online at http://www.blackwell-synergy.com/doi/pdf/10.1111/j.1420-9101.2007.01396.x.

Wilson D S. 2007b. *Evolution for Everyone: How Darwin's Theory Can Change the Way We Think About Our Lives.* New York: Delacorte Press.

Wilson E O. 1971. *The Insect Societies.* Cambridge (MA): Belknap Press of Harvard University Press.

Wilson E O. 1975. *Sociobiology: The New Synthesis.* Cambridge (MA): Belknap Press of Harvard University Press.

Wilson E O, Hölldobler B. 2005. Eusociality: origin and consequences. *Proceedings of the National Academy of Sciences of the United States of America* 102(38): 13367–13371.

Wynne-Edwards V C. 1962. *Animal Dispersion in Relation to Social Behaviour.* Edinburgh (UK): Oliver and Boyd.

Wynne-Edwards V C. 1986. *Evolution Through Group Selection.* Boston (MA): Blackwell Scientific.

ent, fun, readable and slimmer than *War and Peace.'*
Simon Singh

s much, much stranger than science fiction . . .

breath you take contains an atom breathed out by Marilyn Monroe

entire human race would fit in the volume of a sugar cube

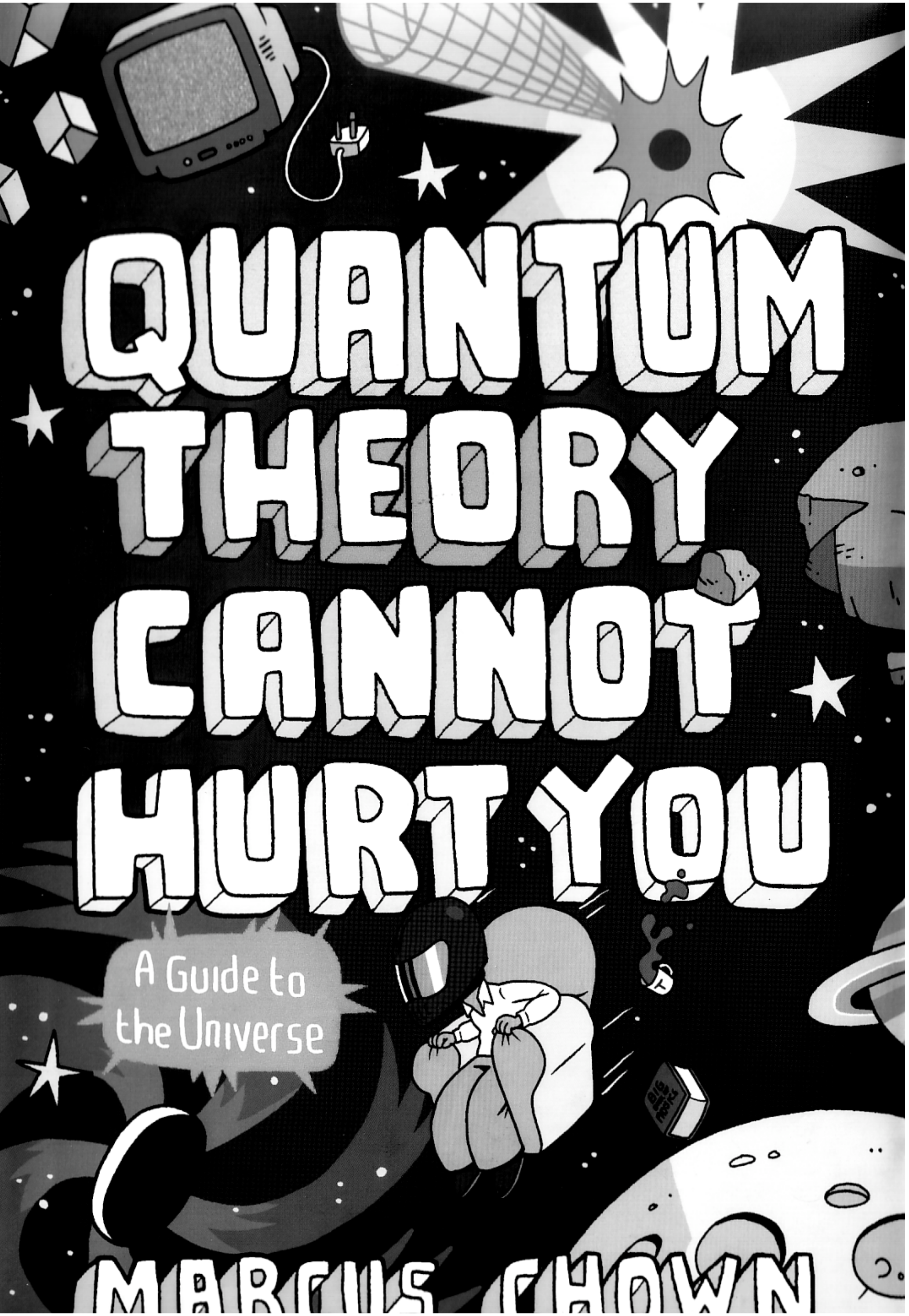
You age faster at the top of a building than at the bottom

are true – but why? Two brilliant ideas – quantum theory
Einstein's general theory of relativity – hold the key.

illuminating and seemingly impossible, *Quantum Theory You* reveals the wonders of modern physics – and explains why the faster you travel, the slimmer you get.

9 780571 235452

QUANTUM THEORY CANNOT HURT YOU
marcus chown



QUANTUM THEORY CANNOT HURT YOU

A Guide to the Universe

MARCUS CHOWN

1

Breathing in Einstein

How we discovered that everything is made of atoms and that atoms are mostly empty space

A hydrogen atom in a cell at the end of my nose was once part of an elephant's trunk.

Jostein Gaarder

We never had any intention of using the weapon. But they were such a terribly troublesome race. They insisted on seeing us as the "enemy" despite all our efforts at reassurance. When they fired their entire nuclear stockpile at our ship, orbiting high above their blue planet, our patience simply ran out.

The weapon was simple but effective. It squeezed out all the empty space from matter.

As the commander of our Sirian expedition examined the shimmering metallic cube, barely 1 centimetre across, he shook his primary head despairingly. Hard to believe that this was all that was left of the "human race"!

If the idea of the entire human race fitting into the volume of a sugar cube sounds like science fiction, think again. It is a remarkable fact that 99.9999999999999 per cent of the volume of ordinary matter is empty space. If there were some way to squeeze all the empty space out of the atoms in our bodies, humanity would indeed fit into the space occupied by a sugar cube.

The appalling emptiness of atoms is only one of the extraordinary characteristics of the building blocks of matter. Another, of course, is their size. It would take 10 million atoms laid end to end to span the width of a single full stop on this page, which raises the question, how did we ever discover that everything is made of atoms in the first place?

The idea that everything is made of atoms was actually first suggested by the Greek philosopher Democritus in about 440 BC.[1] Picking up a rock—or it may have been a branch or a clay pot—he asked himself the question: "If I cut this in half, then in half again, can I go on cutting it in half forever?" His answer was an emphatic *no.* It was inconceivable to him that matter could be subdivided forever. Sooner or later, he reasoned, a tiny grain of matter would be reached that could be cut no smaller. Since the Greek for "uncuttable" was "*a-tomos*," Democritus called the hypothetical building blocks of all matter "atoms."

Since atoms were too small to be seen with the senses, finding evidence for them was always going to be difficult. Nevertheless, a way was found by the 18th-century Swiss mathematician Daniel Bernoulli. Bernoulli realised that, although atoms were impossible to observe directly, it might still be possible to observe them indirectly. In particular, he reasoned that if a large enough number of atoms acted together, they might have a big enough effect to be obvious in the everyday world. All he needed was to find a place in nature where this happened. He found one—in a "gas."

Bernoulli imagined a gas like air or steam as a collection of billions upon billions of atoms in perpetual frenzied motion like a swarm of angry bees. This vivid picture immediately suggested an explanation for the "pressure" of a gas, which kept a balloon inflated

or pushed against the piston of a steam engine. When confined in any container, the atoms of a gas would drum relentlessly on the walls like hailstones on a tin roof. Their combined effect would be to create a jittery force that, to our coarse senses, would seem like a constant force pushing back the walls.

But Bernoulli's microscopic explanation of pressure provided more than a convenient mental picture of what was going on in a gas. Crucially, it led to a specific prediction. If a gas were squeezed into half its original volume, the gas atoms would need to fly only half as far between collisions with the container walls. They would therefore collide twice as frequently with those walls, doubling the pressure. And if the gas were squeezed into a third of its volume, the atoms would collide three times as frequently, trebling the pressure. And so on.

Exactly this behaviour was observed by the English scientist Robert Boyle in 1660. It confirmed Bernoulli's picture of a gas. And since Bernoulli's picture was of tiny grainlike atoms flying hither and thither through empty space, it bolstered the case for the existence of atoms. Despite this success, however, definitive evidence for the existence of atoms did not come until the beginning of the 20th century. It was buried in an obscure phenomenon called Brownian motion.

Brownian motion is named after Robert Brown, a botanist who sailed to Australia on the Flinders expedition of 1801. During his time down under, he classified 4,000 species of antipodean plants; in the process, he discovered the nucleus of living cells. But he is best remembered for his observation in 1827 of pollen grains suspended in water. To Brown, squinting through a magnifying lens, it seemed as if the grains were undergoing a curious jittery motion, zigzagging their way through the liquid like drunkards lurching home from a pub.

Brown never solved the mystery of the wayward pollen grains. That breakthrough had to wait for Albert Einstein, aged 26 and in the midst of the greatest explosion of creativity in the history of science. In his "miraculous year" of 1905, not only did Einstein overthrow

[1] Some of these ideas were covered in my earlier book, *The Magic Furnace* (Vintage, London, 2000). Apologies to those who have read it. In my defense, it is necessary to know some basic things about the atom in order to appreciate the chapters that follow on quantum theory, which is essentially a theory of the atomic world.

Newton, supplanting Newtonian ideas about motion with his special theory of relativity, but he finally penetrated the 80-year-old mystery of Brownian motion.

The reason for the crazy dance of pollen grains, according to Einstein, was that they were under continual machine-gun bombardment by tiny water molecules. Imagine a giant inflatable rubber ball, taller than a person, being pushed about a field by a large number of people. If each person pushes in their own particular direction, without any regard for the others, at any instant there will be slightly more people on one side than another. This imbalance is enough to cause the ball to move erratically about the field. Similarly, the erratic motion of a pollen grain can be caused by slightly more water molecules bombarding it from one side than from another.

Einstein devised a mathematical theory to describe Brownian motion. It predicted how far and how fast the average pollen grain should travel in response to the relentless battering it was receiving from the water molecules all around. Everything hinged on the size of the water molecules, since the bigger they were the bigger would be the imbalance of forces on the pollen grain and the more exaggerated its consequent Brownian motion.

The French physicist Jean Baptiste Perrin compared his observations of water-suspended "gamboge" particles, a yellow gum resin from a Cambodian tree, with the predictions of Einstein's theory. He was able to deduce the size of water molecules and hence the atoms out of which they were built. He concluded that atoms were only about one 10-billionth of a metre across—so small that it would take 10 million, laid end to end, to span the width of a full stop.

Atoms were so small, in fact, that if the billions upon billions of them in a single breath were spread evenly throughout Earth's atmosphere, every breath-sized volume of the atmosphere would end up containing several of those atoms. Put another way, every breath you take contains at least one atom breathed out by Albert Einstein—or Julius Caesar or Marilyn Monroe or even the last Tyrannosaurus Rex to walk on Earth!

What is more, the atoms of Earth's "biosphere" are constantly recycled. When an organism dies, it decays and its constituent atoms are returned to the soil and the atmosphere to be incorporated into plants that are later eaten by animals and humans. "A carbon atom in my cardiac muscle was once in the tail of a dinosaur," writes Norwegian novelist Jostein Gaarder in *Sophie's World.*

Brownian motion was the most powerful evidence for the existence of atoms. Nobody who peered down a microscope and saw the crazy dance of pollen grains under relentless bombardment could doubt that the world was ultimately made from tiny, bulletlike particles. But watching jittery pollen grains—the effect of atoms—was not the same as actually *seeing* atoms. This had to wait until 1980 and the invention of a remarkable device called the scanning tunnelling microscope (STM).

The idea behind the STM, as it became known, was very simple. A blind person can "see" someone's face simply by running a finger over it and building up a picture in their mind. The STM works in a similar way. The difference is that the "finger" is a finger of metal, a tiny stylus reminiscent of an old-fashioned gramophone needle. By dragging the needle across the surface of a material and feeding its up-and-down motion into a computer, it is possible to build up a detailed picture of the undulations of the atomic terrain.[2]

[2] Of course, there is no way a needle can actually feel a surface like a human finger can. However, if the needle is charged with electricity and placed extremely close to a conducting surface, a minuscule but measurable electric current leaps the gap between the tip of the needle and the surface. It is known as a "tunnelling current", and it has a crucial property that can be exploited: the size of the current is extraordinarily sensitive to the width of the gap. If the needle is moved even a shade closer to the surface, the current grows very rapidly; if it is pulled away a fraction, the current plummets. The size of the tunnelling current therefore reveals the distance between the needle tip and the surface. It gives the needle an artificial sense of touch.

Of course, there is a bit more to it than that. Although the principle of the invention was simple, there were formidable practical difficulties in its realisation. For instance, a needle had to be found that was fine enough to "feel" atoms. The Nobel Prize committee certainly recognised the difficulties. It awarded Gerd Binnig and Heinrich Rohrer, the IBM researchers behind the STM, the 1986 Nobel Prize for Physics.

Binnig and Rohrer were the first people in history to actually "see" an atom. Their STM images were some of the most remarkable in the history of science, ranking alongside that of Earth rising above the gray desolation of the Moon or the sweeping spiral staircase of DNA. Atoms looked like tiny footballs. They looked like oranges, stacked in boxes, row on row. But most of all they looked like the tiny hard grains of matter that Democritus had seen so clearly in his mind's eye, 2,400 years before. No one else has ever made a prediction that far in advance of experimental confirmation.

But only one side of the atom was revealed by the STM. As Democritus himself had realised, atoms were a lot more than simply tiny grains in ceaseless motion.

NATURE'S LEGO BRICKS

Atoms are nature's Lego bricks. They come in a variety of different shapes and sizes, and by joining them together in any number of different ways, it is possible to make a rose, a bar of gold, or a human being. Everything is in the combinations.

The American Nobel Prize winner Richard Feynman said: "If in some cataclysm all of scientific knowledge were destroyed and only one sentence passed on to succeeding generations, what statement would convey the most information in the fewest words?" He was in no doubt: "Everything is made of atoms."

The key step in proving that atoms are nature's Lego bricks was identifying the different kinds of atoms. However, the fact that atoms were far too small to be perceived directly by the senses made the task every bit as formidable as proving that atoms were tiny grains of matter in ceaseless motion. The only way to identify different types of atoms was to find substances that were made exclusively out of atoms of a single kind.

In 1789 the French aristocrat Antoine Lavoisier compiled a list of substances that he believed could not, by any means, be broken down into simpler substances. There were 23 "elements" in Lavoisier's list. Though some later turned out not to be elements, many—including gold, silver, iron, and mercury—were indeed elemental. Within 40 years of Lavoisier's death at the guillotine in 1794, the list of elements had grown to include close to 50. Nowadays, we know of 92 naturally occurring elements, from hydrogen, the lightest, to uranium, the heaviest.

But what makes one atom different from another? For instance, how does a hydrogen atom differ from a uranium atom? The answer would come only by probing their internal structures. But atoms are so fantastically small. It seemed impossible that anyone would ever find a way to look inside one. But one man did—a New Zealander named Ernest Rutherford. His ingenious idea was to use atoms to look inside other atoms.

THE MOTH IN THE CATHEDRAL

The phenomenon that laid bare the structure of atoms was radioactivity, discovered by the French chemist Henri Becquerel in 1896. Between 1901 and 1903, Rutherford and the English chemist Frederick Soddy found strong evidence that a radioactive atom is simply a heavy atom that is seething with excess energy. Inevitably, after a second or a year or a million years, it sheds this surplus energy by expelling some kind of particle at high speed. Physicists say it disintegrates, or "decays," into an atom of a slightly lighter element.

One such decay particle was the alpha particle, which Rutherford and the young German physicist Hans Geiger demonstrated was simply an atom of helium, the second lightest element after hydrogen.

In 1903, Rutherford had measured the speed of alpha particles expelled from atoms of radioactive radium. It was an astonishing 25,000 kilometres per second—100,000 times faster than a present-day passenger jet. Here, Rutherford realised, was a perfect bullet to smash into atoms and reveal what was deep inside.

The idea was simple. Fire alpha particles into an atom. If they encountered anything hard blocking their way, they would be deflected from their path. By firing thousands upon thousands of alpha particles and observing how they were deflected, it would be possible to build up a detailed picture of the interior of an atom.

Rutherford's experiment was carried out in 1909 by Geiger and a young New Zealand physicist called Ernest Marsden. Their "alpha-scattering" experiment used a small sample of radium, which fired off alpha particles like microscopic fireworks. The sample was placed behind a lead screen containing a narrow slit, so a thread-thin stream of alpha particles emerged on the far side. It was the world's smallest machine gun, rattling out microscopic bullets.

In the firing line Geiger and Marsden placed a sheet of gold foil only a few thousand atoms thick. It was insubstantial enough that all the alpha particles from the miniature machine gun would pass through. But it was substantial enough that, during their transit, some would pass close enough to gold atoms to be deflected slightly from their path.

At the time of Geiger and Marsden's experiment, one particle from inside the atom had already been identified. The electron had been discovered by the British physicist "J. J." Thomson in 1895. Ridiculously tiny particles—each about 2,000 times smaller than even a hydrogen atom—had turned out to be the elusive particles of electricity. Ripped free from atoms, they surged along a copper wire amid billions of others, creating an electric current.

The electron was the first subatomic particle. It carried a negative electric charge. Nobody knows exactly what electric charge is, only that it comes in two forms: negative and positive. Ordinary matter, which consists of atoms, has no net electrical charge. In ordinary atoms, then, the negative charge of the electrons is always perfectly balanced by the positive charge of something else. It is a characteristic of electrical charge that unlike charges attract each other whereas like charges repel each other. Consequently, there is a force of attraction between an atom's negatively charged electrons and its positively charged something else. It is this attraction that glues the whole thing together.

Not long after the discovery of the electron, Thomson used these insights to concoct the first-ever scientific picture of the atom. He visualised it as a multitude of tiny electrons embedded "like raisins in a plum pudding" in a diffuse ball of positive charge. It was Thomson's plum pudding model of the atom that Geiger and Marsden expected to confirm with their alpha-scattering experiment.

They were to be disappointed.

The thing that blew the plum pudding model out of the water was a rare but remarkable event. One out of every 8,000 alpha particles fired by the miniature machine gun actually bounced back from the gold foil!

According to Thomson's plum pudding model, an atom consisted of a multitude of pin-prick electrons embedded in a diffuse globe of positive charge. The alpha particles that Geiger and Marsden were firing into this flimsy arrangement, on the other hand, were unstoppable subatomic express trains, each as heavy as around 8,000 electrons. The chance of such a massive particle being wildly deflected from its path was about as great as that of a real express train being derailed by a runaway dolls pram. As Rutherford put it: "It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you!"

Geiger and Marsden's extraordinary result could only mean that an atom was not a flimsy thing at all. Something buried deep inside it could stop a subatomic express train dead in its tracks and turn it around. That something could only be a tiny nugget of positive charge sitting at the dead centre of an atom and repelling the positive charge of an incoming alpha particle. Since the nugget was capable of standing up to a massive alpha particle without being knocked to kingdom come, it too must be massive. In fact, it must contain almost all of the mass of an atom.

Rutherford had discovered the atomic nucleus.

The picture of the interior of the atom that emerged was as unlike Thomson's plum pudding picture as was possible to imagine. It was a miniature solar system in which negatively charged electrons were attracted to the positive charge of the nucleus and orbited it like planets around the Sun. The nucleus had to be at least as massive as an alpha particle—and probably a lot more so—for the nucleus with which it collided not to be kicked out of the atom. It therefore had to contain more than 99.9 per cent of the atom's mass.[3]

The nucleus was very, very tiny. Only if nature crammed a large positive charge into a very small volume could a nucleus exert a repulsive force so overwhelming that it could make an alpha particle execute a U-turn. What was most striking about Rutherford's vision of an atom was, therefore, its appalling emptiness. The playwright Tom Stoppard put it beautifully in his play *Hapgood:* "Now make a fist, and if your fist is as big as the nucleus of an atom then the atom is as big as St Paul's, and if it happens to be a hydrogen atom then it has a single electron flitting about like a moth in an empty cathedral, now by the dome, now by the altar."

Despite its appearance of solidity, the familiar world was actually no more substantial than a ghost. Matter, whether in the form of a chair, a human being, or a star, was almost exclusively empty space.

[3] Eventually, physicists would discover that the nucleus contains two particles: positively charged protons and uncharged, or neutral, neutrons. The number of protons in a nucleus is always exactly balanced by an equal number of electrons in orbit about it. The difference between atoms is in the number of protons in their nuclei (and consequently the number of electrons in orbit). For instance, hydrogen has one proton in its nucleus and uranium a whopping 92.

What substance an atom possessed resided in its impossibly small nucleus—100,000 times smaller than a complete atom.

Put another way, matter is spread extremely thinly. If it were possible to squeeze out all the surplus empty space, matter would take up hardly any room at all. In fact, this is perfectly possible. Although an easy way to squeeze the human race down to the size of a sugar cube probably does not exist, a way does exist to squeeze the matter of a massive star into the smallest volume possible. The squeezing is done by tremendously strong gravity, and the result is a neutron star. Such an object packs the enormous mass of a body the size of the Sun into a volume no bigger than Mount Everest.[4]

THE IMPOSSIBLE ATOM

Rutherford's picture of the atom as a miniature solar system with tiny electrons flitting about a dense atomic nucleus like planets around the Sun was a triumph of experimental science. Unfortunately, it had a slight problem. It was totally incompatible with all known physics!

According to Maxwell's theory of electromagnetism—which described all electrical and magnetic phenomena—whenever a charged particle accelerates, changing its speed or direction of motion, it gives out electromagnetic waves—light. An electron is a charged particle. As it circles a nucleus, it perpetually changes its direction; so it should act like a miniature lighthouse, constantly broadcasting light waves into space. The problem is that this would be a catastrophe for any atom. After all, the energy radiated as light has to come from somewhere, and it can only come from the electron itself. Sapped continually of energy, it should spiral ever closer to the centre of the atom. Calculations showed that it would collide with the nucleus within a mere hundred-millionth of a second. By rights, atoms should not exist.

[4] See Chapter 4, "Uncertainty and the Limits of Knowledge."

But atoms do exist. We and the world around us are proof enough of that. Far from expiring in a hundred-millionth of a second, atoms have survived intact since the earliest times of the Universe almost 14 billion years ago. Some crucial ingredient must be missing from Rutherford's picture of the atom. That ingredient is a revolutionary new kind of physics—quantum theory.

2

Why God Plays Dice with the Universe

How we discovered that things in the world of atoms happen for no reason at all

A philosopher once said, "It is necessary for the very existence of science that the same conditions always produce the same results." Well, they don't!

Richard Feynman

It's 2025 and high on a desolate mountain top a giant 100-metre telescope tracks around the night sky. It locks onto a proto-galaxy at the edge of the observable Universe and feeble light, which has been travelling through space since long before Earth was born, is concentrated by the telescope mirror onto ultrasensitive electronic detectors. Inside the telescope dome, seated at a control panel not unlike the console of the starship Enterprise, *the astronomers watch a fuzzy image of the galaxy swim into view on a computer monitor. Someone turns up a loudspeaker and a deafening crackle fills the control room. It sounds like machine gun fire; it sounds like rain drumming on a tin roof. In fact, it is the sound of tiny particles of light raining down on the telescope's detectors from the very depths of space.*

To these astronomers, who spend their careers straining to see the weakest sources of light in the Universe, it is a self-evident fact that light is a stream of tiny bulletlike particles—photons. Not long ago, however, the scientific community had to be dragged kicking and screaming to an acceptance of this idea. In fact, it's fair to say that the discovery that light comes in discrete chunks, or quanta, was the single most shocking discovery in the history of science. It swept away the comfort blanket of pre-20th-century science and exposed physicists to the harsh reality of an *Alice in Wonderland* universe where things happen because they happen, with utter disregard for the civilised laws of cause and effect.

The first person to realise that light was made of photons was Einstein. Only by imagining it as a stream of tiny particles could he make sense of a phenomenon known as the photoelectric effect. When you walk into a supermarket and the doors open for you automatically, they are being controlled by the photoelectric effect. Certain metals, when exposed to light, eject particles of electricity—electrons. When incorporated into a photocell, such a metal generates a small electric current as long as a light beam is falling on it. A shopper who breaks the beam chokes off the current, signalling the supermarket doors to swish aside.

One of the many peculiar characteristics of the photoelectric effect is that, even if a very weak light is used, the electrons are kicked out of the metal instantaneously—that is, with no delay whatsoever.[1] This is inexplicable if light is a wave. The reason is that a wave, being a spread-out thing, will interact with a large number of electrons in the metal. Some will inevitably be kicked out after others. In fact, some of

[1] Another interesting characteristic of the photoelectric effect is that no electrons at all are emitted by the metal if it is illuminated by light with a wavelength—a measure of the distance between successive wave crests—above a certain threshold. This, as Einstein realised, is because photons of light have an energy that goes down with increasing wavelength. And below a certain wavelength the photons have insufficient energy to kick an electron out of the metal.

the electrons could easily be emitted 10 minutes or so after light is shone on the metal.

So how is it possible that the electrons are kicked out of the metal instantaneously? There is only one way—if each electron is kicked out of the metal by *a single particle of light.*

Even stronger evidence that light consists of tiny bulletlike particles comes from the Compton effect. When electrons are exposed to X-rays—a high-energy kind of light—they recoil in exactly the way they would if they were billiard balls being struck by other billiard balls.

On the surface, the discovery that light behaves like a stream of tiny particles may not appear very remarkable or surprising. But it is. The reason is that there is also abundant and compelling evidence that light is something as different from a stream of particles as it is possible to imagine—a wave.

RIPPLES ON A SEA OF SPACE

At the beginning of the 19th century, the English physician Thomas Young, famous for decoding the Rosetta stone independently of the Frenchman Jean François Champollion, took an opaque screen, made two vertical slits in it very close together, and shone light of a single colour onto them. If light were a wave, he reasoned, each slit would serve as a new source of waves, which would spread out on the far side of the screen like concentric ripples on a pond.

A characteristic property exhibited by waves is interference. When two similar waves pass through each other, they reinforce each other where the crest of one wave coincides with the crest of another, and they cancel each other out where the crest of one coincides with the trough of the other. Look at a puddle during a rain shower and you will see the ripples from each raindrop spreading out and "constructively" and "destructively" interfering with each other.

In the path of the light emerging from his two slits Young interposed a second, white, screen. He immediately saw a series of alternating dark and light vertical stripes, much like the lines on a supermarket bar code. This interference pattern was irrefutable evidence that light was a wave. Where the light ripples from the two slits were in step, matching crest for crest, the light was boosted in brightness; where they were out of step, the light was cancelled out.

Using his "double slit" apparatus, Young was able to determine the wavelength of light. He discovered it was a mere thousandth of a millimetre—far smaller than the width of a human hair—explaining why nobody had guessed light was a wave before.

For the next two centuries, Young's picture of light as ripples on a sea of space reigned supreme in explaining all known phenomena involving light. But by the end of the 19th century, trouble was brewing. Although few people noticed at first, the picture of light as a wave and the picture of the atom as a tiny mote of matter were irreconcilable. The difficulty was at the interface, the place where light meets matter.

TWO FACES OF A SINGLE COIN

The interaction between light and matter is of crucial importance to the everyday world. If the atoms in the filament of a bulb did not spit out light, we could not illuminate our homes. If the atoms in the retina of your eye did not absorb light, you would be unable to read these words. The trouble is that the emission and absorption of light by atoms are impossible to understand if light is a wave.

An atom is a highly localised thing, confined to a tiny region of space, whereas a light wave is a spread-out thing that fills a large amount of space. So, when light is absorbed by an atom, how does such a big thing manage to squeeze into such a tiny thing? And when light is emitted by an atom, how does such a small thing manage to cough out such a big thing?

Common sense says that the only way light can be absorbed or emitted by a small localised thing is if it too is a small, localised thing. "Nothing fits inside a snake like another snake," as the saying goes. Light, however, is known to be a wave. The only way out of the conundrum was for physicists to throw up their hands in despair and grudgingly accept that light is both a wave and a particle. But surely something cannot be simultaneously localised and spreadout? In the everyday world, this is perfectly true. Crucially, however, we are not talking about the everyday world; we are talking about the microscopic world.

The microscopic world of atoms and photons turns out to be nothing like the familiar realm of trees and clouds and people. Since it is a domain millions of times smaller than the realm of familiar objects, why should it be? Light really is both a particle and a wave. Or more correctly, light is "something else" for which there is no word in our everyday language and nothing to compare it with in the everyday world. Like a coin with two faces, all we can see are its particle-like face and its wavelike face. What light *actually is* is as unknowable as the colour blue is to a blind man.

Light sometimes behaves like a wave and sometimes like a stream of particles. This was an extremely difficult thing for the physicists of the early 20th century to accept. But they had no choice; it was what nature was telling them. "On Mondays, Wednesdays and Fridays, we teach the wave theory and on Tuesdays, Thursdays and Saturdays the particle theory," joked the English physicist William Bragg in 1921.

Bragg's pragmatism was admirable. Unfortunately, it was not enough to save physics from disaster. As Einstein first realised, the dual wave-particle nature of light was a catastrophe. It was not just impossible to visualise, it was completely incompatible with all physics that had gone before.

WAVING GOODBYE TO CERTAINTY

Take a window. If you look closely you can see a faint reflection of your face. This is because glass is not perfectly transparent. It transmits about 95 per cent of the light striking it while reflecting the remaining 5 per cent. If light is a wave, this is perfectly easy to understand. The wave simply splits into a big wave that goes through the window and a much smaller wave that comes back. Think of the bow wave from a speedboat. If it encounters a half-submerged piece of driftwood, a large part of the wave continues on its way while a small part doubles back on itself.

But while this behaviour is easy to understand if light is a wave, it is extremely difficult to understand if light is a stream of identical bulletlike particles. After all, if all the photons are identical, it stands to reason that each should be affected by the window in an identical way. Think of David Beckham taking a free kick over and over again. If the soccer balls are identical and he kicks each one in exactly the same way, they will all curl through the air and hit the same spot at the back of goal. It's hard to imagine the majority of the balls peppering the same spot while a minority flies off to the corner flag.

How, then, is it possible that a stream of absolutely identical photons can impinge on a window and 95 per cent can go right through while 5 per cent come back? As Einstein realised, there is only one way: if the word "identical" has a very different meaning in the microscopic world than in the everyday world—a diminished, cut-down meaning.

In the microscopic domain, it turns out, identical things do not behave in identical ways in identical circumstances. Instead, they merely have an identical *chance* of behaving in any particular way. Each individual photon arriving at the window has exactly the same *chance* of being transmitted as any of its fellows—95 per cent—and exactly the same *chance* of being reflected—5 per cent. There is absolutely no way to know for certain what will happen to a given photon. Whether it is transmitted or reflected is entirely down to random chance.

In the early 20th century, this unpredictability was something radically new in the world. Imagine a roulette wheel and a ball jouncing around as the wheel spins. We think of the number the ball comes to rest on when the wheel finally halts as inherently unpredictable. But it is not—not really. If it were possible to know the initial trajectory of the ball, the initial speed of the wheel, the way the air currents changed from instant to instant in the casino, and so on, the laws of physics could be used to predict with 100 per cent certainty where the ball will end up. The same is true with the tossing of a coin. If it were possible to know how much force is applied in the flipping, the exact shape of the coin, and so on, the laws of physics could predict with 100 per cent certainty whether the coin will come down heads or tails.

Nothing in the everyday world is fundamentally unpredictable; nothing is truly random. The reason we cannot predict the outcome of a game of roulette or of the toss of a coin is that there is simply too much information for us to take into account. But in principle—and this is the key point—there is nothing to prevent us from predicting both.

Contrast this with the microscopic world of photons. It matters not the slightest how much information we have in our possession. It is impossible to predict whether a given photon will be transmitted or reflected by a window—even in principle. A roulette ball does what it does for a reason—because of the interplay of myriad subtle forces. A photon does what it does for no reason whatsoever! The unpredictability of the microscopic world is fundamental. It is truly something new under the Sun.

And what is true of photons turns out to be true of all the denizens of the microscopic realm. A bomb detonates because its timer tells it to or because a vibration disturbs it or because its chemicals have suddenly become degraded. An unstable, or "radioactive," atom simply detonates. There is absolutely no discernible difference between one that detonates at this moment and an identical atom that waits quietly for 10 million years before blowing itself to pieces. The shocking truth, which stares you in the face every time you look at a window, is that the whole Universe is founded on random chance. So upset was Einstein by this idea that he stuck out his lip and declared: "God does not play dice with the Universe!"

The trouble is He does. As British physicist Stephen Hawking has wryly pointed out: "Not only does God play dice with the Universe, he throws the dice where we cannot see them!"

When Einstein received the Nobel Prize for Physics in 1921 it was not for his more famous theory of relativity but for his explanation of the photoelectric effect. And this was no aberration on the part of the Nobel committee. Einstein himself considered his work on the "quantum" the only thing he ever did in science that was truly revolutionary. And the Nobel committee completely agreed with him.

Quantum theory, born out of the struggle to reconcile light and matter, was fundamentally at odds with all science that had gone before. Physics, pre-1900, was basically a recipe for predicting the future with absolute certainty. If a planet is in a particular place now, in a day's time it will have moved to another place, which can be predicted with 100 per cent confidence by using Newton's laws of motion and the law of gravity. Contrast this with an atom flying through space. Nothing is knowable with certainty. All we can ever predict is its probable path, its probable final position.

Whereas quantum is based on uncertainty, the rest of physics is based on certainty. To say this is a problem for physicists is a bit of an understatement! "Physics has given up on the problem of trying to predict what would happen in a given circumstance," said Richard Feynman. "We can only predict the odds."

All is not lost, however. If the microworld were totally unpredictable, it would be a realm of total chaos. But things are not this bad. Although what atoms and their like get up to is intrinsically unpredictable, it turns out that the unpredictability is at least predictable!

PREDICTING THE UNPREDICTABILITY

Think of the window again. Each photon has a 95 per cent chance of being transmitted and a 5 per cent chance of being reflected. But what determines these probabilities?

Well, the two different pictures of light—as a particle and as a wave—must produce the same outcome. If half the wave goes through and half is reflected, the only way to reconcile the wave and particle pictures is if each individual particle of light has a 50 per cent probability of being transmitted and a 50 per cent probability of being reflected. Similarly, if 95 per cent of the wave is transmitted and 5 per cent is reflected, the corresponding probabilities for the transmission and reflection of individual photons must be 95 per cent and 5 per cent.

To get agreement between the two pictures of light, the particle-like aspect of light must somehow be "informed" about how to behave by its wavelike aspect. In other words, in the microscopic domain, waves do not simply behave like particles; those particles behave like waves as well! There is perfect symmetry. In fact, in a sense this statement is all you need to know about quantum theory (apart from a few details). Everything else follows unavoidably. All the weirdness, all the amazing richness of the microscopic world, is a direct consequence of this wave-particle "duality" of the basic building blocks of reality.

But how exactly does light's wavelike aspect inform its particle-like aspect about how to behave? This is not an easy question to answer.

Light reveals itself either as a stream of particles or as a wave. We never see both sides of the coin at the same time. So when we observe light as a stream of particles, there is no wave in existence to inform those particles about how to behave. Physicists therefore have a problem in explaining the fact that photons do things—for instance, fly through windows—as if directed by a wave.

They solve the problem in a peculiar way. In the absence of a real wave, they imagine an abstract wave—a mathematical wave. If this sounds ludicrous, this was pretty much the reaction of physicists when the idea was first proposed by the Austrian physicist Erwin Schrödinger in the 1920s. Schrödinger imagined an abstract mathematical wave that spread through space, encountering obstacles and being reflected and transmitted, just like a water wave spreading on a pond. In places where the height of the wave was large, the probability of finding a particle was highest, and in locations where it was small, the probability was lowest. In this way Schrödinger's wave of probability christened the wave function, informed a particle what to do, and not just a photon—any microscopic particle, from an atom to a constituent of an atom like an electron.

There is a subtlety here. Physicists could make Schrödinger's picture accord with reality only if the probability of finding a particle at any point was related to the square of the height of the probability wave at that point. In other words, if the probability wave at some point in space is twice as high as it is at another point in space, the particle is four times as likely to be found there than at the other place.

The fact that it is the square of the probability wave and not the probability wave itself that has real physical meaning to this day causes debate about whether the wave is a real thing lurking beneath the skin of the world or just a convenient mathematical device for calculating things. Most but not all people favour the latter.

The probability wave is crucially important because it makes a connection between the wavelike aspect of matter and familiar waves of all kinds, from water waves to sound waves to earthquake waves. All obey a so-called wave equation. This describes how they ripple through space and allows physicists to predict the wave height at any location at any time. Schrödinger's great triumph was to find the wave equation that described the behaviour of the probability wave of atoms and their like.

By using the Schrödinger equation, it is possible to determine the probability of finding a particle at any location in space at any time. For instance, it can be used to describe photons impinging on the obstacle of a windowpane and to predict the 95 per cent probability of finding one on the far side of the pane. In fact, the Schrödinger equation can be used to predict the probability of any particle, be it a photon or an atom, doing just about anything. It provides the crucial bridge to the microscopic world, allowing physicists to predict everything that happens there—if not with 100 per cent certainty, at least with predictable uncertainty!

Where is all this talk of probability waves leading? Well, the fact that waves behave like particles in the microscopic world leads unavoidably to the realisation that the microscopic world dances to an entirely different tune than that of the everyday world. It is governed by random unpredictability. This in itself was a shocking, confidence-draining blow to physicists and their belief in a predictable, clockwork universe. But this, it turns out, is only the beginning. Nature had many more shocks in store. The fact that waves not only behave as particles but also that those particles behave as waves leads to the realisation that all the things that familiar waves, like water waves and sound waves, can do, so too can the probability waves that inform the behaviour of atoms, photons, and their kin.

So what? Well, waves can do an awful lot of different things. And each of these things turns out to have a semi-miraculous consequence in the microscopic world. The most straightforward thing waves can do is exist as superpositions. Remarkably, this enables an atom to be in two places at once, the equivalent of you being in London and New York at the same time.

3

The Schizophrenic Atom

How an atom can be in many places at once and do many things at once

> *If you imagine the difference between an abacus and the world's fastest supercomputer, you would still not have the barest inkling of how much more powerful a quantum computer could be compared with the computers we have today.*
>
> Julian Brown

It's 2041. A boy sits at a computer in his bedroom. It's not an ordinary computer. It's a quantum computer. The boy gives the computer a task . . . and instantly it splits into thousands upon thousands of versions of itself, each of which works on a separate strand of the problem. Finally, after just a few seconds, the strands come back together and a single answer flashes on the computer display. It's an answer that all the normal computers in the world put together would have taken a trillion trillion years to find. Satisfied, the boy shuts the computer down and goes out to play, his night's homework done.

Surely, no computer could possibly do what the boy's computer has just done? Not only could a computer do such a thing, crude versions are already in existence today. The only thing in serious dispute is whether such a quantum computer merely behaves like a vast multiplicity of computers or whether, as some believe, it literally exploits the computing power of multiple copies of itself existing in parallel realities, or universes.

The key property of a quantum computer—the ability to do many calculations at once—follows directly from two things that waves—and therefore microscopic particles such as atoms and photons, which behave like waves—can do. The first of those things can be seen in the case of ocean waves.

On the ocean there are both big waves and small ripples. But as anyone who has watched a heavy sea on a breezy day knows, you can also get big, rolling waves with tiny ripples superimposed on them. This is a general property of all waves. If two different waves can exist, so too can a combination, or superposition, of the waves. The fact that superpositions can exist is pretty innocuous in the everyday world. However, in the world of atoms and their constituents, its implications are nothing short of earth-shattering.

Think again of a photon impinging on a windowpane. The photon is informed about what to do by a probability wave, described by the Schrödinger equation. Since the photon can either be transmitted or reflected, the Schrödinger equation must permit the existence of two waves—one corresponding to the photon going through the window and another corresponding to the photon bouncing back. Nothing surprising here. However, remember that, if two waves are permitted to exist, a superposition of them is also permitted to exist. For waves at sea such a combination is nothing out of the ordinary. But here the combination corresponds to something quite extraordinary—the photon being both transmitted and reflected. In other words, the photon can be on both sides of the windowpane simultaneously!

And this unbelievable property follows unavoidably from just two facts: that photons are described by waves and that superpositions of waves are possible.

This is no theoretical fantasy. In experiments it is actually possible to observe a photon or an atom being in two places at once—the everyday equivalent of you being in San Francisco and Sydney at the same time. (More accurately, it is possible to observe the *consequences* of a photon or an atom being in two places at once.) And since there is no limit to the number of waves that can be superposed, a photon or an atom can be in three places, 10 places, a million places at once.

But the probability wave associated with a microscopic particle does more than inform it where it could be *located*. It informs it *how to behave* in all circumstances—telling a photon, for instance, whether or not to be transmitted or reflected by a pane of glass. Consequently, atoms and their like can not only be in many places at once, they can *do many things at once*, the equivalent of you cleaning the house, walking the dog, and doing the weekly supermarket shopping all at the same time. This is the secret behind the prodigious power of a quantum computer. It exploits the ability of atoms to do many things at once, to do many calculations at once.

DOING MANY THINGS AT ONCE

The basic elements of a conventional computer are transistors. These have two distinct voltage states, one of which is used to represent the binary digit, or bit, "0", the other to represent a "1." A row of such zeros and ones can represent a large number, which in the computer can be added, subtracted, multiplied, and divided by another large number.[1] But in a quantum computer the basic elements—which may be single atoms—can be in a superposition of states. In other words, they can represent a zero and a one simultaneously. To distinguish them from normal bits, physicists call such schizophrenic entities quantum bits, or qubits.

[1] Binary was invented by the 17th-century mathematician Gottfried Leibniz. It is a way of representing numbers as a strings of zeros and ones. Usually, we use decimal, or base 10. The right-hand digit represents the ones, the next digit the tens, the next the 10 × 10s, and so on. So, for instance, 9,217 means 7 + 1 × 10 + 2 × (10 × 10) + 9 × (10 × 10 × 10). In binary, or base 2, the right-hand digit represents the ones, the next digit the twos, the next the 2 × 2s, and so on. So for instance, 1101 means 1 + 0 × 2 + 1 × (2 × 2) + 1 × (2 × 2 × 2), which in decimal is 13.

One qubit can be in two states (0 or 1), two qubits in four (00 or 01 or 10 or 11), three qubits in eight, and so on. Consequently, when you calculate with a single qubit, you can do two calculations simultaneously, with two qubits four calculations, with three eight, and so on. If this doesn't impress you, with 10 qubits you could do 1,024 calculations all at once, with 100 qubits 100 billion billion billion! Not surprisingly, physicists positively salivate at the prospect of quantum computers. For some calculations, they could massively outperform conventional computers, making conventional personal computers appear positively retarded.

But for a quantum computer to work, wave superpositions are not sufficient on their own. They need another essential wave ingredient: interference.

The interference of light observed by Thomas Young in the 18th century was the key observation that convinced everyone that light was a wave. When, at the beginning of the 20th century, light was also shown to behave like a stream of particles, Young's double slit experiment assumed a new and unexpected importance—as a means of exposing the central peculiarity of the microscopic world.

INTERFERENCE IS THE KEY

In the modern incarnation of Young's experiment, a double slit in an opaque screen is illuminated with light, which is undeniably a stream of particles. In practice, this means using a light source so feeble that it spits out photons one at a time. Sensitive detectors at different positions on the second screen count the arrival of photons. After the experiment has been running for a while, the detectors show something remarkable. Some places on the screen get peppered with photons while other places are completely avoided. What is more, the places that are peppered by photons and the places that are avoided alternate, forming vertical stripes—exactly as in Young's original experiment.

But wait a minute! In Young's experiment the dark and light bands are caused by interference. And a fundamental feature of interference is that it involves the mingling of two sets of waves from the same source—the light from one slit with the light from the other slit. But in this case the photons are arriving at the double slit one at a time. Each photon is completely alone, with no other photon to mingle with. How, then, can there be any interference? How can it know where its fellow photons will land?

There would appear to be only one way—if each photon somehow goes through both slits simultaneously. Then it can interfere with itself. In other words, each photon must be in a superposition of two states—one a wave corresponding to a photon going through the left-hand slit and the other a wave corresponding to a photon going through the right-hand slit.

The double slit experiment can be done with photons or atoms or any other microscopic particles. It shows graphically how the behaviour of such particles—where they can and cannot strike the second screen—is orchestrated by their wavelike alter ego. But this is not all the double slit experiment demonstrates. Crucially, it shows that the individual waves that make up a superposition are not passive but can actively interfere with each other. It is this ability of the individual states of a superposition to interfere with each other that is the absolute key to the microscopic world, spawning all manner of weird quantum phenomena.

Take quantum computers. The reason they can carry out many calculations at once is because they can exist in a superposition of states. For instance, a 10-element quantum computer is simultaneously in 1,024 states and can therefore carry out 1,024 calculations at once. But all the parallel strands of a calculation are of absolutely no use unless they get woven together. Interference is the means by which this is accomplished. It is the means by which the 1,024 states of the superposition can interact and influence each other. Because of interference, the single answer coughed out by the quantum computer is able to reflect and synthesise what was going on in all those 1,024 parallel calculations.

Think of a problem divided into 1,024 separate pieces and one person working on each piece. For the problem to be solved, the 1,024 people must communicate with each other and exchange results. This is what interference makes possible in a quantum computer.

An important point worth making here is that, although superpositions are a fundamental feature of the microscopic world, it is a curious property of reality that they are never actually observed. All we ever see are the consequences of their existence—what results when the individual waves of a superposition *interfere* with each other. In the case of the double slit experiment, for instance, all we ever see is an interference pattern, from which we infer that an electron was in a superposition in which it went through both slits simultaneously. It is impossible to actually *catch* an electron going through both slits at once. This is what was meant by the earlier statement that it is possible only to observe the *consequences* of an atom being in two places at once, not it actually being in two places at once.

MULTIPLE UNIVERSES

The extraordinary ability of quantum computers to do enormous numbers of calculations simultaneously poses a puzzle. Though practical quantum computers are currently at a primitive stage, manipulating only a handful of qubits, it is nevertheless possible to imagine a quantum computer that can do billions, trillions, or quadrillions of calculations simultaneously. In fact, it is quite possible that in 30 or 40 years we will be able to build a quantum computer that can do more calculations simultaneously than there are particles in the Universe. This hypothetical situation poses a sticky question: Where exactly will such a computer be doing its calculations? After all, if such a computer can do more calculations simultaneously than there are particles in the Universe, it stands to reason that the Universe has insufficient computing resources to carry them out.

One extraordinary possibility, which provides a way out of the conundrum, is that a quantum computer does its calculations in parallel realities or universes. The idea goes back to a Princeton graduate student named Hugh Everett III, who, in 1957, wondered why quantum theory is such a brilliant description of the microscopic world of atoms but we never actually see superpositions. Everett's extraordinary answer was that each state of the superposition exists in a totally separate reality. In other words, there exists a multiplicity of realities—a *multiverse*—where all possible quantum events occur.

Although Everett proposed his "Many Worlds" idea long before the advent of quantum computers, it can shed some helpful light on them. According to the Many Worlds idea, when a quantum computer is given a problem, it splits into multiple versions of itself, each living in a separate reality. This is why the boy's quantum personal computer at the start of this chapter split into so many copies. Each version of the computer works on a strand of the problem, and the strands are brought together by interference. In Everett's picture, therefore, interference has a very special significance. It is the all-important *bridge* between separate universes, the means by which they interact and influence each other.

Everett had no idea *where* all the parallel universes were located. And, frankly, nor do the modern-day proponents of the Many Worlds idea. As Douglas Adams wryly observed in *The Hitchhiker's Guide to the Galaxy:* "There are two things you should remember when dealing with parallel universes. One, they're not really parallel, and two, they're not really universes!"

Despite such puzzles, half a century after Everett proposed the Many Worlds idea, it is undergoing an upsurge in popularity. An increasing number of physicists, most notably David Deutsch of the University of Oxford, are taking it seriously. "The quantum theory of parallel universes is not some troublesome, optional interpretation emerging from arcane theoretical considerations," says Deutsch in his book, *The Fabric of Reality.* "It is *the* explanation—the only one that is tenable—of a remarkable and counterintuitive reality."

If you go along with Deutsch—and the Many Worlds idea predicts exactly the same outcome for every conceivable experiment as more conventional interpretations of quantum theory—then quantum computers are something radically new under the Sun. They are the very first machines humans have ever built that exploit the resources of multiple realities. Even if you do not believe the Many Worlds idea, it still provides a simple and intuitive way of imagining what is going on in the mysterious quantum world. For instance, in the double slit experiment, it is not necessary to imagine a single photon going through both slits simultaneously and interfering with itself. Instead, a photon going through one slit interferes with another photon going through the other slit. What other photon, you may ask? A photon in a neighbouring universe, of course!

WHY ARE ONLY SMALL THINGS QUANTUM?

Quantum computers are extremely difficult to build. The reason is that the ability of the individual states of a quantum superposition to interfere with each other is destroyed, or severely degraded, by the environment. This destruction can be vividly seen in the double slit experiment.

If some kind of particle detector is used to spot a particle going through one of the slits, the interference stripes on the screen immediately vanish, to be replaced by more or less uniform illumination. The act of observing which slit the particle goes through is all that is needed to destroy the superposition in which it goes through both slits simultaneously. And a particle going through one slit only is as likely to exhibit interference as you are to hear the sound of one hand clapping.

What has really happened here is that an attempt has been made to locate, or measure, the particle by the outside world. Knowledge of the superposition by the outside world is all that is needed to destroy it. It is almost as if quantum superpositions are a secret. Of course, once the world knows about the secret, the secret no longer exists!

Superpositions are *continually* being measured by their environment. And it takes only a single photon to bounce off a superposition and take information about it to the rest of the world to destroy the superposition. This process of natural measurement is called decoherence. It is the ultimate reason we do not see weird quantum behaviour in the everyday world.[2] Although naively we may think of quantum behaviour as a property of small things like atoms but not of big things like people and trees, this is not necessarily so. Quantum behaviour is actually a property of isolated things. The reason we see it in the microscopic world but not in the everyday world is simply because it is easier to isolate a small thing from its surroundings than a big thing.

The price of quantum schizophrenia is therefore isolation. As long as a microscopic particle like an atom can remain isolated from the outside world, it can do many different things at once. This is not difficult in the microscopic world, where quantum schizophrenia is an everyday phenomenon. However, in the large-scale world in which we live, it is nearly impossible, with countless quadrillions of photons bouncing off every object every second.

Keeping a quantum computer isolated from its surroundings is the main obstacle facing physicists in trying to construct such a machine. So far, the biggest quantum computer they have managed to build has been composed of only 10 atoms, storing 10 qubits. Keeping 10 atoms isolated from their surroundings for any length of time takes all their ingenuity. If a single photon bounces off the computer, 10 schizophrenic atoms instantly become 10 ordinary atoms.

[2] I am totally aware that all this talk of quantumness being a "secret" that is destroyed if the rest of the world learns about it is a complete fudge. But it is sufficient for our discussion here. Decoherence, the means by which the quantum world, with its schizophrenic superpositions, becomes the everyday world where trees and people are never in two places at once, is a can of worms with which the experts are still wrestling. For a real explanation, see Chapter 5, "The Telepathic Universe."

Decoherence illustrates a limitation of quantum computers not often publicised amid the hype surrounding such devices. To extract an answer, someone from the outside world—you—must interact with it, and this necessarily destroys the superposition. The quantum computer reverts to being an ordinary computer in a single state. A 10-qubit machine, instead of spitting out the answers to 1,024 separate calculations, spits out just one.

Quantum computers are therefore restricted to parallel calculations that output only a single answer. Consequently, only a limited number of problems are suited to solution by quantum computer, and much ingenuity is required to find them. They are not, as is often claimed, the greatest thing since sliced bread. Nevertheless, when a problem is found that plays to the strengths of a quantum computer, it can massively outperform a conventional computer, calculating in seconds what otherwise might take longer than the lifetime of the Universe.

On the other hand, decoherence, which is the greatest enemy of those struggling to build quantum computers, is also their greatest friend. It is because of decoherence, after all, that the giant superposition of a quantum computer with all its mutually interfering strands is finally destroyed; it is only by being destroyed—reduced to a single state representing a single answer—that anything useful comes out of such a machine. The world of the quantum is indeed a paradoxical one!

4

Uncertainty and the Limits of Knowledge

Why we can never know all we would like to know about atoms and why this fact makes atoms possible

> *Passing farther through the quantum land our travelers met quite a lot of other interesting phenomena, such as quantum mosquitoes, which could scarcely be located at all, owing to their small mass.*
>
> George Gamow

He must be going mad. Only moments before he had parked his shiny red Ferrari in the garage. He had even stood there on the driveway, admiring his pride and joy until the last possible moment, as the automatic door swung shut. But then as he crunched across the gravel to his front door there had been a curious rustling of the air, a faint tremor of the ground. He had wheeled round. And there, squatting back on his driveway, in front of the still-locked garage doors, was his beautiful red Ferrari!

Such Houdini-like feats of escapology are never of course seen in the everyday world. In the realm of the ultrasmall, however, they are a common occurrence. One instant an atom can be locked up in a microscopic prison; the next it has shed its shackles and slipped away silently into the night.

This miraculous ability to escape escape-proof prisons is entirely due to the wavelike face of microscopic particles, which enables atoms and their constituents to do all the things that waves can do. And one of the many things waves can do is penetrate apparently impenetrable barriers. This is not an obvious or well-known wave property. But it can be demonstrated by a light beam travelling through a block of glass and trying to escape into the air beyond.

The key thing is what happens at the edge of the glass block, the boundary where the glass meets the air. If the light happens to strike the boundary at a shallow angle, it gets reflected back into the glass block and fails to escape into the air beyond. In effect, it is imprisoned in the glass. However, something radically different happens if another block of glass is brought close to the boundary, leaving a small gap of air between the two blocks. Just as before, some of the light is reflected back into the glass. But—and this is the crucial thing—some of the light now leaps the air gap and travels into the second glass block.

The parallel between the Ferrari escaping its garage and the light escaping the block of glass may not be very obvious. However, for all intents and purposes, the air gap should be just as impenetrable a barrier to the light as the garage walls are to the Ferrari.

The reason the light wave can penetrate the barrier and escape from the block of glass is that a wave is not a localised thing but something spread out through space. So when the light waves strike the glass-air boundary and are reflected back into the glass, they are not actually reflected from the exact boundary of the glass. Instead, they penetrate a short distance into the air beyond. Consequently, if they encounter another block of glass before they can turn back, they can continue on their way. Place a second glass block within a hair's breadth of the first and, hey presto, the light jumps the air gap and escapes its prison.

This ability to penetrate an apparently impenetrable barrier is common to all types of waves, from light waves to sound waves to the probability waves associated with atoms. It therefore manifests itself in the microscopic world. Arguably, the most striking example is the phenomenon of alpha decay in which an alpha particle breaks out of the apparently escape-proof prison of an atomic nucleus.

BREAKING OUT OF A NUCLEUS

An alpha particle is the nucleus of a helium atom. An unstable, or radioactive, nucleus sometimes spits out an alpha particle in a desperate attempt to turn itself into a lighter and more stable nucleus. The process poses a big puzzle, however. By rights, an alpha particle should not be able to get out of a nucleus.

Think of an Olympic high jumper penned in by a 5-metre-high metal fence. Even though he is one of the best high jumpers in the world, there is no way he can jump over a fence that high. No human being alive has sufficient strength in their legs. Well, an alpha particle inside an atomic nucleus finds itself in a similar position. The barrier that pens it in is created by the nuclear forces that operate inside a nucleus, but it is just as impenetrable a barrier to the alpha particle as the solid metal fence is to the high jumper.

Contrary to all expectations, however, alpha particles do escape from atomic nuclei. And their escape is entirely due to their wavelike face. Like light waves trapped in a glass block, they can penetrate an apparently impenetrable barrier and slip away quietly into the outside world.

This process is called quantum tunnelling and alpha particles are said to "tunnel" out of an atomic nucleus. Tunnelling is actually an instance of a more general phenomenon known as uncertainty, which puts a fundamental limit on what we can and cannot know about the microscopic world. The double slit experiment is an excellent demonstration of uncertainty.

THE HEISENBERG UNCERTAINTY PRINCIPLE

The reason a microscopic particle like an electron can go through both slits in the screen simultaneously is that it can exist as a superposition of two waves—one wave corresponding to the particle going through one slit and the other to the particle going through the other slit. But that is not sufficient to guarantee that its schizophrenic behaviour will be noticed. For that to happen, an interference pattern must appear on the second screen. But this, of course, requires the individual waves in the superposition to interfere. The fact that interference is a crucial ingredient for the electron to exhibit weird quantum behaviour turns out to have profound implications for what nature permits us to know about the electron.

Say in the double slit experiment we try to locate the slit each electron goes through. If we succeed, the interference pattern on the second screen disappears. After all, interference requires that two things mingle. If the electron and its associated probability wave go through only one slit, there is only one thing.

How, in practice, could we locate which slit an electron goes through? Well, to make the double slit experiment a bit easier to visualise, think of an electron as a bullet from a machine gun and the screen as a thick metal sheet with two vertical parallel slits. When bullets are fired at the screen, some enter the slits and go through. Think of the slits as deep channels cut through the thick metal. The bullets ricochet off the internal walls of the channels and by this means reach the second screen. They can obviously hit any point on the second screen. But, for simplicity, imagine they end up at the midpoint of the second screen. Also for simplicity, say that at this point the probability waves associated with the bullets interfere constructively, so it is a place that gets peppered with lots of bullets.

Now, when a bullet ricochets off the inside of a slit, it causes the metal screen to recoil in the opposite direction. It's the same if you are playing tennis and a fast serve ricochets off your racquet. Your racquet recoils in the opposite direction. Crucially, the recoil of the screen can be used to deduce which slit a bullet goes through. After all, if the screen moves to the left, the bullet must have gone through the left-hand slit; if it moves to the right, it must have been the right-hand slit.

However, we know that if we locate which slit a bullet goes through, it destroys the interference pattern on the second screen. This is straightforward to understand from the wave point of view. We are as unlikely to see one thing interfere with itself as we are to hear the sound of one hand clapping. But how do we make sense of things from the equally valid particle point of view?

Remember that the interference pattern on the second screen is like a supermarket bar code. It consists of vertical "stripes" where no bullets hit, alternating with vertical stripes where lots of bullets hit. For simplicity, think of the stripes as black and white. The key question therefore is: From the bullet's point of view, what would it take to destroy the interference pattern?

The answer is a little bit of sideways jitter. If each bullet, instead of flying unerringly towards a black stripe, possesses a little sideways jitter in its trajectory so that it can hit either the black stripe or an adjacent white stripe, this will be sufficient to "smear out" the interference pattern. Stripes that were formerly white will become blacker, and stripes that were formerly black will become whiter. The net result will be a uniform gray. The interference pattern will be smeared out.

Because it must be impossible to tell whether a given bullet will hit a black stripe or an adjacent white stripe (or vice versa), the jittery sideways motion of each bullet must be entirely unpredictable. And all this must come to pass for no other reason than that we are locating which slit each bullet goes through by the recoil of the screen.

In other words, the very act of pinning down the location of a particle like an electron adds unpredictable jitter, making its velocity uncertain. And the opposite is true as well. The act of pinning down the velocity of a particle makes its location uncertain. The first person to recognise and quantify this effect was the German physicist Werner Heisenberg, and it is called the Heisenberg uncertainty principle in his honour.

According to the uncertainty principle, it is impossible to know both the location and the velocity of a microscopic particle with complete certainty. There is a trade-off, however. The more precisely its location is pinned down, the more uncertain is its velocity. And the more precisely its velocity is pinned down, the more uncertain its location.

Imagine if this constraint also applied to what we could know about the everyday world. If we had precise knowledge of the speed of a jet aeroplane, we would not be able to tell whether it was over London or New York. And if we had precise knowledge of the location of the aeroplane, we would be unable to tell whether it was cruising at 1,000 kilometres per hour or 1 kilometre per hour—and about to plummet out of the sky.

The uncertainty principle exists to *protect* quantum theory. If you could measure the properties of atoms and their like better than the uncertainty principle permits, you would destroy their wave behaviour—specifically, interference. And without interference, quantum theory would be impossible. Measuring the position and velocity of a particle with greater accuracy than the uncertainty principle dictates must therefore be impossible. Because of the Heisenberg uncertainty principle, when we try to look closely at the microscopic world, it starts to get fuzzy, like a newspaper picture that has been overmagnified. Infuriatingly, nature does not permit us to measure precisely all we would like to measure. There is a limit to our knowledge.

This limit is not simply a quirk of the double slit experiment. It is fundamental. As Richard Feynman remarked: "No one has ever found (or even thought of) a way around the uncertainty principle. Nor are they ever likely to."

It is because alpha particles have a wavelike character that they can escape the apparently escape-proof prison of an atomic nucleus.

However, the Heisenberg uncertainty principle makes it possible to understand the phenomenon from the particle point of view.

GOING WHERE NO HIGH JUMPER HAS GONE BEFORE

Recall that an alpha particle in a nucleus is like an Olympic high jumper corralled by a 5-metre-high fence. Common sense says that it is moving about inside the nucleus with insufficient speed to launch itself over the barrier. But common sense applies only to the everyday world, not to the microscopic world. Ensnared in its nuclear prison, the alpha particle is very localised in space—that is, its position is pinned down with great accuracy. According to the Heisenberg uncertainty principle, then, its velocity must necessarily be very uncertain. It could, in other words, be much greater than we think. And if it is greater, then, contrary to all expectations, the alpha particle can leap out of the nucleus—a feat comparable to the Olympic high jumper jumping the 5-metre fence.

Alpha particles emerge into the world outside their prison as surprisingly as the Ferrari emerged into the world outside its garage. And this "tunnelling" is due to the Heisenberg uncertainty principle. But tunnelling is a two-way process. Not only can subatomic particles like alpha particles tunnel out of a nucleus, they can tunnel into it too. In fact, such tunnelling in reverse helps explain a great mystery: why the Sun shines.

TUNNELLING IN THE SUN

The Sun generates heat by gluing together protons—the nuclei of hydrogen atoms—to make the nuclei of helium atoms.[1] This nuclear fusion produces as a by-product a dam burst of nuclear binding energy, which ultimately emerges from the Sun as sunlight.

[1] See Chapter 8, "$E = mc^2$ and the Weight of Sunshine."

But hydrogen fusion has a problem. The force of attraction that glues together protons—the "strong nuclear force"—has an extremely short range. For two protons in the Sun to come under its influence and be snapped together, they must pass extremely close to each other. But two protons, by virtue of their similar electric charge, repel each other ferociously. To overcome this fierce repulsion, the protons must collide at enormous speed. In practice, this requires the core of the Sun, where nuclear fusion goes on, to be at an extremely high temperature.

Physicists calculated the necessary temperature in the 1920s, just as soon as it was suspected that the Sun was running on hydrogen fusion. It turned out to be roughly 10 billion degrees. This, however, posed a problem. The temperature at the heart of the Sun was known to be only about 15 million degrees—roughly a thousand times lower. By rights, the Sun should not be shining at all. Enter the German physicist Fritz Houtermans and the English astronomer Robert Atkinson.

When a proton in the core of the Sun approaches another proton and is pushed back by its fierce repulsion, it is just as if it encounters a high brick wall surrounding the second proton. At the 15 million degrees temperature in the heart of the Sun, the proton would appear to be moving far too slowly to jump the wall. However, the Heisenberg uncertainty principle changes everything.

In 1929, Houtermans and Atkinson carried out the relevant calculations. They discovered that the first proton can tunnel through the apparently impenetrable barrier around the second proton and successfully fuse with it even at the ultralow temperature of 15 million degrees. What is more, this explains perfectly the observed heat output of the Sun.

The night after Houtermans and Atkinson did the calculation, Houtermans reportedly tried to impress his girlfriend with a line that nobody in history had used before. As they stood beneath a perfect moonless sky, he boasted that he was the only person in the world who knew why the stars were shining. It must have worked. Two years later, Charlotte Riefenstahl agreed to marry him. (Actually, she married him twice, but that's another story.)

Sunlight apart, the Heisenberg uncertainty principle explains something much closer to home: the very existence of the atoms in our bodies.

UNCERTAINTY AND THE EXISTENCE OF ATOMS

By 1911 the Cambridge experiments of New Zealand physicist Ernest Rutherford had revealed the atom as resembling a miniature solar system. Tiny electrons flitted about a compact atomic nucleus much like planets around the Sun. However, according to Maxwell's theory of electromagnetism, an orbiting electron should radiate light energy and, within a mere hundred-millionth of a second, spiral into the nucleus. "Atoms," as Richard Feynman pointed out, "are completely impossible from the classical point of view." But atoms do exist. And the explanation comes from quantum theory.

An electron cannot get too close to a nucleus because, if it did, its location in space would be very precisely known. But according to the Heisenberg uncertainty principle, this would mean that its velocity would be very uncertain. It could become enormously huge.

Imagine an angry bee in a shrinking box. The smaller the box gets, the angrier the bee and the more violently it batters itself against the walls of its prison. This is pretty much the way an electron behaves in an atom. If it were squeezed into the nucleus itself, it would acquire an enormous speed—far too great to stay confined in the nucleus.

The Heisenberg uncertainty principle, which explains why electrons do not spiral into their nuclei, is therefore the ultimate reason why the ground beneath our feet is solid. But the principle does more than simply explain the existence of atoms and the solidity of matter. It explains why atoms are so big—or at least so much bigger than the nuclei at their cores.

WHY ATOMS ARE SO BIG

Recall that a typical atom is about 100,000 times bigger than the nucleus at its centre. Understanding why there is such a fantastic amount of empty space in atoms requires being a bit more precise about the Heisenberg uncertainty principle. Strictly speaking, it says that it is a particle's position and momentum—rather than just its velocity—that cannot simultaneously be determined with 100 per cent certainty.

The momentum of a particle is the product of its mass and velocity. It's really just a measure of how difficult it is to stop something that is moving. A train, for instance, has a lot of momentum compared to a car, even if the car is going faster. A proton in an atomic nucleus is about 2,000 times more massive than an electron. According to the Heisenberg uncertainty principle, then, if a proton and an electron are confined in the same volume of space, the electron will be moving about 2,000 times faster.

Already, we get an inkling of why the electrons in an atom must have a far bigger volume to fly about in than the protons and neutrons in the nucleus. But atoms are not just 2,000 times bigger than their nuclei; they are more like 100,000 times bigger. Why?

The answer is that an electron in an atom and a proton in a nucleus are not in the grip of the same force. While the nuclear particles are held by the powerful "strong nuclear" force, the electrons are held by the much weaker electric force. Think of the electrons flying about the nucleus attached to gossamer threads of elastic while the protons and the neutrons are constrained by elastic 50 times thicker. Here is the explanation for why the atom is a whopping 100,000 times bigger than the nucleus.

But the electrons in an atom do not orbit at one particular distance from the nucleus. They are permitted to orbit at a range of distances. Explaining this requires resorting to yet another wave picture—this one involving organ pipes!

OF ATOMS AND ORGAN PIPES

There are always many different ways of looking at things in the quantum world, each a glimpse of a truth that is frustratingly elusive. One way is to think of the probability waves associated with an atom's electrons as being like sound waves confined to an organ pipe. It is not possible to make just any note with the organ pipe. The sound can vibrate in only a limited number of different ways, each with a definite pitch, or frequency.

This turns out to be a general property of waves, not just sound waves. In a confined space they can exist only at particular, definite frequencies.

Now think of an electron in an atom. It behaves like a wave. And it is gripped tightly by the electrical force of the atomic nucleus. This may not be exactly the same as being trapped in a physical container. However, it confines the electron wave as surely as the wall of an organ pipe confines a sound wave. The electron wave can therefore exist at only certain frequencies.

The frequencies of the sound waves in an organ pipe and of the electron waves in an atom depend on the characteristics of the organ pipe—a small organ pipe, for instance, produces higher-pitched notes than a big organ pipe—and on the characteristics of the electrical force of the atomic nucleus. In general, though, there is lowest, or fundamental, frequency and a series of higher-frequency "overtones."

A higher-frequency wave has more peaks and troughs in a given space. It is choppier, more violent. In the case of an atom, such a wave corresponds to a faster-moving, more energetic electron. And a faster-moving, more energetic electron is able to defy the electrical attraction of the nucleus and orbit farther away.

The picture that emerges is of an electron that is permitted to orbit at only certain special distances from the nucleus. This is quite unlike our solar system where a planet such as Earth could, in principle, orbit at any distance whatsoever from the Sun.

This property highlights another important difference between the microscopic world of atoms and the everyday world. In the everyday world, all things are continuous—a planet can orbit the Sun anywhere it likes, people can be any weight they like—whereas things in the microscopic world are discontinuous—an electron can exist in only certain orbits around a nucleus, light and matter can come in only certain indivisible chunks. Physicists call the chunks quanta—which is why the physics of the microscopic world is known as quantum theory.

The innermost orbit of an electron in an atom is determined by the Heisenberg uncertainty principle—by its hornetlike resistance to being confined in a small space. But the Heisenberg uncertainty principle does not simply prevent small things like atoms from shrinking without limit—ultimately explaining the solidity of matter. It also prevents far bigger things from shrinking without limit. The far bigger things in question are stars.

UNCERTAINTY AND STARS

A star is a giant ball of gas held together by the gravitational pull of its own matter. That pull is constantly trying to shrink the star and, if unopposed, would very quickly collapse it down to the merest speck—a black hole. For the Sun this would take less than half an hour. Since the Sun is very definitely not shrinking down to a speck, there must be another force counteracting gravity. There is. It comes from the hot matter inside. The Sun—along with every other normal star—is in a delicate state of balance, with the inward force of gravity exactly matched by the outward force of its hot interior.

This balance, however, is temporary. The outward force can be maintained only while there is fuel to burn and keep the star hot. Sooner or later, the fuel will run out. For the Sun this will occur in about another 5 billion years. When this happens, gravity will be king. Unopposed, it will crush the star, shrinking it ever smaller.

But all is not lost. In the dense, hot environment inside a star, frequent and violent collisions between high-speed atoms strip them of their electrons, creating a plasma, a gas of atomic nuclei mixed in with a gas of electrons. It is the tiny electrons that unexpectedly come to the rescue of the fast-shrinking star. As the electrons in the star's matter are jammed ever closer together, they buzz about ever more violently because of the Heisenberg uncertainty principle. They batter anything trying to confine them, and this collective battering results in a tremendous outward force. Eventually, it is enough to slow and halt the shrinkage of the star.

A new balance is struck with the inward pull of gravity balanced not by the outward force of the star's hot matter but by the naked force of its electrons. Physicists call it degeneracy pressure. But it's just a fancy term for the resistance of electrons to being squeezed too close together. A star supported against gravity by electron pressure is known as a white dwarf. Little more than the size of Earth and occupying about a millionth of the star's former volume, a white dwarf is an enormously dense object. A sugarcube of its matter weighs as much as a car!

One day the Sun will become a white dwarf. Such stars have no means of replenishing their lost heat. They are nothing more than stellar embers, cooling inexorably and gradually fading from view. But the electron pressure that prevents white dwarfs from shrinking under their own gravity has its limits. The more massive a star, the stronger its self-gravity. If the star is massive enough, its gravity will be powerful enough to overcome even the stiff resistance of the star's electrons.

In fact, the star is sabotaged from both outside and inside. The stronger the gravity of a star, the more it squeezes the gas inside. And the more a gas is squeezed, the hotter it gets, as anyone who has used a bicycle pump knows. Since heat is nothing more than the microscopic jiggling of matter, the electrons inside the star fly about ever faster—so fast, in fact, that the effects of relativity become important.[2] The electrons get more massive rather than much faster, which means they are less effective at battering the walls of their prison.

The star suffers a double whammy—crushed by stronger gravity and simultaneously robbed of the ability to fight back. The two effects combine to ensure that the heaviest a white dwarf can be is a mere 40 per cent more massive than the Sun. If a star is heavier than this "Chandrasekhar limit", electron pressure is powerless to halt its headlong collapse and it just goes on shrinking.

But, once again, all is not lost. Eventually, the star shrinks so much that its electrons, despite their tremendous aversion to being confined in a small volume, are actually squeezed into the atomic nuclei. There they react with protons to form neutrons, so that the whole star becomes one giant mass of neutrons.

Recall that all particles of matter—not just electrons—resist being confined because of the Heisenberg uncertainty principle. Neutrons are thousands of times more massive than electrons. They therefore have to be squeezed into a volume thousands of times smaller to begin to put up significant resistance. In fact, they have to be squeezed together until they are virtually touching before they finally halt the shrinkage of the star.

A star supported against gravity by neutron degeneracy pressure is known as a neutron star. In effect, it is a huge atomic nucleus with all the empty space squeezed out of its matter. Since atoms are mostly empty space, with their nuclei 100,000 times smaller than their surrounding cloud of orbiting electrons, neutron stars are 100,000 times smaller than a normal star. This makes them only about 15 kilometres across, not much bigger than Mount Everest. So dense is a neutron star that a sugarcube of its matter weighs as much as the entire human race. (This, of course, is an illustration of just how much empty space there is in all of us. Squeeze it all out and humanity would fit in your hand.)

[2] See Chapter 7, "The Death of Space and Time."

Such stars are thought to form violently in supernova explosions. While the outer regions of a star are blown into space, the inner core shrinks to form a neutron star. Neutron stars, being tiny and cold, ought to be difficult to spot. However, they are born spinning very fast and produce lighthouse beams of radio waves that flash around the sky. Such pulsating neutron stars, or simply pulsars, semaphore their existence to astronomers.

UNCERTAINTY AND THE VACUUM

White dwarfs and neutron stars apart, perhaps the most remarkable consequence of the Heisenberg uncertainty principle is the modern picture of empty space. It simply cannot be empty!

The Heisenberg uncertainty principle can be reformulated to say that it is impossible to simultaneously measure the energy of a particle and the interval of time for which it has been in existence. Consequently, if we consider what happens in a region of empty space in a very tiny interval of time, there will be a large uncertainty in the energy content of that region. In other words, energy can appear out of nothing!

Now, mass is a form of energy.[3] This means that mass too can appear out of nothing. The proviso is that it can appear only for a mere split second before disappearing again. The laws of nature, which usually prevent things from appearing out of nothing, appear to turn a blind eye to events that happen too quickly. It's rather like a teenager's dad not noticing his son has borrowed the car for the night as long as it gets put back in the garage before daybreak.

In practice, mass is conjured out of empty space in the form of microscopic particles of matter. The quantum vacuum is actually a seething morass of microscopic particles such as electrons popping

[3] See Chapter 8, "$E = mc^2$ and the Weight of Sunlight."

into existence and then vanishing again.[4] And this is no mere theory. It actually has observable consequences. The roiling sea of the quantum vacuum actually buffets the outer electrons in atoms, very slightly changing the energy of the light they give out.[5]

The fact that the laws of nature permit something to come out of nothing has not escaped cosmologists, people who think about the origin of the Universe. Could it be, they wonder, that the entire Universe is nothing more than a quantum fluctuation of the vacuum? It's an extraordinary thought.

[4] Actually, every particle created is created alongside its antiparticle, a particle with opposite properties. So a negatively charged electron is always created with a positively charged positron.

[5] This effect is called the Lamb shift.