"Brilliant... the most gripping 400 pages I've read in years."—*The Times* (London)

I AM A STRANGE LOOP

DOUGLAS HOFSTADTER

Praise for *I Am A Strange Loop*

"[F]ascinating . . . original and thought-provoking . . . [T]here are many pleasures in *I Am a Strange Loop.*"

— *Wall Street Journal*

"*I Am a Strange Loop* scales some lofty conceptual heights, but it remains very personal, and it's deeply colored by the facts of Hofstadter's later life. In 1993 Hofstadter's wife Carol died suddenly of a brain tumor at only 42, leaving him with two young children to care for . . . *I Am a Strange Loop* is a work of rigorous thinking."

— *Time*

"Almost thirty years after the publication of his well-loved *Gödel, Escher, Bach*, Hofstadter revisits some of the same themes. The purpose of the new book is to make inroads into the nexus of self, self-awareness and consciousness by examining self-referential structures in areas as diverse as art and mathematics. Hofstadter is the man for the job. His treatment of issues is approachable and personal, you might even say subjective. His discussion is never over technical and his prose never over-bearing. He stays close to the surface of real life at all times, even as he discusses matters of the highest level of abstraction, and his book is full of fresh and rich real-life examples that give texture and authenticity to the discussion."

— *Times Literary Supplement*, London

"[P]leasant and intriguing . . . Hofstadter is a supremely skillful master of an educational alchemy that can, at the turn of the page, transform the most abstract and complex of thoughts into a digestible idea that is both fun and interesting . . . Hofstadter's good humor and easygoing style make it a real pleasure to read from start to finish."

— *Times Higher Education Supplement,* London

"*I Am a Strange Loop* contains many profound and unique insights on the question of who we are. In addition, it is a delightful read."

— *Physics Today*

"*I Am a Strange Loop* is vintage Hofstadter: earnest, deep, overflowing with ideas, building its argument into the experience of reading it — for if our souls can incorporate those of others, then *I Am a Strange Loop* can transmit Hofstadter's into ours. And indeed, it is impossible to come away from this book without having introduced elements of his point of view into our own. It may not make us kinder or more compassionate, but we will never look at the world, inside or out, in the same way again."

— *Los Angeles Times Book Review*

"Nearly thirty years after his best-selling book *Gödel, Escher, Bach*, cognitive scientist and polymath Douglas Hofstadter has returned to his extraordinary theory of self."

— *New Scientist*

"*I Am a Strange Loop* is thoughtful, amusing and infectiously enthusiastic."

— *Bloomberg News*

"[P]rovocative and heroically humane . . . it's impossible not to experience this book as a tender, remarkably personal and poignant effort to understand the death of his wife from cancer in 1993 — and to grasp how consciousness mediates our otherwise ineffable relationships. In the end, Hofstadter's view is deeply philosophical rather than scientific. It's hopeful and romantic as well, as his model allows one consciousness to create and maintain within itself true representations of the essence of another."

— *Publishers Weekly* Starred Review

"[Hofstadter's] new book, as brilliant and provocative as earlier ones, is a colorful mix of speculations with passages of autobiography."

— Martin Gardner in *Notices of the American Mathematical Society*

"Why am I inside this body and not in a different one? This is among the most irresistible and fascinating questions humanity has ever asked, according to Douglas Hofstadter. His latest book *I Am a Strange Loop* asks many more

challenging questions: Are our thoughts made of molecules? Could a machine be confused? Could a machine *know* it was confused? — until it ties you in loops. If you enjoy such brain-bending questions and are willing to struggle with some deep mathematical ideas along the way, then you'll certainly enjoy this book . . . (I)f this book works its magic on you, you will no longer want to ask 'why am I inside this body and not a different one?' because you'll know what it means to be just a strange loop."

— *BBC Focus*

"Hofstadter introduces new ideas about the self-referential structure of consciousness and offers a multifaceted examination of what an 'I' is. He conveys abstract, complicated ideas in a relaxed, conversational manner and uses many first-person stories and personal examples as well as two Platonic dialogs. Though Hofstadter admits he writes for the general educated public, he also hopes to reach professional philosophers interested in the epistemological implications of selfhood."

— *Library Journal*

"Hofstadter explains the dynamics of [the] reflective self in refreshingly lucid language, enlivened with personal anecdotes that translate arcane formulas into the wagging tail on a golden retriever or the smile on Hopalong Cassidy. Nonspecialists are thus able to assess the divide between human and animal minds, and even to plumb the mental links binding the living to the dead . . . [E]ven skeptics will appreciate the way he forces us to think deeper thoughts about thought."

— *Booklist* Starred Review

I AM A STRANGE LOOP

DOUGLAS HOFSTADTER

A Member of the Perseus Books Group
New York

To my sister Laura,
who can understand,

and to our sister Molly,
who cannot.

A note from the Publisher

Doug Hofstadter, who over the years has been a friend to Basic Books in so many ways, has kindly lent us this page to remember a late colleague. We gratefully dedicate this book

To Liz Maguire
1958–2006
who lives on in all of us.

WORDS OF THANKS

SINCE my teen-age years, I have been fascinated by what the mind is and does, and have pondered such riddles for many decades. Some of my conclusions have come from personal experiences and private musings, but of course I have been profoundly marked by the ideas of many other people, stretching way back to elementary school, if not earlier.

Among the well-known authors who have most influenced my thinking on the interwoven topics of minds, brains, patterns, symbols, self-reference, and consciousness are, in some vague semblance of chronological order: Ernest Nagel, James R. Newman, Kurt Gödel, Martin Gardner, Raymond Smullyan, John Pfeiffer, Wilder Penfield, Patrick Suppes, David Hamburg, Albert Hastorf, M. C. Escher, Howard DeLong, Richard C. Jeffrey, Ray Hyman, Karen Horney, Mikhail Bongard, Alan Turing, Gregory Chaitin, Stanislaw Ulam, Leslie A. Hart, Roger Sperry, Jacques Monod, Raj Reddy, Victor Lesser, Marvin Minsky, Margaret Boden, Terry Winograd, Donald Norman, Eliot Hearst, Daniel Dennett, Stanislaw Lem, Richard Dawkins, Allen Wheelis, John Holland, Robert Axelrod, Gilles Fauconnier, Paolo Bozzi, Giuseppe Longo, Valentino Braitenberg, Derek Parfit, Daniel Kahneman, Anne Treisman, Mark Turner, and Jean Aitchison. Books and articles by many of these authors are cited in the bibliography. Over the years, I have come to know quite a few of these individuals, and I count the friendships thus formed among the great joys of my life.

On a more local level, I have been influenced over a lifetime by thousands of intense conversations, phone calls, letters, and emails with family members, friends, students, and colleagues. Once again, listed in some rough semblance of chronological order, these people would include: Nancy Hofstadter, Robert Hofstadter, Laura Hofstadter, Peter Jones, Robert Boeninger, Charles Brenner, Larry Tesler, Michael Goldhaber, David Policansky, Peter S. Smith, Inga Karliner, Francisco Claro, Peter Rimbey, Paul Csonka, P. David Jennings, David Justman, J. Scott Buresh, Sydney Arkowitz, Robert Wolf, Philip Taylor, Scott Kim, Pentti Kanerva, William Gosper, Donald Byrd, J. Michael Dunn, Daniel Friedman, Marsha Meredith, Gray Clossman, Ann Trail, Susan Wunder, David Moser, Carol Brush Hofstadter, Leonard Shar, Paul

Smolensky, David Leake, Peter Suber, Greg Huber, Bernard Greenberg, Marek Lugowski, Joe Becker, Melanie Mitchell, Robert French, David Rogers, Benedetto Scimemi, Daniel Defays, William Cavnar, Michael Gasser, Robert Goldstone, David Chalmers, Gary McGraw, John Rehling, James Marshall, Wang Pei, Achille Varzi, Oliviero Stock, Harry Foundalis, Hamid Ekbia, Marilyn Stone, Kellie Gutman, James Muller, Alexandre Linhares, Christoph Weidemann, Nathaniel Shar, Jeremy Shar, Alberto Parmeggiani, Alex Passi, Francesco Bianchini, Francisco Lara-Dammer, Damien Sullivan, Abhijit Mahabal, Caroline Strobbe, Emmanuel Sander, Glen Worthey — and of course Carol's and my two children, Danny and Monica Hofstadter.

I feel deep gratitude to Indiana University for having so generously supported me personally and my group of researchers (the Fluid Analogies Research Group, affectionately known as "FARG") for such a long time. Some of the key people at IU who have kept the FARGonauts afloat over the past twenty years are Helga Keller, Mortimer Lowengrub, Thomas Ehrlich, Kenneth Gros Louis, Kumble Subbaswamy, Robert Goldstone, Richard Shiffrin, J. Michael Dunn, and Andrew Hanson. All of them have been intellectual companions and staunch supporters, some for decades, and I am lucky to be able to count them among my colleagues.

I have long felt part of the family at Basic Books, and am grateful for the support of many people there for nearly thirty years. In the past few years I have worked closely with William Frucht, and I truly appreciate his open-mindedness, his excellent advice, and his unflagging enthusiasm.

A few people have helped me enormously on this book. Ken Williford and Uriah Kriegel launched it; Kellie Gutman, Scott Buresh, Bill Frucht, David Moser, and Laura Hofstadter all read chunks of it and gave superb critical advice; and Helga Keller chased permissions far and wide. I thank them all for going "way ABCD" — way above and beyond the call of duty.

The many friends mentioned above, and some others not mentioned, form a "cloud" in which I float; sometimes I think of them as the "metropolitan area" of which I, construed narrowly, am just the zone inside the official city limits. Everyone has friends, and in that sense I am no different from anyone else, but this cloud is *my* cloud, and it somehow defines me, and I am proud of it and proud of them all. And so I say to this cloud of friends, with all my heart, "Thank you so very much, one and all!"

PREFACE

An Author and His Book

Facing the Physicality of Consciousness

FROM an early age onwards, I pondered what my mind was and, by analogy, what all minds are. I remember trying to understand how I came up with the puns I concocted, the mathematical ideas I invented, the speech errors I committed, the curious analogies I dreamt up, and so forth. I wondered what it would be like to be a girl, to be a native speaker of another language, to be Einstein, to be a dog, to be an eagle, even to be a mosquito. By and large, it was a joyous existence.

When I was twelve, a deep shadow fell over our family. My parents, as well as my seven-year-old sister Laura and I, faced the harsh reality that the youngest child in our family, Molly, then only three years old, had something terribly wrong with her. No one knew what it was, but Molly wasn't able to understand language or to speak (nor is she to this day, and we never did find out why). She moved through the world with ease, even with charm and grace, but she used no words at all. It was so sad.

For years, our parents explored every avenue imaginable, including the possibility of some kind of brain surgery, and as their quest for a cure or at least some kind of explanation grew ever more desperate, my own anguished thinking about Molly's plight and the frightening idea of people opening up my tiny sister's head and peering in at the mysterious stuff that filled it (an avenue never explored, in the end) gave me the impetus to read a couple of lay-level books about the human brain. Doing so had a huge impact on my life, since it forced me to consider, for the first time, the physical basis of consciousness and of being — or of having — an "I", which I found disorienting, dizzying, and profoundly eerie.

Right around that time, toward the end of my high-school years, I encountered the mysterious metamathematical revelations of the great

Austrian logician Kurt Gödel and I also learned how to program, using Stanford University's only computer, a Burroughs 220, which was located in the deliciously obscure basement of decrepit old Encina Hall. I rapidly became addicted to this "Giant Electronic Brain", whose orange lights flickered in strange magical patterns revealing its "thoughts", and which, at my behest, discovered beautiful abstract mathematical structures and composed whimsical nonsensical passages in various foreign languages that I was studying. I simultaneously grew obsessed with symbolic logic, whose arcane symbols danced in strange magical patterns reflecting truths, falsities, hypotheticals, possibilities, and counterfactualities, and which, I was sure, afforded profound glimpses into the hidden wellsprings of human thought. As a result of these relentlessly churning thoughts about symbols and meanings, patterns and ideas, machines and mentality, neural impulses and mortal souls, all hell broke loose in my adolescent mind/brain.

The Mirage

One day when I was around sixteen or seventeen, musing intensely on these swirling clouds of ideas that gripped me emotionally no less than intellectually, it dawned on me — and it has ever since seemed to me — that what we call "consciousness" was a kind of mirage. It had to be a very peculiar kind of mirage, to be sure, since it was a mirage that perceived itself, and of course it didn't *believe* that it was perceiving a mirage, but no matter — it still *was* a mirage. It was almost as if this slippery phenomenon called "consciousness" lifted itself up by its own bootstraps, almost as if it made itself out of nothing, and then disintegrated back into nothing whenever one looked at it more closely.

So caught up was I in trying to understand what being alive, being human, and being conscious are all about that I felt driven to try to capture my elusive thoughts on paper lest they flit away forever, and so I sat down and wrote a dialogue between two hypothetical contemporary philosophers whom I flippantly named "Plato" and "Socrates" (I knew almost nothing about the real Plato and Socrates). This may have been the first serious piece of writing I ever did; in any case, I was proud of it, and never threw it away. Although I now see my dialogue between these two pseudo-Greek philosophers as pretty immature and awkward, not to mention extremely sketchy, I decided nonetheless to include it herein as my Prologue, because it hints at many of the ideas to come, and I think it sets a pleasing and provocative tone for the rest of the book.

A Shout into a Chasm

When, some ten years or so later, I started working on my first book, whose title I imagined would be "Gödel's Theorem and the Human Brain", my overarching goal was to relate the concept of a human self and the mystery of consciousness to Gödel's stunning discovery of a majestic wraparound self-referential structure (a "strange loop", as I later came to call it) in the very midst of a formidable bastion from which self-reference had been strictly banished by its audacious architects. I found the parallel between Gödel's miraculous manufacture of self-reference out of a substrate of meaningless symbols and the miraculous appearance of selves and souls in substrates consisting of inanimate matter so compelling that I was convinced that here lay the secret of our sense of "I", and thus my book *Gödel, Escher, Bach* came about (and acquired a catchier title).

That book, which appeared in 1979, couldn't have enjoyed a greater success, and indeed yours truly owes much of the pathway of his life since then to its success. And yet, despite the book's popularity, it always troubled me that the fundamental message of *GEB* (as I always call it, and as it is generally called) seemed to go largely unnoticed. People liked the book for all sorts of reasons, but seldom if ever for its most central *raison d'être*! Years went by, and I came out with other books that alluded to and added to that core message, but still there didn't seem to be much understanding out there of what I had really been trying to say in *GEB.*

In 1999, *GEB* celebrated its twentieth anniversary, and the folks at Basic Books suggested that I write a preface for a special new edition. I liked the idea, so I took them up on it. In my preface, I told all sorts of tales about the book and its vicissitudes, and among other things I described my frustration with its reception, ending with the following plaint: "It sometimes feels as if I had shouted a deeply cherished message out into an empty chasm and nobody heard me."

Well, one day in the spring of 2003, I received a very kind email message from two young philosophers named Ken Williford and Uriah Kriegel, inviting me to contribute a chapter to an anthology they were putting together on what they called "the self-referentialist theory (or theories)" of consciousness. They urged me to participate, and they even quoted back to me that very lamentation of mine from my preface, and they suggested that this opportunity would afford me a real chance to change things. I was genuinely gratified by their sincere interest in my core message and moved by their personal warmth, and I saw that indeed, contributing to their volume would be a grand occasion for me to try once again to articulate my ideas about self and

consciousness for exactly the right audience of specialists — philosophers of mind. And so it wasn't too hard for me to decide to accept their invitation.

From the Majestic Dolomites to Gentle Bloomington

I started writing my chapter in a quiet and simple hotel room in the beautiful Alpine village of Anterselva di Mezzo, located in the Italian Dolomites, only a few stone's throws from the Austrian border. Inspired by the loveliness of the setting, I quickly dashed off ten or fifteen pages, thinking I might already have reached the halfway point. Then I returned home to Bloomington, Indiana, where I kept on plugging away.

It took me a good deal longer than I had expected to finish it (some of my readers will recognize this as a quintessential example of Hofstadter's Law, which states, "It always takes longer than you think it will take, even when you take into account Hofstadter's Law"), and worse, the chapter wound up being four times longer than the specified limit — a disaster! But when they finally received it, Ken and Uriah were very pleased with what I had written and were most tolerant of my indiscretions; indeed, so keen were they to have a contribution from me in their book that they said they could accept an extra-long chapter, and Ken in particular helped me cut it down to half its length, which was a real labor of love on his part.

In the meantime, I was starting to realize that what I had on my hands could be more than a book chapter — it could become a book unto itself. And so what had begun as a single project fissioned into two. I gave my chapter the title "What is it like to be a strange loop?", alluding to a famous article on the mystery of consciousness called "What is it like to be a bat?" by the philosopher of mind Thomas Nagel, while the book-to-be was given the shorter, sweeter title "I Am a Strange Loop".

In Ken Williford and Uriah Kriegel's anthology, *Self-Representational Approaches to Consciousness,* which appeared in the spring of 2006, my essay was placed at the very end, in a two-chapter section entitled "Beyond Philosophy" (why it qualified as lying "beyond philosophy" is beyond me, but I rather like the idea nonetheless). I don't know if, in that distinguished but rather specialized setting, this set of ideas will have much impact on anyone, but I certainly hope that in this book, its more fully worked-out and more visible incarnation, it will be able to reach all sorts of people, both inside and outside of philosophy, both young and old, both specialists and novices, and will give them new imagery about selves and souls (not to mention loops!). In any case, I owe a great deal to Ken and Uriah for having provided the initial spark that

gave rise to this book, as well as for giving me much encouragement along the way.

And so, after just about forty-five years (good grief!), I've come full circle, writing once again about souls, selves, and consciousness, banging up against the same mysteriousness and eerieness that I first experienced when I was a teenager horrified and yet riveted by the awful and awesome physicality of that which makes us be what we are.

An Author and His Audience

Despite its title, this book is not about me, but about the concept of "I". It's thus about you, reader, every bit as much as it is about me. I could just as well have called it "You Are a Strange Loop". But the truth of the matter is that, in order to suggest the book's topic and goal more clearly, I should probably have called it "'I' Is a Strange Loop" — but can you imagine a clunkier title? Might as well call it "I Am a Lead Balloon".

In any case, this book is about the venerable topic of what an "I" is. And what is its audience? Well, as always, I write in order to reach a general educated public. I almost never write for specialists, and in a way that's because I'm not really a specialist myself. Oh, I take it back; that's unfair. After all, at this point in my life, I have spent nearly thirty years working with my graduate students on computational models of analogy-making and creativity, observing and cataloguing cognitive errors of all sorts, collecting examples of categorization and analogy, studying the centrality of analogies in physics and math, musing on the mechanisms of humor, pondering how concepts are created and memories are retrieved, exploring all sorts of aspects of words, idioms, languages, and translation, and so on — and over these three decades I have taught seminars on many aspects of thinking and how we perceive the world.

So yes, in the end, I am a kind of specialist — I specialize in thinking about thinking. Indeed, as I stated earlier, this topic has fueled my fire ever since I was a teenager. And one of my firmest conclusions is that we always think by seeking and drawing parallels to things we know from our past, and that we therefore communicate best when we exploit examples, analogies, and metaphors galore, when we avoid abstract generalities, when we use very down-to-earth, concrete, and simple language, and when we talk directly about our own experiences.

The Horsies-and-Doggies Religion

Over the years, I have fallen into a style of self-expression that I call the "horsies-and-doggies" style, a phrase inspired by a charming episode in the famous cartoon "Peanuts", which I've reproduced on the following page.

I often feel just the way that Charlie Brown feels in that last frame — like someone whose ideas are anything but "in the clouds", someone who is so down-to-earth as to be embarrassed by it. I realize that some of my readers have gotten an impression of me as someone with a mind that enormously savors and indefatigably pursues the highest of abstractions, but that is a very mistaken image. I'm just the opposite, and I hope that reading this book will make that evident.

I don't have the foggiest idea why I wrongly remembered the poignant phrase that Charlie Brown utters here, but in any case the slight variant "horsies and doggies" long ago became a fixture in my own speech, and so, for better or for worse, that's the standard phrase I always use to describe my teaching style, my speaking style, and my writing style.

In part because of the success of *Gödel, Escher, Bach,* I have had the good fortune of being given a great deal of freedom by the two universities on whose faculties I have served — Indiana University (for roughly twenty-five years) and the University of Michigan (for four years, in the 1980's). Their wonderful generosity has given me the luxury of being able to explore my variegated interests without being under the infamous publish-or-perish pressures, or perhaps even worse, the relentless pressures of grant-chasing.

I have not followed the standard academic route, which involves publishing paper after paper in professional journals. To be sure, I have published some "real" papers, but mostly I have concentrated on expressing myself through books, and these books have always been written with an eye to maximal clarity.

Clarity, simplicity, and concreteness have coalesced into a kind of religion for me — a set of never-forgotten guiding principles. Fortunately, a large number of thoughtful people appreciate analogies, metaphors, and examples, as well as a relative lack of jargon, and last but not least, accounts from a first-person stance. In any case, it is for people who appreciate that way of writing that this book, like all my others, has been written. I believe that this group includes not only outsiders and amateurs, but also many professional philosophers of mind.

If I tell many first-person stories in this book, it is not because I am obsessed with my own life or delude myself about its importance, but simply because it is the life I know best, and it provides all sorts of examples that I suspect are typical of most people's lives. I believe most people understand abstract ideas most clearly if they hear them through stories, and so I try to convey difficult and abstract ideas through the medium of my own life. I wish that more thinkers wrote in a first-person fashion.

Although I hope to reach philosophers with this book's ideas, I don't think that I write very much like a philosopher. It seems to me that many philosophers believe that, like mathematicians, they can actually *prove* the points they believe in, and to that end, they often try to use highly rigorous and technical language, and sometimes they attempt to anticipate and to counter all possible counter-arguments. I admire such self-confidence, but I am a bit less optimistic and a bit more fatalistic. I don't think one can truly prove anything in philosophy; I think one can merely try to convince, and probably one will wind up convincing only those people who started out fairly close to the position one is advocating. As a result of this mild brand of fatalism, my strategy for conveying my points is based more on metaphor and analogy than on attempts at rigor. Indeed, this book is a gigantic salad bowl full of metaphors and analogies. Some will savor my metaphor salad, while others

will find it too… well, too metaphorical. But I particularly hope that *you,* dear eater, will find it seasoned to your taste.

A Few Last Random Observations

I take analogies very seriously, so much so that I went to a great deal of trouble to index a large number of the analogies in my "salad". There are thus two main headings in the index for my lists of examples. One is "analogies, serious examples of"; the other is "throwaway analogies, random examples of". I made this droll distinction because whereas many of my analogies play key roles in conveying ideas, some are there just to add spice. There's another point to be made, though: in the final analysis, virtually every thought in this book (or in any book) is an analogy, as it involves recognizing something as being a variety of something else. Thus every time I write "similarly" or "by contrast", there is an implicit analogy, and every time I pick a word or phrase (*e.g.,* "salad", "storehouse", "bottom line"), I am making an analogy to something in my life's storehouse of experiences. The bottom line is, *every* thought herein could be listed under "analogies". However, I refrained from making my index that detailed.

I initially thought this book was just going to be a distilled retelling of the central message of *GEB,* employing little or no formal notation and not indulging in Pushkinian digressions into such variegated topics as Zen Buddhism, molecular biology, recursion, artificial intelligence, and so forth. In other words, I thought I had already fully stated in *GEB* and my other books what I intended to (re)state here, but to my surprise, as I started to write, I saw new ideas sprouting everywhere under foot. That was a relief, and made me feel that my new book was more than just a rehash of an earlier book (or books).

Among the keys to *GEB*'s success was its alternation between chapters and dialogues, but I didn't intend, thirty years later, to copycat myself with another such alternation. I was in a different frame of mind, and I wanted this book to reflect that. But as I was approaching the end, I wanted to try to compare my ideas with well-known ideas in the philosophy of mind, and so I started saying things like, "Skeptics might reply as follows…" After I had written such phrases a few times, I realized I had inadvertently fallen into writing a dialogue between myself and a hypothetical skeptical reader, so I invented a pair of oddly-named characters and let them have at each other for what turned out to be one of the longest chapters in the book. It's not intended to be uproariously funny, although I hope my readers will occasionally smile here and there as they read

it. In any case, fans of the dialogue form, take heart — there are two dialogues in this book.

I am a lifelong lover of form–content interplay, and this book is no exception. As with several of my previous books, I have had the chance to typeset it down to the finest level of detail, and my quest for visual elegance on each page has had countless repercussions on how I phrase my ideas. To some this may sound like the tail wagging the dog, but I think that attention to form improves anyone's writing. I hope that reading this book not only is stimulating intellectually but also is a pleasant visual experience.

A Useful Youthfulness

GEB was written by someone pretty young (I was twenty-seven when I started working on it and twenty-eight when I completed the first draft — all written out in pen on lined paper), and although at that tender age I had already experienced my fair or unfair share of suffering, sadness, and moral soul-searching, one doesn't find too much allusion to those aspects of life in the book. In this book, though, written by someone who has known considerably more suffering, sadness, and soul-searching, those hard aspects of life are much more frequently touched on. I think that's one of the things about growing older — one's writing becomes more inward, more reflective, perhaps wiser, or perhaps just sadder.

I have long been struck by the poetic title of André Malraux's famous novel *La Condition humaine.* I guess each of us has a personal sense of what this evocative phrase means, and I would characterize *I Am a Strange Loop* as being my own best shot at describing what "the human condition" is.

One of my favorite blurbs for *GEB* came from the physicist and writer Jeremy Bernstein, and in part it said, "It has a youthful vitality and a wonderful brilliance…" True music to my ears! But unfortunately this flattering phrase got garbled at some point, and as a result there are now thousands of copies of *GEB* floating around on whose back cover Bernstein proclaims, "It has a useful vitality…" What a letdown, compared with a "youthful" vitality! And yet perhaps this new book, in its older, more sober style, will someday be described by someone somewhere as having a "useful" vitality. I guess worse things could be said about a book.

And so now I will stop talking about my book, and will let my book talk for itself. In it I hope you will discover messages imbued with interest and novelty, and even with a useful, if no longer youthful, vitality. I hope that reading this book will make you reflect in fresh ways on what being human is all about — in

fact, on what just-plain *being* is all about. And I hope that when you put the book down, you will perhaps be able to imagine that you, too, are a strange loop. Now that would please me no end.

— Bloomington, Indiana

December, MMVI.

PROLOGUE

An Affable Locking of Horns

[As I stated in the Preface, I wrote this dialogue when I was a teenager, and it was my first, youthful attempt at grappling with these difficult ideas.]

Dramatis personæ:

Plato: a seeker of truth who suspects consciousness is an illusion
Socrates: a seeker of truth who believes in consciousness' reality

PLATO: But what then do you mean by "life", Socrates? To my mind, a living creature is a body which, after birth, grows, eats, learns how to react to various stimuli, and which is ultimately capable of reproduction.
SOCRATES: I find it interesting, Plato, that you say a living creature *is* a body, rather than *has* a body. For surely, many people today would say that there are at least some living creatures that have souls independent of their bodies.
PLATO: Yes, and with those I would agree. I should have said that living creatures *have* bodies.
SOCRATES: Then you would agree that fleas and mice have souls, however insignificant.
PLATO: My definition does require that, yes.
SOCRATES: And do trees have souls, and blades of grass?
PLATO: You have used words to put me in this situation, Socrates. I will revise what I said — only animals have souls.
SOCRATES: But no, I have not only used words, for there is no distinction to be found between plants and animals, if you examine small enough creatures.

PLATO: You mean there are some creatures sharing the properties of plant and animal? Yes, I guess I can imagine such a thing, myself. Now I suppose you will force me into saying that only humans have souls. SOCRATES: No, on the contrary, I will ask you, what animals do you usually consider to have souls?

PLATO: Why, all higher animals — those which are able to think. SOCRATES: Then, at least higher animals are alive. Now can you truly consider a stalk of grass to be a living creature like yourself?

PLATO: Let me put it this way, Socrates: I can only imagine true life with a soul, and so I must discard grass as true life, though I could say it has the symptoms of life.

SOCRATES: I see. So you would classify soulless creatures as only *appearing* alive, and creatures with souls as *true* life. Then am I right if I say that your question "What is true life?" depends on the understanding of the soul?

PLATO: Yes, that is right.

SOCRATES: And you have said that you consider the soul as the ability to think?

PLATO: Yes.

SOCRATES: Then you are really seeking the answer to "What is thinking?"

PLATO: I have followed each step of your argument, Socrates, but this conclusion makes me uneasy.

SOCRATES: It has not been *my* argument, Plato. You have provided all the facts, and I have only drawn logical conclusions from them. It is curious, how one often mistrusts one's own opinions if they are stated by someone else.

PLATO: You are right, Socrates. And surely it is no simple task to explain thinking. It seems to me that the purest thought is the *knowing* of something; for clearly, to know something is more than just to write it down or to assert it. These can be done if one knows something; and one can learn to know something from hearing it asserted or from seeing it written. Yet knowing is more than this — it is conviction — but I am only using a synonym. I find it beyond me to understand what knowing is, Socrates.

SOCRATES: That is an interesting thought, Plato. Do you say that knowing is not so familiar as we think it is?

PLATO: Yes. Because we humans have knowledge, or convictions, we are humans, yet when we try to analyze knowing itself, it recedes, and evades us.

SOCRATES: Then had one not better be suspicious of what we call "knowing", or "conviction", and not take it so much for granted?

PLATO: Precisely. We must be cautious in saying "I know", and we must ponder what it truly means to say "I know" when our minds would have us say it.

SOCRATES: True. If I asked you, “Are you alive?”, you would doubtless reply, “Yes, I am alive.” And if I asked you, “How do you know that you are alive?”, you would say “I *feel* it, I *know* I am alive — indeed, is not knowing and feeling one is alive *being* alive?” Is that not right?
PLATO: Yes, I would certainly say something to that effect.
SOCRATES: Now let us suppose that a machine had been constructed which was capable of constructing English sentences and answering questions. And suppose I asked this English machine, “Are you alive?” and suppose it gave me precisely the same answers as you did. What would you say as to the validity of its answers?
PLATO: I would first of all object that no machine can know what words are, or mean. A machine merely deals with words in an abstract mechanical fashion, much as canning machines put fruit in cans.
SOCRATES: I do not accept your objections for two reasons. Surely you do not contend that the basic unit of human thought is the word? For it is well known that humans have nerve cells, the laws of whose operation are arithmetical. Secondly, you cautioned earlier that we must be wary of the verb “to know”, yet here you use it quite nonchalantly. What makes you say that no machine could ever “know” what words are, or mean?
PLATO: Socrates, do you argue that machines can know facts, as we humans do?
SOCRATES: You declared just now that you yourself cannot even explain what knowing is. How did you learn the verb “to know” as a child?
PLATO: Evidently, I assimilated it from hearing it used around me.
SOCRATES: Then it was by automatic action that you gained control of it.
PLATO: No… Well, perhaps I see what you mean. I grew accustomed to hearing it in certain contexts, and thus came to be able to use it myself in those contexts, in a more or less automatic fashion.
SOCRATES: Much as you use language now — without having to reflect on each word?
PLATO: Yes, exactly.
SOCRATES: Thus now, if you say, “I know I am alive”, that sentence is merely a reflex coming from your brain, and is not a product of conscious thought.
PLATO: No, no! You or I have used faulty logic. Not all thoughts I utter are simply products of reflex actions. Some thoughts I think about *consciously* before uttering.
SOCRATES: In what sense do you think consciously about them?
PLATO: I don’t know. I suppose that I try to find the correct words to describe them.
SOCRATES: What guides you to the correct words?

PLATO: Why, I search logically for synonyms, similar words, and so on, with which I am familiar.
SOCRATES: In other words, *habit* guides your thought.
PLATO: Yes, my thought is guided by the habit of connecting words with one another systematically.
SOCRATES: Then once again, these conscious thoughts are produced by reflex action.
PLATO: I do not see how I can know I am conscious, how I can feel alive, if this is true, yet I have followed your argument.
SOCRATES: But this argument itself shows that your reaction is merely habit, or reflex action, and that no conscious thought is leading you to say you know you are alive. If you stop to consider it, do you really understand what you mean by saying such a sentence? Or does it just come into your mind without your thinking consciously of it?
PLATO: Indeed, I am so confused I scarcely know.
SOCRATES: It becomes interesting to see how one's mind fails when working in new channels. Do you see how little you understand of that sentence "I am alive"?
PLATO: Yes, it is truly a sentence which, I must admit, is not so obvious to understand.
SOCRATES: I think it is in the same way as you fashioned that sentence that many of our actions come about — we think they arise through conscious thought, yet, on careful analysis, each bit of that thought is seen to be automatic and without consciousness.
PLATO: Then feeling one is alive is merely an illusion propagated by a reflex that urges one to utter, without understanding, such a sentence, and a truly living creature is reduced to a collection of complex reflexes. Then you have told me, Socrates, what you think life is.

CHAPTER 1

On Souls and Their Sizes

Soul-Shards

ONE gloomy day in early 1991, a couple of months after my father died, I was standing in the kitchen of my parents' house, and my mother, looking at a sweet and touching photograph of my father taken perhaps fifteen years earlier, said to me, with a note of despair, "What meaning does that photograph have? None at all. It's just a flat piece of paper with dark spots on it here and there. It's useless." The bleakness of my mother's grief-drenched remark set my head spinning because I knew instinctively that I disagreed with her, but I did not quite know how to express to her the way I felt the photograph should be considered.

After a few minutes of emotional pondering — soul-searching, quite literally — I hit upon an analogy that I felt could convey to my mother my point of view, and which I hoped might lend her at least a tiny degree of consolation. What I said to her was along the following lines.

"In the living room we have a book of the Chopin études for piano. All of its pages are just pieces of paper with dark marks on them, just as two-dimensional and flat and foldable as the photograph of Dad — and yet, think of the powerful effect that they have had on people all over the world for 150 years now. Thanks to those black marks on those flat sheets of paper, untold thousands of people have collectively spent millions of hours moving their fingers over the keyboards of pianos in complicated patterns, producing sounds that give them indescribable pleasure and a sense of great meaning. Those pianists in turn have conveyed to many millions of listeners, including you and me, the profound emotions that churned in Frédéric Chopin's heart, thus affording all of us some partial access to Chopin's interiority — to the experience of living in the head, or rather the soul, of Frédéric Chopin. The

marks on those sheets of paper are no less than soul-shards — scattered remnants of the shattered soul of Frédéric Chopin. Each of those strange geometries of notes has a unique power to bring back to life, inside our brains, some tiny fragment of the internal experiences of another human being — his sufferings, his joys, his deepest passions and tensions — and we thereby know, at least in part, what it was like to be that human being, and many people feel intense love for him. In just as potent a fashion, looking at that photograph of Dad brings back, to us who knew him intimately, the clearest memory of his smile and his gentleness, activates inside our living brains some of the most central representations of him that survive in us, makes little fragments of his soul dance again, but in the medium of brains other than his own. Like the score to a Chopin étude, that photograph is a soul-shard of someone departed, and it is something we should cherish as long as we live."

Although the above is a bit more flowery than what I said to my mother, it gives the essence of my message. I don't know what effect it had on her feelings about the picture, but that photo is still there, on a counter in her kitchen, and every time I look at it, I remember that exchange.

What Is It Like to Be a Tomato?

I slice up and devour tomatoes without the slightest sense of guilt. I do not go to bed uneasily after having consumed a fresh tomato. It does not occur to me to ask myself *which* tomato I ate, or whether by eating it I have snuffed an inner light, nor do I believe it is meaningful to try to imagine how the tomato felt as it was sitting on my plate being sliced apart. To me, a tomato is a desireless, soulless, non-conscious entity, and I have no qualms about doing with its "body" as I like. Indeed, a tomato is nothing but its body. There is no "mind–body problem" for tomatoes. (I hope, dear reader, that we agree on this much!)

I also swat mosquitoes without a qualm, though I try to avoid stepping on ants, and when there is an insect other than a mosquito in the house, I usually try to capture it and carry it outside, where I let it go unharmed. I eat chicken and fish sometimes [Note: This is no longer the case — see the Post Scriptum to this chapter], but many years ago I stopped eating the flesh of mammals. No beef, no ham, no bacon, no spam, no pork, no lamb — no thank you, ma'am! Mind you, I would still enjoy the *taste* of a BLT or well-done burger, but for moral reasons, I simply don't partake of them. I don't want to go on a crusade here, but I do need to talk a little bit about my vegetarian leanings, because

they have everything to do with souls.

Guinea Pig

When I was fifteen, I had a summer job punching buttons on a Friden mechanical calculator in a physiology lab at Stanford University. (This was back in those days when there was but one computer on the whole Stanford campus and few scientists even knew of its existence, let alone thought about using it for their calculations.) It was pretty grueling work to do such "number-punching" for hours on end, and one day, Nancy, the graduate student for whose research project I was doing all this, asked me if, for relief, I'd like to try my hand at other kinds of tasks around the lab. I said "Sure!", and so that afternoon she escorted me up to the fourth floor of the physiology building and showed me the cages where they kept the animals — literally guinea pigs — that they used in their experiments. I still remember the pungent smell and the scurrying-about of all those little orange-furred rodents.

The next afternoon, Nancy asked me if I would please go up to the top floor and bring down two animals for her next round of experiments. I didn't have a chance to reply, however, for no sooner had I started to imagine myself reaching into one of those cages and selecting two small soft furry beings to be snuffed than my head began spinning, and in a flash I fainted right away, banging my head on the concrete floor. The next thing I knew, I was looking up into the face of the lab's director, George Feigen, a dear old family friend, who was deeply concerned that I might have injured myself in the fall. Luckily I was fine, and I slowly stood up and then rode my bike home for the rest of the day. Nobody ever asked me again to pick animals to be sacrificed for the sake of science.

Pig

Oddly enough, despite that extremely troubling head-on encounter with the concept of taking the life of a living creature, I kept on eating hamburgers and other kinds of meat for several years. I don't think I thought about it very much, since none of my friends did, and certainly no one talked about it. Meat-eating was just a background fact in the life of everyone I knew. Moreover, I admit with shame that in my mind, back in those days, the word "vegetarian" conjured up an image of weird, sternly moralistic nutcases (the movie *The*

Seven Year Itch has a terrific scene in a vegetarian restaurant in Manhattan that conveys this stereotype to a tee). But one day when I was twenty-one, I read a short story called “Pig” by the Norwegian–English writer Roald Dahl, and this story had a profound effect on my life — and through me, on the lives of other creatures as well.

“Pig” starts off lightly and amusingly — a naïve young man named Lexington, raised as a strict vegetarian by his Aunt Glosspan (“Pangloss” reversed), discovers after her death that he loves the taste of meat (though he doesn’t know what it is that he’s eating). Soon, as in all Dahl stories, things take weird twists.

Driven by curiosity about this tasty substance called “pork”, Lexington, on the recommendation of a new friend, decides to take a tour of a slaughterhouse. We join him as he sits in the waiting room with other tourists. He idly watches as various waiting parties are called, one by one, to take their tours. Eventually, Lexington’s turn comes, and he is escorted from the waiting room into the shackling area where he watches pigs being hoisted by their back legs onto hooks on a moving chain, getting their throats slit, and, with blood gushing out, proceeding head downwards down the “disassembly line” to fall into a cauldron of boiling water where their body hair is removed, after which their heads and limbs are chopped off and they are prepared for being gutted and sent off, in neat little cellophane-wrapped packages, to supermarkets all over the country, where they will sit in glass cases, along with other rose-colored rivals, waiting for purchasers to admire them and hopefully to select them to take home.

As he is observing all this with detached fascination, Lexington himself is suddenly yanked by the leg and flipped upside down, and he realizes that he too is now dangling from the moving chain, just like the pigs he’s been watching. His placidity all gone, he yells out, “There has been a frightful mistake!”, but the workers ignore his cries. Soon the chain pulls him alongside a friendly-looking chap who Lexington hopes will grasp the situation’s absurdity, but instead, the gentle “sticker” grasps Lexington’s ear, pulls the dangling lad a bit closer, and then, smiling at him with loving kindness, deftly slits the boy’s jugular vein wide open with a razor-sharp knife blade. As young Lexington continues his unanticipated inverted journey, his powerful heart pumps his blood out of his throat and onto the concrete floor, and even though he is upside down and losing awareness rapidly, he dimly perceives the pigs ahead of him dropping, one by one, into the steaming cauldron. One of them, oddly enough, seems to have white gloves on its two front trotters, and he is reminded of the glove-clad young woman who had just preceded him from the waiting room into the tour area. And with that curious final thought, Lexington

woozily slips out of this, “the best of all possible worlds”, into the next.

The closing scene of “Pig” reverberated in my head for a long time. In my mind, I kept on flipping back and forth between being an upside-down oinking pig on a hook and being Lexington, spilling into the cauldron…

Revulsion, Revelation, Revolution

A month or two after reading this haunting story, I accompanied my parents and my sister Laura to the city of Cagliari, at the southern end of the rugged island of Sardinia, where my father was participating in a physics conference. To wind up the meeting in grand local style, the organizers had planned a sumptuous banquet in a park on the outskirts of Cagliari, in which a suckling piglet was to be roasted and then sliced apart in front of all the diners. As honored guests of the conference, we were all expected to take part in this venerated Sardinian tradition. I, however, was deeply under the influence of the Dahl story I had recently read, and I simply could not envision participating in such a ritual. In my new frame of mind, I couldn’t even imagine how anybody could wish to be there, let alone partake of the piglet’s body. It turned out that my sister Laura was also horrified by the prospect, and so the two of us stayed behind in our hotel and were very happy to eat some pasta and vegetables.

The one–two punch of the Norwegian “Pig” and the Sardinian piglet resulted in my following my sister’s lead in completely giving up meat-eating. I also refused to buy leather shoes or belts. Soon I became a fervent proselytizer for my new credo, and I remember how gratified I was that I managed to sway a couple of my friends for a few months, although to my disappointment, they gradually gave up on it.

In those days, I often wondered how some of my personal idols — Albert Einstein, for instance — could have been meat-eaters. I found no explanation, although recently, to my great pleasure, a Web search yielded hints that Einstein’s sympathies were, in fact, toward vegetarianism, and not for health reasons but out of compassion towards living beings. But I didn’t know that fact back then, and in any case many other heroes of mine were certainly carnivores who knew exactly what they were doing. Such facts saddened and confused me.

Reversion, Re-evolution

The very strange thing is that only a few years later, I, too, found the pressures of daily life in American society so strong that I gave up on my once-passionate vegetarianism, and for a while all my intense ruminations went totally underground. I think that the me of the mid-sixties would have found this reversal totally unfathomable, and yet the two versions of me had both lived in the very same skull. Was I really the same person?

Several years passed this way, almost as if I had never had any epiphany, but then one day, when I was a beginning assistant professor at Indiana University, I met a highly thoughtful woman who had adopted the same vegetarian philosophy as I once had, and had done so for similar reasons, but she had stuck to it for a longer time than I had. Sue and I became good friends, and I admired the purity of her stance. Our friendship caused me to think it all through once more, and in short order I had swung back to my post-"Pig" stance of no killing at all.

Over the next several years there came a few more oscillations, but by my late thirties I had finally settled into a stable state — a compromise representing my evolving intuition that there are souls of different sizes. Though it was anything but crystal-clear to me, I was willing to accept the vague idea that some souls, provided they were "small enough", could legitimately be sacrificed for the sake of the desires of "larger" souls, such as mine and those of other human beings. Although drawing the dividing line at mammals was clearly somewhat arbitrary (as any such dividing line must be), that became my new credo and I stuck with it for two more decades.

The Mystery of Inanimate Flesh

We English speakers do not eat pig or cow; we eat pork and beef. We do eat chicken — but we don't eat chickens. One time the very young daughter of a friend of mine exclaimed with great mirth to her father that the word for a certain farm bird that clucks and lays eggs was also the word for a substance that she often found on her plate at dinner time. She found this a most humorous coincidence, similar to the humorous coincidence that "calf " means both a young cow and a part of one's leg. She was upset, needless to say, when she was told that the tasty foodstuff and the clucky egg-layer were one and the same thing.

Presumably we all go through much the same confusion when, as children, we discover we are eating animals that our culture tells us are cute — lambs, bunnies, calves, chicks, and so forth. I remember, albeit dimly, my own

genuine childhood confusion at this mystery, but since meat-eating was such a bland commonplace, I usually swept it under the rug and didn't give it much thought.

Nonetheless, grocery stores had an annoying way of bringing the issue up very vividly. There were big display cases with all sorts of slimy-looking blobs of various strange colors, labeled "liver", "tripe", "heart", and "kidney", and sometimes even "tongue" and "brain". Not only did these sound like animal parts, they *looked* like them as well. Fortunately, what was called "ground beef " didn't look terribly much like an animal part, and I say "fortunately" because it tasted so good. Wouldn't want to be talked out of *that*! Bacon tasted great too, and strips of the stuff were so thin and, once cooked, so crunchy, that they hardly conjured up thoughts of an animal at all. How fortunate!

It was the unloading docks at the rear of grocery stores that made the mystery come back with a vengeance. Sometimes a big truck would pull up and when its rear doors swung open, I would see huge hunks of flesh and bones dangling lifelessly on scary-looking metal hooks. I would watch with morbid curiosity as these carcasses were carried into the back of the store and attached to hooks that slid along overhead rails, so that they could be moved around easily. All this made the preadolescent me very uneasy, and as I gazed at a carcass, I could not help musing, "Who was that animal?" I wasn't wondering about its *name,* because I knew that farm animals didn't have names; I was grasping at something more philosophical — how it had felt to be *that* animal as opposed to some *other* one. What was the unique inner light that had suddenly gone off when this animal had been slaughtered?

When I went to Europe as a teenager, the issue was raised more starkly. There, lifeless animal bodies (usually skinned, headless and tailless, but sometimes not) were on display in front of all customers. My most vivid recollection is of one grocery store that, around the Christmas season, featured the severed head of a pig on a table in the middle of an aisle. If you chanced to approach it from the rear, you would see a flat cross-section showing all the inner structures of that pig's neck, exactly as if it had been guillotined. There were all the dense communication lines that had once connected all the far-flung parts of this individual's body to the central "headquarters" in its head. Seen from the other side, this pig had what looked like a smile frozen on its face, and that gave me the creeps.

Once again, I couldn't help wondering, "Who once had been in that head? Who had lived there? Who had looked out through those eyes, heard through those ears? Who had this hunk of flesh really been? Was it a male or a female?" No answers came, of course, and no other customers seemed to pay any attention to this display. It seemed to me that nobody else was facing the

intense questions of life, death, and "porcinal identity" that this silent, still head provoked so powerfully and agitatedly inside mine.

I sometimes asked myself the analogous question if I squished an ant or clothes moth or mosquito — but not so often. My instincts told me that there was less meaning to the question "Who is 'in there'?" in such cases. Nonetheless, the sight of a partly squished insect writhing around on the floor would always give rise to some soul-searching.

And indeed, the reason I have raised all these grim images is not in order to crusade for a cause to which probably most of my readers have already given considerable thought; it is, rather, to raise the burning issue of what a "soul" is, and who or what possesses one. It is an issue that concerns everyone throughout their life — implicitly at the very least, and for many people quite explicitly — and it is the core issue of this book.

Give Me Some Men Who Are Stouter-souled Men

I alluded earlier to my deep love for the music of Chopin. In my teens and twenties, I played a lot of Chopin on the piano, often out of the bright yellow editions published by G. Schirmer in New York City. Each of those volumes opened with an essay penned in the early 1900's by the American critic James Huneker. Today, many people would find Huneker's prose overblown, but I did not; its unrestrained emotionality resonated with my perception of Chopin's music, and I still love his style of writing and his rich metaphors. In his preface to the volume of Chopin's études, Huneker asserts of the eleventh étude in Opus 25, in A minor (a titanic outburst often called the "Winter Wind", though that was certainly neither Chopin's title nor his image for it), the following striking thought: "Small-souled men, no matter how agile their fingers, should not attempt it."

I personally can attest to the terrifying technical difficulty of this incredible surging piece of music, having valiantly attempted to learn it when I was around sixteen and having sadly been forced to give it up in mid-stream, since playing just the first page up to speed (which I finally managed to do after several weeks of unbelievably arduous practice) made my right hand throb with pain. But the technical difficulty is, of course, not what Huneker was referring to. Quite rightly, he is saying that the piece is majestic and noble, but more controversially, he is drawing a dividing line between different levels or "sizes" of human souls, suggesting that some people are simply not up to playing this piece, not because of any physical limitations of their bodies, but

because their souls are not "large enough". (I won't bother to criticize the sexism of Huneker's words; that was par for the course in those days.)

This kind of sentiment does not go down well in today's egalitarian America. It would not play in Peoria. Quite frankly, it rings terribly elitist, perhaps even repugnant, to our modern democratic ears. And yet I have to admit that I somewhat agree with Huneker, and I can't help wondering if we don't all of us implicitly believe in the validity of something vaguely like the idea of "small-souled" and "large-souled" human beings. In fact, I can't help suggesting that this is indeed the belief of almost all of us, no matter how egalitarian we publicly profess to be.

Small-souled and Large-souled Humans

Some of us believe in capital punishment — the intentional public squelching of a human soul, no matter how ardently that soul would plead for mercy, would tremble, would shake, would shriek, would desperately struggle to escape, on being led down the corridor to the site of their doom.

Some of us, perhaps almost all of us, believe that it is legitimate to kill enemy soldiers in a war, as if war were a special circumstance that shrinks the sizes of enemy souls.

In earlier days, perhaps some of us would have believed (as did George Washington, Thomas Jefferson, and Benjamin Franklin, each in their own way, at least for some period of time) that it was not immoral to own slaves and to buy and sell them, breaking up families willy-nilly, just as we do today with, for example, horses, dogs, and cats.

Some religious people believe that atheists, agnostics, and followers of other faiths — and worst of all, traitors who have abandoned "the" faith — have no souls at all, and are therefore eminently deserving of death.

Some people (including some women) believe that women have no souls — or perhaps, a little more generously, that women have "smaller souls" than men do.

Some of us (myself included) believe that the late President Reagan was essentially "all gone" many years before his body gave up the ghost, and more generally we believe that people in the final stages of Alzheimer's disease are essentially all gone. It strikes us that although there is a human brain couched inside each of those cranial shells, something has gone away from that brain — something essential, something that contains the secrets of that person's soul. The "I" has either wholly or partly vanished, gone down the drain, never

to be found again.

Some of us (again, I count myself in this group) believe that neither a just-fertilized egg nor a five-month old fetus possesses a full human soul, and that, in some sense, a potential mother's life counts more than the life of that small creature, alive though it indisputably is.

Hattie the Chocolate Labrador

Kellie: After brunch we're going out to see Lynne's turkey, which we haven't seen yet.
Doug: *Which,* or *whom*?
Kellie: *Which,* I'd say. A turkey's not a *whom.*
Doug: I see… So is Hattie a *whom,* or a *which*?
Kellie: Oh, she's a *whom,* no doubt.

Ollie the Golden Retriever

Doug: So how did Ollie enjoy the outing this afternoon at Lake Griffy?
Danny: Oh, he had a pretty good time, but he didn't play much with the other dogs. He liked playing with the people, though.
Doug: Really? How come?
Danny: Ollie's a people person.

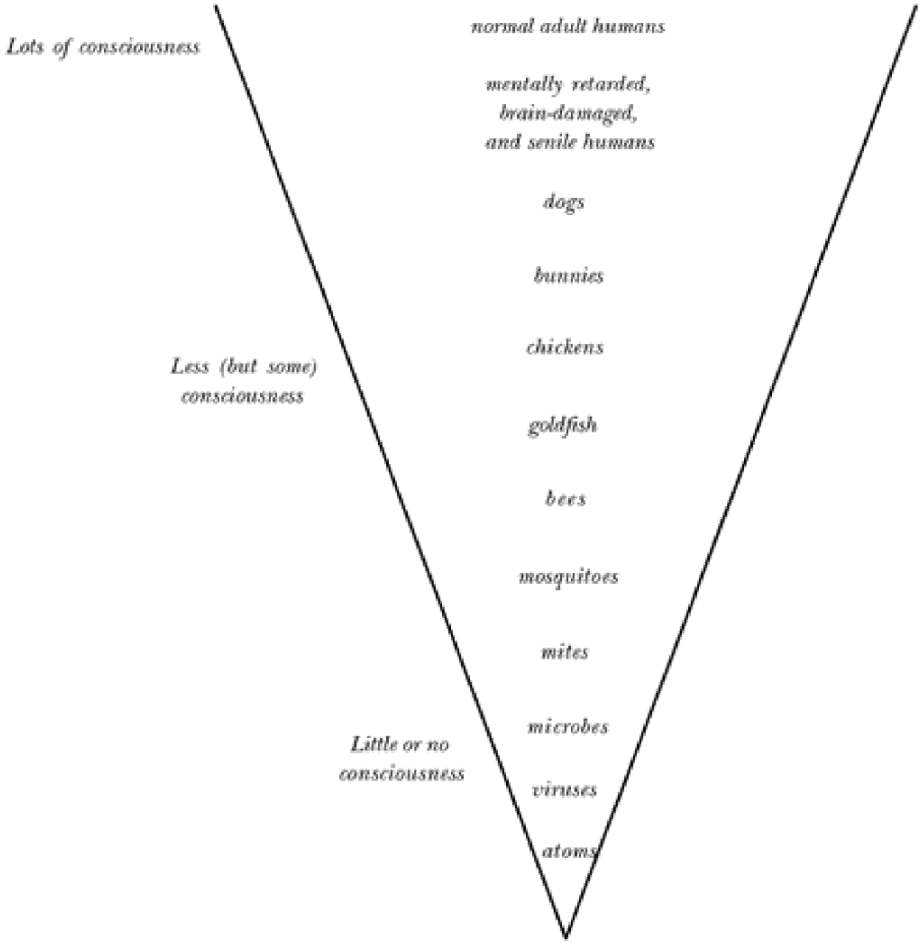
Where to Draw that Fateful, Fatal Line?

All human beings — at least all sufficiently large-souled ones — have to make up their minds about such matters as the swatting of mosquitoes or flies, the setting of mousetraps, the eating of rabbits or lobsters or turkeys or pigs, perhaps even of dogs or horses, the purchase of mink stoles or ivory statues, the usage of leather suitcases or crocodile belts, even the penicillin based attack on swarms of bacteria that have invaded their body, and on and on. The world imposes large and small moral dilemmas on us all the time — at the very least, meal after meal — and we are all forced to take a stand. Does a baby lamb have a soul that matters, or is the taste of lamb chops just too delicious to worry one's head over that? Does a trout that went for the bait and is now

helplessly thrashing about on the end of a nylon line deserve to survive, or should it just be given one sharp thwack on the head and "put out of its misery" so that we can savor the indescribable and yet strangely predictable soft, flaky texture of its white muscles? Do grasshoppers and mosquitoes and even bacteria have a tiny little "light on" inside, no matter how dim, or is it all dark "in there"? (In *where*?) Why do I not eat dogs? Who was the pig whose bacon I am enjoying for breakfast? Which tomato is it that I am munching on? Should we chop down that magnificent elm in our front yard? And while I'm at it, shall I yank out the wild blackberry bush? And all the weeds growing right by it?

What gives us word-users the right to make life-and-death decisions concerning other living creatures that have no words? Why do we find ourselves in positions of such anguish (at least for some of us)? In the final analysis, it is simply because *might makes right,* and we humans, thanks to the intelligence afforded us by the complexity of our brains and our embeddedness in rich languages and cultures, are indeed high and mighty, relative to the "lower" animals (and vegetables). By virtue of our might, we are forced to establish some sort of ranking of creatures, whether we do so as a result of long and careful personal reflections or simply go along with the compelling flow of the masses. Are cows just as comfortably killable as mosquitoes? Would you feel any less troubled by swatting a fly preening on a wall than by beheading a chicken quivering on a block? Obviously, such questions can be endlessly proliferated (note the ironic spelling of this verb), but I will not do so here.

Below, I have inserted my own personal "consciousness cone". It is not meant to be exact; it is merely suggestive, but I submit that some comparable structure exists inside your head, as well as in the head of each language-endowed human being, although in most cases it is seldom if ever subjected to intense scrutiny, because it is not even explicitly formulated.



Interiority — What Has it, and to What Degree?

It is most unlikely that you, a reader of this book, have missed all the *Star Wars* movies, with their rather unforgettable characters C-3PO and R2-D2. Absurdly unrealistic though these two robots are, especially as perceived by someone like myself who has worked for decades trying to understand just the most primordial mechanisms of human intelligence by building computational models thereof, they nonetheless serve one very useful purpose — they are mind-openers. Seeing C-3PO and R2-D2 “in flesh and blood” on the screen

makes us realize that whenever we look at an entity made of metal or plastic, we are not inherently destined to jump reflexively to the dogmatic conclusion, "That thing is necessarily an inanimate object since it is made of 'the wrong stuff '." Rather, we find, perhaps to our own surprise, that we are easily able to imagine a thinking, feeling entity made of cold, rigid, unfleshlike stuff.

In one of the *Star Wars* films, I recall seeing a huge squadron of hundreds of uniformly marching robots — and when I say "uniformly", I mean *really* uniformly, with all of them strutting in perfect synchrony, and all of them featuring identical, impassive, vacuous, mechanical facial expressions. I suspect that thanks to this unmistakable image of absolute interchangeability, virtually no viewer feels the slightest twinge of sadness when a bomb falls on the charging platoon and all of its members — these factory-made "creatures" — are instantly blown to smithereens. After all, in diametric opposition to C-3PO and R2-D2, *these* robots are not creatures at all — they are just hunks of metal! There is no more *interiority* to these metallic shells than there is to a can-opener or a car or a battleship, a fact revealed to us by their perfect identically. Or else, if perchance there is inside of them some *tiny* degree of interiority, it is on the same order as the interiority of an ant. These metallic marchers are mere soldier robots, members of a drone like caste in some larger robot colony, and are merely following out, in their zombie-ish way, the inflexible mechanical drives implanted in their circuitry. *If* there is interiority somewhere in there, it is of a negligible level.

What is it, then, that gives us the undeniable sense that C-3PO and R2-D2 have a "light on" inside, that there is lots of genuine interiority inside their inorganic crania, located somewhere behind their funny circular "eyes"? Where does our undeniable sense of their "I" 's come from? And contrariwise, what was it that was *lacking* in former President Reagan in his last years and in that mass of identical blown-up soldier robots, and what is it that is *not* lacking in Hattie the chocolate labrador and in R2-D2 the robot, that makes all the difference to us?

The Gradual Growth of a Soul

I stated above that I am among those who reject the notion that a full-fledged human soul comes into being the moment that a human sperm joins a human ovum to form a human zygote. By contrast, I believe that a human soul — and, by the way, it is my aim in this book to make clear what I mean by this slippery, shifting word, often rife with religious connotations, but here not

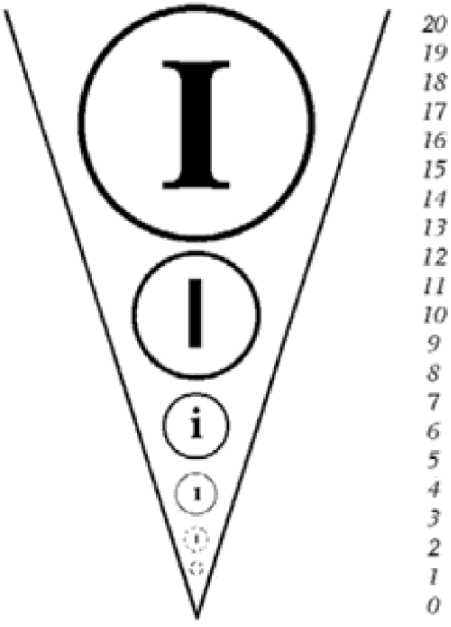
having any — comes slowly into being over the course of years of development. It may sound crass to put it this way, but I would like to suggest, at least metaphorically, a numerical scale of "degrees of souledness". We can initially imagine it as running from 0 to 100, and the units of this scale can be called, just for the fun of it, "hunekers". Thus you and I, dear reader, both possess 100 hunekers of souledness, or thereabouts. Shake!

Oops! I just realized that I have committed an error that comes from long years of indoctrination into the admirable egalitarian traditions of my native land — namely, I unconsciously assumed that there is a value at which souledness "maxes out", and that all normal adults reach that ceiling and can go no higher. Why, though, should I make any such assumption? Why could souledness not be like tallness? There is an average tallness for adults, but there is also a considerable spread around that average. Why should there not likewise be an average degree of souledness for adults (100 hunekers, say), plus a wide range around that average, maybe (as for IQ) going as high as 150 or 200 hunekers in rare cases, and down to 50 or lower in others?

If that's how things are, then I retract my reflexive claim that you and I, dear reader, share 100 hunekers of souledness. Instead, I'd like to suggest that we both have considerably *higher* readings than that on the hunekometer! (I hope you agree.) However, this is starting to feel like dangerous moral territory, verging on the suggestion that some people are *worth more* than others — a thought that is anathema in our society (and which troubles me, as well), so I won't spend much time here trying to figure out how to calculate a person's souledness value in hunekers.

It strikes me that when sperm joins ovum, the resulting infinitesimal bio-blob has a soul-value of essentially zero hunekers. What has happened, however, is that a dynamic, snowballing entity has come into existence that over a period of years will be capable of developing a complex set of internal structures or patterns — and the presence, to a higher and higher degree, of those intricate patterns is what would endow that entity (or rather, the enormously more complex entities into which it slowly metamorphoses, step by step) with an ever-larger value along the Huneker soul-scale, homing in on a value somewhere in the vicinity of 100.

The cone shown on the following page gives a crude but vivid sense of how I might attach huneker values to human beings of ages from zero to twenty (or alternatively, to just one human being, but at different stages).



In short, I would here argue, echoing and generalizing the provocative statement by James Huneker, that “souledness” is by no means an off–on, black-and-white, discrete variable having just two possible states like a bit, a pixel, or a light bulb, but rather is a shaded, blurry numerical variable that ranges continuously across different species and varieties of object, and that also can rise or fall over time as a result of the growth or decay, within the entity in question, of a special kind of subtle pattern (the elucidation of whose nature will keep us busy for much of this book). I would also argue that most people’s largely unconscious prejudices about whether to eat or not to eat this or that food, whether to buy or not to buy this or that article of clothing, whether to swat or not to swat this or that insect, whether to root or not to root for this or that species of robot in a sci-fi film, whether to be sad or not to be sad if a human character in a film or a novel meets with a violent end, whether to claim or not to claim that a particular senescent person “is no longer there”, and so forth, reflect precisely this kind of numerical continuum in their minds, whether they admit it or not.

You might wonder whether my having drawn a cone that impenitently depicts “degrees of souledness” during the development of a given human being implies that I would be more willing, if placed under enormous pressure (as in the film *Sophie’s Choice*), to extinguish the life of a two-year-old child

than the life of a twenty-year-old adult. The answer is, “No, it does not.” Even though I sincerely believe there is much more of a soul in the twenty-year-old than in the two-year-old (a view that will no doubt dismay many readers), I nonetheless have enormous respect for the *potential* of the two-year-old to *develop* a much larger soul over the course of a dozen or so years. In addition, I have been built, by the mechanisms of billions of years of evolution, to perceive in the two-year-old what, for lack of a better word, I will call “cuteness”, and the perceived presence of that quality grants the two-year-old an amazingly strong shell of protectedness against attacks not just by me, but by humans of all ages, sexes, and persuasions.

Lights On?

The central aim of this book is to try to pinpoint the nature of that “special kind of subtle pattern” that I have come to believe underlies, or gives rise to, what I have here been calling a “soul” or an “I”. I could just as well have spoken of “having a light on inside”, “possessing interiority”, or that old standby, “being conscious”.

Philosophers of mind often use the terms “possessing intentionality” (which means having beliefs and desires and fears and so forth) or “having semantics” (which means the ability to genuinely think *about* things, as contrasted with the “mere” ability to juggle meaningless tokens in complicated patterns — a distinction that I raised in the dialogue between my versions of Socrates and Plato).

Although each of these terms puts the focus on a slightly different aspect of the elusive abstraction that concerns us, they are all, from my perspective, pretty much interchangeable. And for all of these terms, I reiterate that they have to be understood as coming in *degrees* along a sliding scale, rather than as on/off, black/white, yes/no switches.

Post Scriptum

The first draft of this chapter was written two years ago, and although it discussed meat-eating and vegetarianism, it had far less on the topic than this final version does. Some months later, while I was “fleshing it out” by summarizing the short story “Pig”, I suddenly found myself questioning the dividing line that I had carefully drawn two decades earlier and had lived with

ever since (although occasionally somewhat uneasily) — namely, the line between mammals and other animals.

All at once, I started feeling distinctly uncomfortable with the idea of eating chicken and fish, even though I had done so for some twenty years, and so, catching myself by surprise, I stopped "cold turkey". And by a remarkable coincidence, my two children independently came to similar conclusions at almost exactly the same time, so that over a period of just a couple of weeks our family's diet was transmuted into a completely vegetarian one. I've returned to the same spot as I was in when I was twenty-one in Sardinia, and it's the spot I plan to stay in.

Writing this chapter thus gave rise to a totally unexpected boomerang effect on its author — and as we shall see in later chapters, such an unpredictable bouncing-back of choices one has just made, followed by the incorporation of their repercussions into one's self-model, serves as an excellent example of the meaning of the motto "I am a strange loop."

CHAPTER 2

This Teetering Bulb of Dread and Dream

What Is a "Brain Structure"?

I HAVE often been asked, when people hear that my research amounts to a quest after the hidden machinery of human thought, "Oh, so that means that you study the brain?"

One part of me wants to reply, "No, no — I think about *thinking.* I think about how concepts and words are related, what 'thinking in French' is, what underlies slips of the tongue and other types of errors, how one event effortlessly reminds us of another, how we recognize written letters and words, how we understand sloppily spoken, slurred, slangy speech, how we toss off untold numbers of utterly bland-seeming yet never-before-made analogies and occasionally come up with sparklingly original ones, how each of our concepts grows in subtlety and fluidity over our lifetime, and so forth. I don't think *in the least* about the brain — I leave the wet, messy, tangled web of the brain to the neurophysiologists."

Another part of me, however, wants to reply, "*Of course* I think about the human brain. *By definition,* I think about the brain, since the human brain is precisely the machinery that carries out human thinking."

This amusing contradiction has forced me to ask myself, "What do I mean, and what do other people mean, by 'brain research'?", and this leads naturally to the question, "What are the structures in the brain that someone could in principle study?" Most neuroscientists, if they were asked such a question, would make a list that would include (at least some of) the following items (listed roughly in order of physical size):

amino acids
neurotransmitters

DNA and RNA
synapses
dendrites
neurons
Hebbian neural assemblies
columns in the visual cortex
area 19 of the visual cortex
the entire visual cortex
the left hemisphere

Although these are all legitimate and important objects of neurological study, to me this list betrays a limited point of view. Saying that studying the brain is limited to the study of physical entities such as these would be like saying that literary criticism must focus on paper and bookbinding, ink and its chemistry, page sizes and margin widths, typefaces and paragraph lengths, and so forth. But what about the high abstractions that are the heart of literature — plot and character, style and point of view, irony and humor, allusion and metaphor, empathy and distance, and so on? Where did these crucial essences disappear in the list of topics for literary critics?

My point is simple: abstractions are central, whether in the study of literature or in the study of the brain. Accordingly, I herewith propose a list of abstractions that "researchers of the brain" should be just as concerned with:

the concept "dog"
the associative link between the concepts "dog" and "bark"
object files (as proposed by Anne Treisman)
frames (as proposed by Marvin Minsky)
memory organization packets (as proposed by Roger Schank)
long-term memory and short-term memory
episodic memory and melodic memory
analogical bridges (as proposed by my own research group)
mental spaces (as proposed by Gilles Fauconnier)
memes (as proposed by Richard Dawkins)
the ego, id, and superego (as proposed by Sigmund Freud)
the grammar of one's native language
sense of humor
"I"

I could extend this list arbitrarily. It is merely suggestive, intended to convey my thesis that the term "brain structure" should include items of this general sort. It goes without saying that some of the above-listed theoretical notions are unlikely to have lasting validity, while others may be increasingly confirmed

by various types of research. Just as the notion of "gene" as an invisible entity that enabled the passing-on of traits from parents to progeny was proposed and studied scientifically long before any physical object could be identified as an actual carrier of such traits, and just as the notion of "atoms" as the building blocks of all physical objects was proposed and studied scientifically long before individual atoms were isolated and internally probed, so any of the notions listed above might legitimately be considered as invisible structures for brain researchers to try to pinpoint physically in the human brain.

Although I'm convinced that finding the exact physical incarnation of any such structure in "the human brain" (is there only one?) would be an amazing stride forward, I nonetheless don't see why physical mapping should constitute the be-all and end-all of neurological inquiry. Why couldn't the establishment of various sorts of precise relationships among the above-listed kinds of entities, prior to (or after) physical identification, be just as validly considered brain research? This is how scientific research on genes and atoms went on for many decades before genes and atoms were confirmed as physical objects and their inner structure was probed.

A Simple Analogy between Heart and Brain

I wish to offer a simple but crucial analogy between the study of the brain and the study of the heart. In our day, we all take for granted that bodies and their organs are made of cells. Thus a heart is made of many billions of cells. But concentrating on a heart at that microscopic scale, though obviously important, risks missing the big picture, which is that *a heart is a pump.* Analogously, *a brain is a thinking machine,* and if we're interested in understanding what thinking is, we don't want to focus on the trees (or their leaves!) at the expense of the forest. The big picture will become clear only when we focus on the brain's large-scale architecture, rather than doing ever more fine-grained analyses of its building blocks.

At some point a billion years or so ago, natural selection, in its usual random-walk fashion, bumped into cells that contracted rhythmically, and little beings possessing such cells did well for themselves because the cells' contractions helped send useful stuff here and there inside the being itself. Thus, by accident, were pumps born, and in the abstract design space of all such proto-pumps, nature favored designs that were more efficient. The inner workings of the pulsating cells making up those pumps had been found, in essence, and the cells' innards thus ceased being the crucial variables that

were selected for. It was a brand-new game, in which rival *architectures* of hearts became the chief contenders for selection by nature, and on that new level, ever more complex patterns quickly evolved.

For this reason, heart surgeons don't think about the details of heart cells but concentrate instead on large architectural structures in the heart, just as car buyers don't think about the physics of protons and neutrons or the chemistry of alloys, but concentrate instead on high abstractions such as comfort, safety, fuel efficiency, maneuverability, sexiness, and so forth. And thus, to close out my heart–brain analogy, the bottom line is simply that the microscopic level may well be — or rather, almost certainly is — the wrong level in the brain on which to look, if we are seeking to explain such enormously abstract phenomena as concepts, ideas, prototypes, stereotypes, analogies, abstraction, remembering, forgetting, confusing, comparing, creativity, consciousness, sympathy, empathy, and the like.

Can Toilet Paper Think?

Simple though this analogy is, its bottom line seems sadly to sail right by many philosophers, brain researchers, psychologists, and others interested in the relationship between brain and mind. For instance, consider the case of John Searle, a philosopher who has spent much of his career heaping scorn on artificial-intelligence research and computational models of thinking, taking special delight in mocking Turing machines.

A momentary digression… Turing machines are extremely simple idealized computers whose memory consists of an infinitely long (*i.e.,* arbitrarily extensible) "tape" of so-called "cells", each of which is just a square that either is blank or has a dot inside it. A Turing machine comes with a movable "head", which looks at any one square at a time, and can "read" the cell (*i.e.,* tell if it has a dot or not) and "write" on it (*i.e.,* put a dot there, or erase a dot). Lastly, a Turing machine has, stored in its "head", a fixed list of instructions telling the head under which conditions to move left one cell or right one cell, or to make a new dot or to erase an old dot. Though the basic operations of all Turing machines are supremely trivial, any computation of any sort can be carried out by an appropriate Turing machine (numbers being represented by adjacent dot-filled cells, so that "•••" flanked by blanks would represent the integer 3).

Back now to philosopher John Searle. He has gotten a lot of mileage out of the fact that a Turing machine is an abstract machine, and therefore could, in principle, be built out of any materials whatsoever. In a ploy that, in my opinion,

should fool only third-graders but that unfortunately takes in great multitudes of his professional colleagues, he pokes merciless fun at the idea that *thinking* could ever be implemented in a system made of such far-fetched physical substrates as *toilet paper and pebbles* (the tape would be an infinite roll of toilet paper, and a pebble on a square of paper would act as the dot in a cell), or *Tinkertoys,* or a vast assemblage of *beer cans and ping-pong balls* bashing together.

In his vivid writings, Searle gives the appearance of tossing off these humorous images lightheartedly and spontaneously, but in fact he is carefully and premeditatedly instilling in his readers a profound prejudice, or perhaps merely profiting from a preexistent prejudice. After all, it *does* sound preposterous to propose "thinking toilet paper" (no matter how long the roll might be, and regardless of whether pebbles are thrown in for good measure), or "thinking beer cans", "thinking Tinkertoys", and so forth. The light-hearted, apparently spontaneous images that Searle puts up for mockery are in reality skillfully calculated to make his readers scoff at such notions without giving them further thought — and sadly, they often work.

The Terribly Thirsty Beer Can

Indeed, Searle goes very far in his attempt to ridicule the systems that he portrays in this humorous fashion. For example, to ridicule the notion that a gigantic system of interacting beer cans might "have experiences" (yet another term for consciousness), he takes *thirst* as the experience in question, and then, in what seems like a casual allusion to something obvious to everyone, he drops the idea that in such a system there would have to be *one particular can* that would "pop up" (whatever that might mean, since he conveniently leaves out all description of how these beer cans might interact) on which the English words "I am thirsty" are written. The popping-up of this single beer can (a micro-element of a vast system, and thus comparable to, say, one neuron or one synapse in a brain) is meant to constitute the system's experience of thirst. In fact, Searle has chosen this silly image very deliberately, because he knows that no one would attribute it the slightest amount of plausibility. How could a metallic beer can possibly experience thirst? And how would its "popping up" *constitute* thirst? And why should the words "I am thirsty" written on a beer can be taken any more seriously than the words "I want to be washed" scribbled on a truck caked in mud?

The sad truth is that this image is the most ludicrous possible distortion of

computer-based research aimed at understanding how cognition and sensation take place in minds. It could be criticized in any number of ways, but the key sleight of hand that I would like to focus on here is how Searle casually states that the experience claimed for this beer-can brain model is localized to *one single beer can,* and how he carefully avoids any suggestion that one might instead seek the system's experience of thirst in a more complex, more global, high-level property of the beer cans' configuration.

When one seriously tries to think of how a beer-can model of thinking or sensation might be implemented, the "thinking" and the "feeling", no matter how superficial they might be, would not be localized phenomena associated with a single beer can. They would be vast processes involving millions or billions or trillions of beer cans, and the state of "experiencing thirst" would not reside in three English words pre-painted on the side of a single beer can that popped up, but in a very intricate pattern involving huge numbers of beer cans. In short, Searle is merely mocking a trivial target of his own invention. No serious modeler of mental processes would ever propose the idea of one lonely beer can (or neuron) for each sensation or concept, and so Searle's cheap shot misses the mark by a wide margin.

It's also worth noting that Searle's image of the "single beer can as thirst-experiencer" is but a distorted replay of a long-discredited idea in neurology — that of the "grandmother cell". This is the idea that your visual recognition of your grandmother would take place if and only if one special cell in your brain were activated, that cell constituting your brain's physical representation of your grandmother. What significant difference is there between a grandmother cell and a thirst can? None at all. And yet, because John Searle has a gift for catchy imagery, his specious ideas have, over the years, had a great deal of impact on many professional colleagues, graduate students, and lay people.

It's not my aim here to attack Searle in detail (that would take a whole dreary chapter), but to point out how widespread is the tacit assumption that the level of the most primordial physical components of a brain must *also* be the level at which the brain's most complex and elusive mental properties reside. Just as many aspects of a mineral (its density, its color, its magnetism or lack thereof, its optical reflectivity, its thermal and electrical conductivity, its elasticity, its heat capacity, how fast sound spreads through it, and on and on) are properties that come from how its billions of atomic constituents interact and form high-level patterns, so mental properties of the brain reside not on the level of a single tiny constituent but on the level of *vast abstract patterns* involving those constituents.

Dealing with brains as multi-level systems is essential if we are to make even the slightest progress in analyzing elusive mental phenomena such as

perception, concepts, thinking, consciousness, "I", free will, and so forth. Trying to localize a concept or a sensation or a memory (etc.) down to a single neuron makes no sense at all. Even localization to a higher level of structure, such as a column in the cerebral cortex (these are small structures containing on the order of forty neurons, and they exhibit a more complex collective behavior than single neurons do), makes no sense when it comes to aspects of thinking like analogy-making or the spontaneous bubbling-up of episodes from long ago.

Levels and Forces in the Brain

I once saw a book whose title was "Molecular Gods: How Molecules Determine Our Behavior". Although I didn't buy it, its title stimulated many thoughts in my brain. (What is *a thought in a brain*? Is a thought really *inside* a brain? Is a thought made of molecules?) Indeed, the very fact that I soon placed the book back up on the shelf is a perfect example of the kinds of thoughts that its title triggered in my brain. What exactly determined my behavior that day (*e.g.,* my interest in the book, my pondering about its title, my decision not to buy it)? Was it some *molecules* inside my brain that made me reshelve it? Or was it some *ideas* in my brain? What is the proper way to talk about what was going on in my head as I first flipped through that book and then put it back?

At the time, I was reading books by many different writers on the brain, and in one of them I came across a chapter by the neurologist Roger Sperry, which not only was written with a special zest but also expressed a point of view that resonated strongly with my own intuitions. I would like to quote here a short passage from Sperry's essay "Mind, Brain, and Humanist Values", which I find particularly provocative.

> In my own hypothetical brain model, conscious awareness does get representation as a very real causal agent and rates an important place in the causal sequence and chain of control in brain events, in which it appears as an active, operational force....
>
> To put it very simply, it comes down to the issue of who pushes whom around in the population of causal forces that occupy the cranium. It is a matter, in other words, of straightening out the peck-order hierarchy among intracranial control agents. There exists within the cranium a whole world of diverse causal forces; what is more, there are forces within forces within forces, as in no other cubic half-foot of universe that

we know….

To make a long story short, if one keeps climbing upward in the chain of command within the brain, one finds at the very top those over-all organizational forces and dynamic properties of the large patterns of cerebral excitation that are correlated with mental states or psychic activity…. Near the apex of this command system in the brain…. we find ideas.

Man over the chimpanzee has ideas and ideals. In the brain model proposed here, the causal potency of an idea, or an ideal, becomes just as real as that of a molecule, a cell, or a nerve impulse. Ideas cause ideas and help evolve new ideas. They interact with each other and with other mental forces in the same brain, in neighboring brains, and, thanks to global communication, in far distant, foreign brains. And they also interact with the external surroundings to produce *in toto* a burst-wise advance in evolution that is far beyond anything to hit the evolutionary scene yet, including the emergence of the living cell.

Who Shoves Whom Around Inside the Cranium?

Yes, reader, I ask you: Who shoves whom around in the tangled megaganglion that is your brain, and who shoves whom around in "this teetering bulb of dread and dream" that is mine? (The marvelously evocative phrase in quotes, serving also as this chapter's title, is taken from "The Floor" by American poet Russell Edson.)

Sperry's pecking-order query puts its finger on what we need to know about ourselves — or, more pointedly, about our *selves.* What was *really* going on in that fine brain on that fine day when, allegedly, something calling itself "I" did something called "deciding", after which a jointed appendage moved in a fluid fashion and a book found itself back where it had been just a few seconds before? Was there truly something referable-to as "I" that was "shoving around" various physical brain structures, resulting in the sending of certain carefully coordinated messages through nerve fibers and the consequent moving of shoulder, elbow, wrist, and fingers in a certain complex pattern that left the book upright in its original spot — or, contrariwise, were there merely myriads of microscopic physical processes (quantum-mechanical collisions involving electrons, photons, gluons, quarks, and so forth) taking place in that localized region of the spatiotemporal continuum that poet Edson dubbed a "teetering bulb"?

as well as its flipped-around version: *Statistical mentalics can be bypassed by talking at the level of thinkodynamics.*

What do I mean by these two terms, "thinkodynamics" and "statistical mentalics"? It is pretty straightforward. Thinkodynamics is analogous to thermodynamics; it involves large-scale structures and patterns in the brain, and makes no reference to microscopic events such as neural firings. Thinkodynamics is what psychologists study: how people make choices, commit errors, perceive patterns, experience novel remindings, and so on.

By contrast, by "mentalics" I mean the small-scale phenomena that neurologists traditionally study: how neurotransmitters cross synapses, how cells are wired together, how cell assemblies reverberate in synchrony, and so forth. And by "statistical mentalics", I mean the averaged-out, collective behavior of these very small entities — in other words, the behavior of a huge swarm as a whole, as opposed to a tiny buzz inside it.

However, as neurologist Sperry made very clear in the passage cited above, there is not, in the brain, just one single natural upward jump, as there is in a gas, all the way from the basic constituents to the whole thing; rather, there are many way-stations in the upward passage from mentalics to thinkodynamics, and this means that it is particularly hard for us to see, or even to imagine, the ground-level, neural-level explanation for why a certain professor of cognitive science once chose to reshelve a certain book on the brain, or once refrained from swatting a certain fly, or once broke out in giggles during a solemn ceremony, or once exclaimed, lamenting the departure of a cherished co-worker, "She'll be hard shoes to fill!"

The pressures of daily life require us, force us, to talk about events *at the level on which we directly perceive them.* Access at that level is what our sensory organs, our language, and our culture provide us with. From earliest childhood on, we are handed concepts such as "milk", "finger", "wall", "mosquito", "sting", "itch", "swat", and so on, on a silver platter. We perceive the world in terms of such notions, not in terms of microscopic notions like "proboscis" and "hair follicle", let alone "cytoplasm", "ribosome", "peptide bond", or "carbon atom". We can of course acquire such notions later, and some of us master them profoundly, but they can never replace the silver-platter ones we grew up with. In sum, then, we are victims of our macroscopicness, and cannot escape from the trap of using everyday words to describe the events that we witness, and perceive as *real.*

This is why it is much more natural for us to say that a war was triggered for religious or economic reasons than to try to imagine a war as a vast pattern of interacting elementary particles and to think of what triggered it in similar terms — even though physicists may insist that that is the only "true" level of

explanation for it, in the sense that no information would be thrown away if we were to speak at that level. But having such phenomenal accuracy is, alas (or rather, "Thank God!"), not our fate.

We mortals are condemned *not* to speak at that level of no information loss. We *necessarily* simplify, and indeed, vastly so. But that sacrifice is also our glory. Drastic simplification is what allows us to reduce situations to their bare bones, to discover abstract essences, to put our fingers on what matters, to understand phenomena at amazingly high levels, to survive reliably in this world, and to formulate literature, art, music, and science.

CHAPTER 3

The Causal Potency of Patterns

The Prime Mover

AS THE rest of this book depends on having a clear sense for the interrelationships between different levels of description of entities that think, I would like to introduce here a few concrete metaphors that have helped me a great deal in developing my intuitions on this elusive subject.

My first example involves the familiar notion of a chain of falling dominos. However, I'll jazz up the standard image a bit by stipulating that each domino is spring-loaded in a clever fashion (details do not concern us) so that whenever it gets knocked down by its neighbor, after a short "refractory" period it flips back up to its vertical state, all set to be knocked down once more. With such a system, we can implement a mechanical computer that works by sending signals down stretches of dominos that can bifurcate or join together; thus signals can propagate in loops, jointly trigger other signals, and so forth. Relative timing, of course, will be of the essence, but once again, details do not concern us. The basic idea is just that we can imagine a network of precisely timed domino chains that amounts to a computer program for carrying out a particular computation, such as determining if a given input is a prime number or not. (John Searle, so fond of unusual substrates for computation, should like this "domino chainium" thought experiment!)

Let us thus imagine that we can give a specific numerical "input" to the chainium by taking any positive integer we are interested in — 641, say — and placing exactly that many dominos end to end in a "reserved" stretch of the network. Now, when we tip over the chainium's first domino, a Rube Goldberg–type series of events will take place in which domino after domino will fall, including, shortly after the outset, all 641 of the dominos constituting our input stretch, and as a consequence various loops will be triggered, with

some loop presumably testing the input number for divisibility by 2, another for divisibility by 3, and so forth. If ever a divisor is found, then a signal will be sent down one particular stretch — let's call it the "divisor stretch" — and when we see that stretch falling, we will know that the input number has some divisor and thus is not prime. By contrast, if the input has no divisor, then the divisor stretch will never be triggered and we will know the input is prime.

Suppose an observer is standing by when the domino chainium is given 641 as input. The observer, who has not been told what the chainium was made for, watches keenly for while, then points at one of the dominos in the divisor stretch and asks with curiosity, "How come that domino there is never falling?"

Let me contrast two very different types of answer that someone might give. The first type of answer — myopic to the point of silliness — would be, "Because its predecessor never falls, you dummy!" To be sure, this is correct as far as it goes, but it doesn't go very far. It just pushes the buck to a different domino, and thus begs the question.

The second type of answer would be, "Because 641 is prime." Now this answer, while just as correct (indeed, in some sense it is far more on the mark), has the curious property of not talking about anything physical at all. Not only has the focus moved upwards to collective properties of the chainium, but those properties somehow transcend the physical and have to do with pure abstractions, such as primality.

The second answer bypasses all the physics of gravity and domino chains and makes reference only to concepts that belong to a completely different domain of discourse. The domain of prime numbers is as remote from the physics of toppling dominos as is the physics of quarks and gluons from the Cold War's "domino theory" of how communism would inevitably topple country after neighboring country in Southeast Asia. In both cases, the two domains of discourse are many levels apart, and one is purely local and physical, while the other is global and organizational.

Before passing on to other metaphors, I'd just like to point out that although here, 641's primality was used as an explanation for why a certain domino did *not* fall, it could equally well serve as the explanation for why a different domino *did* fall. In particular, in the domino chainium, there could be a stretch called the "prime stretch" whose dominos all topple when the set of potential divisors has been exhausted, which means that the input has been determined to be prime.

The point of this example is that 641's primality is the best explanation, perhaps even the *only* explanation, for why certain dominos *did* fall and certain other ones *did not* fall. In a word, 641 is the prime mover. So I ask: Who shoves whom around inside the domino chainium?

The Causal Potency of Collective Phenomena

My next metaphor was dreamt up on an afternoon not long ago when I was caught in a horrendous traffic jam on some freeway out in the countryside, with several lanes of nearly touching cars all sitting stock still. For some reason I was reminded of big-city traffic jams where you often hear people honking angrily at each other, and I imagined myself suddenly starting to honk my horn over and over again at the car in front of me, as if to say, "Get out of my way, lunkhead!"

The thought of myself (or anyone) taking such an outrageously childish action made me smile, but when I considered it a bit longer, I saw that there might be a slim rationale for honking that way. After all, if the next car were magically to poof right out of existence, I could fill the gap and thus make one car-length's worth of progress. Now a car poofing out of existence is not too terribly likely, and one car-length is not much progress, but somehow, through this image, the idea of honking became just barely comprehensible to me. And then I remembered my domino chainium and the silly superlocal answer, "That domino didn't fall because its neighbor didn't fall, you dummy!" This myopic answer and my fleeting thought of honking at the car just ahead of me seemed to be cut from the same cloth.

As I continued to sit in this traffic jam, twiddling my thumbs instead of honking, I let these thoughts continue, in their bully-like fashion, to push my helpless neurons around. I imagined a counterfactual situation in which the highway was shrouded in the densest pea-soup fog imaginable, so that I could barely make out the rear of the car ahead of me. In such a case, honking my horn wouldn't be quite so blockheaded. For all I know, that car alone might well be the entire cause of my being stuck, and if only it would just get out of the way, I could go sailing down the highway!

If you're totally fog-bound like that, or if you're incredibly myopic, then you might think to yourself, "It's all my neighbor's fault!", and there's at least a small chance that you're right. But if you have a larger field of view and can see hordes of immobilized cars on all sides, then honking at your immediate predecessor is an absurdity, for it's obvious that the problem is not local. The root problem lies at some level of discourse other than that of cars. Though you may not know its nature, some higher-level, more abstract reason must lie behind this traffic jam.

Perhaps a very critical baseball game just finished three miles up the road. Perhaps it's 7:30 on a weekday morning and you're heading towards Silicon Valley. Perhaps there's a huge blizzard ten miles ahead. Or it may be

The Strange Irrelevance of Lower Levels

This idea — that the bottom level, though 100 percent *responsible* for what is happening, is nonetheless *irrelevant* to what happens — sounds almost paradoxical, and yet it is an everyday truism. Since I want this to be crystal-clear, let me illustrate it with one more example.

Consider the day when, at age eight, I first heard the fourth étude of Chopin's Opus 25 on my parents' record player, and instantly fell in love with it. Now suppose that my mother had placed the needle in the groove a millisecond later. One thing for sure is that all the molecules in the room would have moved completely differently. If you had been one of those molecules, you would have had a wildly different life story. Thanks to that millisecond delay, you would have careened and bashed into completely different molecules in utterly different places, spun off in totally different directions, and on and on, *ad infinitum.* No matter which molecule you were in the room, your life story would have turned out unimaginably different. But would any of that have made an iota of difference to the life story of the kid listening to the music? No — not the teensiest, tiniest iota of difference. All that would have mattered was that Opus 25, number 4 got transmitted faithfully through the air, and *that* would most surely have happened. *My* life story would not have been changed in any way, shape, or form if my mother had put the needle down in the groove a millisecond earlier or later. Or a second earlier or later.

Although the air molecules were crucial mediating agents for a series of high-level events involving a certain kid and a certain piece of music, their precise behavior was not crucial. Indeed, saying it was "not crucial" is a ridiculous understatement. Those air molecules could have done exactly the same kid–music job in an astronomical number of different but humanly indistinguishable fashions. The lower-level laws of their collisions played a role only in that they gave rise to predictable high-level events (propagation of the notes in the Chopin étude to little Douggie's ear). But the positions, speeds, directions, even the chemical identity of the molecules — all of this was changeable, and the high-level events would have been the same. It would have been the same music to my ears. One can even imagine that the microscopic laws of physics could have been different — what matters is not the detailed laws but merely the fact that they reliably give rise to stable statistical consequences.

Flip a quarter a million times and you'll very reliably get within one percent of 500,000 heads. Flip a penny the same number of times, and the same statement holds. Use a different coin on every flip — dimes, quarters, new

pennies, old pennies, buffalo nickels, silver dollars, you name it — and still you'll get the same result. Shave your penny so that its outline is hexagonal instead of circular — no difference. Replace the hexagonal outline by an elephant shape. Dip the penny in apple butter before each flip. Bat the penny high into the air with a baseball bat instead of tossing it up. Flip the penny in helium gas instead of air. Do the experiment on Mars instead of Earth. These and countless other variations on the theme will not have any effect on the fact that out of a million tosses, within one percent of 500,000 will wind up heads. That high-level statistical outcome is robust and invariant against the details of the substrate and the microscopic laws governing the flips and bounces; the high-level outcome is insulated and sealed off from the microscopic level. It is a fact in its own right, at its own level.

That is what it means to say that although what happens on the lower level is *responsible* for what happens on the higher level, it is nonetheless *irrelevant* to the higher level. The higher level can blithely ignore the processes on the lower level. As I put it in Chapter 2, "Our existence as animals whose perception is limited to the world of everyday macroscopic objects forces us, quite obviously, to function without any reference to entities and processes at microscopic levels. No one really knew the slightest thing about atoms until only about a hundred years ago, and yet people got along perfectly well."

A Hat-tip to the Spectrum of Unpredictability

I am not suggesting that the invisible, swarming, chaotic, microscopic level of the world can be totally swept under the rug and forgotten. Although in many circumstances we rely on the familiar macroworld to be completely predictable to us, there are many other circumstances where we are very aware of not being able to predict what will happen. Let me first, however, make a little list of some sample predictables that we rely on unthinkingly all the time.

When we turn our car's steering wheel, we know for sure where our car will go; we don't worry that a band of recalcitrant little molecules might mutiny and sabotage our turn. When we turn a burner to "high" under a saucepan filled with water, we know that the water will boil within a few minutes. We can't predict the pattern of bubbles inside the boiling water, but we really don't give a hoot about that. When we take a soup can down from the shelf in the grocery store and place it in our cart, we know for sure that it will not turn into a bag of potato chips, will not burn our hand, will not be so heavy that we cannot lift it, will not slip through the grill of the cart, will sit still if placed vertically, and so

forth. To be sure, if we lay the soup can down horizontally and start wheeling the cart around the store, the can will roll about in the cart in ways that are not predictable to us, though they lie completely within the bounds of our expectations and have little interest or import to us, aside from being mildly annoying.

When we speak words, we know that they will reach the ears of our listeners without being changed by the intermediary pressure waves into other words, will even come through with the exact intonations that we impart to them. When we pour milk into a glass, we know just how far to tilt the milk container to get the desired amount of flow without spilling a drop. We control the milk and we get exactly the result we want.

There is no surprise in any of this! And I could extend this list forever, and it would soon grow very boring, because you know it all instinctively and take it totally for granted. Every day of our lives, we all depend in a million tacit ways on innumerable rock-solid predictabilities about how things happen in the visible, tangible world (the solidity of rocks being yet another of those countless rock-solid predictabilities).

On the other hand, there's also plenty of unpredictability "up here" in the macroworld. How about a second list, giving typical unpredictables?

When we toss a basketball towards a basket, we don't have any idea whether it will go through or not. It might bounce off the backboard and then teeter for a couple of seconds on the rim, keeping us in suspense and perhaps even holding an entire crowd in tremendous, tingling tension. A championship basketball game could go one way or the other, depending on a microscopic difference in the position of the pinky of the player who makes a desperate last-second shot.

When we begin to utter a thought, we have no idea what words we will wind up using nor which grammatical pathways we will wind up following, nor can we predict the speech errors or the facts about our unconscious mind that our little slips will reveal. Usually such revelations will make little difference, but once in a while — in a job interview, say — they can have huge repercussions. Think of how people jump on a politician whose unconscious mind chooses a word loaded with political undertones (*e.g.,* "the crusade against terrorism").

When we ski down a slope, we don't know if we're going to fall on our next turn or not. Every turn is a risk — slight for some, large for others. A broken bone can come from an event whose cause we will never fathom, because it is so deeply hidden in detailed interactions between the snow and our ski. And the tiniest detail about the manner in which we fall can make all the difference as to whether we suffer a life-changing multiple break or a just a trivial hairline fracture.

The macroscopic world as experienced by humans is, in short, an intimate mixture ranging from the most predictable events all the way to wildly unpredictable ones. Our first few years of life familiarize us with this spectrum, and the degree of predictability of most types of actions that we undertake becomes second nature to us. By the time we emerge from childhood, we have acquired a reflex-level intuition for where most of our everyday world's loci of unpredictability lie, and the more unpredictable end of this spectrum simultaneously beckons to us and frightens us. We're pulled by but fearful of risk-taking. That is the nature of life.

The Careenium

I now move to a somewhat more complex metaphor for thinking about the multiple levels of causality in our brains and minds (and eventually, if you will indulge me in this terminology, in our souls). Imagine an elaborate frictionless pool table with not just sixteen balls on it, but myriads of extremely tiny marbles, called "sims" (an acronym for "small interacting marbles"). These sims bash into each other and also bounce off the walls, careening about rather wildly in their perfectly flat world — and since it is frictionless, they just keep on careening and careening, never stopping.

So far our setup sounds like a two-dimensional ideal gas, but now we'll posit a little extra complexity. The sims are also magnetic (so let's switch to "simms", with the extra "m" for "magnetic"), and when they hit each other at lowish velocities, they can stick together to form clusters, which I hope you will pardon me for calling "simmballs". A simmball consists of a very large number of simms (a thousand, a million, I don't care), and on its periphery it frequently loses a few simms while gaining others. There are thus two extremely different types of denizen of this system: tiny, light, zipping simms, and giant, ponderous, nearly-immobile simmballs.

The dynamics taking place on this pool table — hereinafter called the "careenium" — thus involves simms crashing into each other and also into simmballs. To be sure, the details of the physics involve transfers of momentum, angular momentum, kinetic energy, and rotational energy, just as in a standard gas, but we won't even think about that, because this is just a *thought* experiment (in two senses of the term). All that matters for our purposes is that there are these collisions taking place all the time.

Simmballism

Why the corny pun on "symbol"? Because I now add a little more complexity to our system. The vertical walls that constitute the system's boundaries react sensitively to outside events (*e.g.,* someone touching the outside of the table, or even a breeze) by momentarily flexing inward a bit. This flexing, whose nature retains some traces of the external causing event, of course affects the motions of the simms that bounce internally off that section of wall, and indirectly this will be registered in the slow motions of the nearest simmballs as well, thus allowing the simmballs to *internalize* the event. We can posit that one particular simmball always reacts in some standard fashion to breezes, another to sharp blows, and so forth. Without going into details, we can even posit that the configurations of simmballs *reflect the history* of the impinging outer-world events. In short, for someone who looked at the simmballs and knew how to read their configuration, the simmballs would be *symbolic,* in the sense of *encoding events.* That's why the corny pun.

Of course this image is far-fetched, but remember that the careenium is merely intended as a useful metaphor for understanding our brains, and the fact is that our brains, too, are rather far-fetched, in the sense that they too contain tiny events (neuron firings) and larger events (patterns of neuron firings), and the latter presumably somehow have *representational* qualities, allowing us to register and also to remember things that happen outside of our crania. Such internalization of the outer world in symbolic patterns in a brain is a pretty far-fetched idea, when you think about it, and yet we know it somehow came to exist, thanks to the pressures of evolution. If you wish, then, feel free to imagine that careenia, too, evolved. You can think of them as emerging as the end result of billions of more primitive systems fighting for survival in the world. But the evolutionary origins of our careenium need not concern us here. The key idea is that whereas no simm on its own encodes anything or plays a symbolic role, the simmballs, on their far more macroscopic level, *do* encode and *are* symbolic.

Taking the Reductionistic View of the Careenium

The first inclination of a modern physicist who heard this story might be reductionistic, in the sense of pooh-poohing the large simmballs as mere *epiphenomena,* meaning that although they are undeniably *there,* they are not essential to an understanding of the system, since they are composed of

an electric fan. The second shift is that we spatially back away or zoom out, thus rendering simms too small to be seen, and so the simmballs alone necessarily become our focus of attention.

Now we see a completely different type of dynamics on the table. Instead of seeing simms bashing into what look like large stationary blobs, we realize that these blobs are not stationary at all but have a lively life of their own, moving back and forth across the table and interacting with each other, as if there were nothing else on the table but them. Of course we know that deep down, this is all happening thanks to the teeny-weeny simms' bashing-about, *but we cannot see the simms any more.* In our new way of seeing things, their frenetic careening-about on the table forms nothing but a stationary gray background.

Think of how the water in a glass sitting on a table seems completely still to us. If our eyes could shift levels (think of the twist that zooms binoculars in or out) and allow us to peer at the water at the micro-level, we would realize that it is not peaceful at all, but a crazy tumult of bashings of water molecules. In fact, if colloidal particles are added to a glass of water, then it becomes a locus of Brownian motion, which is an incessant random jiggling of the colloidal particles, due to a myriad of imperceptible collisions with the water molecules, which are far tinier. (The colloidal particles here play the role of simmballs, and the water molecules play the role of simms.) The effect, which is visible under a microscope, was explained in great detail in 1905 by Albert Einstein using the theory of molecules, which at the time were only hypothetical entities, but Einstein's explanation was so far-reaching (and, most crucially, consistent with experimental data) that it became one of the most important confirmations that molecules do exist.

Who Shoves Whom Around inside the Careenium?

And so we finally have come to the crux of the matter: *Which of these two views of the careenium is the truth?* Or, to echo the key question posed by Roger Sperry, *Who shoves whom around in the population of causal forces that occupy the careenium?* In one view, the meaningless tiny simms are the primary entities, zipping around like mad, and in so doing they very slowly push the heavy, passive simmballs about, hither and thither. In this view, it is the tiny simms that shove the big simmballs around, and that is all there is to it. In fact, in this view the simmballs are not even recognized as separate entities, since anything we might say about their actions is just a shorthand way of talking about what simms do. From this perspective, there are no simmballs,

no symbols, no ideas, no thoughts going on — just a great deal of tumultuous, pointless careening-about of tiny, shiny, magnetic spheres.

In the other view, speeded up and zoomed out, all that is left of the shiny tiny simms is a featureless gray soup, and the interest resides solely in the simmballs, which give every appearance of richly interacting with each other. One sees groups of simmballs triggering other simmballs in a kind of "logic" that has nothing to do with the soup churning around them, except in the rather pedestrian sense that the simmballs derive their *energy* from that omnipresent soup. Indeed, the simmballs' logic, not surprisingly, has to do with the *concepts* that the simmballs symbolize.

The Dance of the Simmballs

From our higher-level macroscopic vantage point as we hover above the table, we can see *ideas* giving rise to other *ideas,* we can see one symbolic event *reminding* the system of another symbolic event, we can see elaborate patterns of simmballs coming together and forming even larger patterns that constitute *analogies* — in short, we can visually eavesdrop on the logic of a thinking mind taking place in the patterned dance of the simmballs. And in this latter view, *it is the simmballs that shove each other about,* at their own isolated symbolic level.

The simms are still there, to be sure, but they are simply serving the simmballs' dance, allowing it to happen, with the microdetails of their bashings being no more relevant to the ongoing process of cognition than the microdetails of the bashings of air molecules are relevant to the turning of the blades of a windmill. Any old air-molecule bashings will do — the windmill will turn no matter what, thanks to the aerodynamic nature of its blades. Likewise, any old simm-bashings will do — the "thoughtmill" will churn no matter what, thanks to the symbolic nature of its simmballs.

If any of this strikes you as too far-fetched to be plausible, just return to the human brain and consider what must be going on inside it in order to allow our thinking's logic to take place. What else is going on inside every human cranium but some story like this?

Of course we have come back to the question that that long-ago shelved book's title made me ask, and the question that Roger Sperry also asked: Who is shoving whom about in here? And the answer is that it all depends on what level you choose to focus on. Just as, on one level, the primality of 641 could legitimately be said to be shoving about dominos in the domino-chain network,

so here there is a level on which the meanings attached to various simmballs can legitimately be said to be shoving other simmballs about. If this all seems topsy-turvy, it certainly is — but it is nonetheless completely consistent with the fundamental causality of the laws of physics.

CHAPTER 4

Loops, Goals, and Loopholes

The First Flushes of Desire

WHEN the first mechanical systems with feedback in them were designed, a set of radically new ideas began coming into focus for humanity. Among the earliest of such systems was James Watt's steam-engine governor; subsequent ones, which are numberless, include the float-ball mechanism governing the refilling of a flush toilet, the technology inside a heat-seeking missile, and the thermostat. Since the flush toilet is probably the most familiar and the easiest to understand, let's consider it for a moment.

A flush toilet has a pipe that feeds water into the tank, and as the water level rises, it lifts a hollow float. Attached to the rising float is a rigid rod whose far end is fixed, so that the rod's angle of tilt reflects the amount of water in the tank. This variable angle controls a valve that regulates the flow of water in the pipe. Thus at a critical level of filling, the angle reaches a critical value and the valve closes totally, thereby shutting off all flow in the pipe. However, if there is leakage from the tank, the water level gradually falls, and of course the float falls with it, the valve opens, and the inflow of water is thereby turned back on. Thus one sometimes gets into cyclic situations where, because a little rubber gizmo didn't land exactly centered on the tank's drain right after a flush, the tank slowly leaks for a few minutes, then suddenly fills for a few seconds, then again slowly leaks for a few minutes, then again fills for a few seconds, and so on, in a cyclic pattern that somewhat resembles breathing, and that never stops — that is, not until someone jiggles the toilet handle, thus jiggling the rubber gizmo, hopefully making it land properly on the drain, thus fixing the leak.

Once a friend of mine who was watching my house while I was away for a few weeks' vacation flushed the toilet on the first day and, by chance, the little

rubber gizmo didn't fall centered, so this cycle was entered. My friend diligently returned a few times to check out the house but he never noticed anything untoward, so the toilet tank kept on leaking and refilling periodically for my entire absence, and as a result I had a $300 water bill. No wonder people are suspicious of feedback loops!

We might anthropomorphically describe a flush toilet as a system that is "trying" to make the water reach and stay at a certain level. Of course, it's easy to bypass such anthropomorphic language since we effortlessly see how the mechanism works, and it's pretty clear that such a simple system has no desires; even so, when working on a toilet whose tank has sprung a leak, one might be tempted to say the toilet is "trying" get the water up to the mark but "can't". One doesn't *truly* impute desires or frustrations to the device — it's just a manner of speaking — but it is a convenient shorthand.

A Soccer Ball Named Desire

Why does this move to a goal-oriented — that is, *teleological* — shorthand seem appealing to us for a system endowed with feedback, but not so appealing for a less structured system? It all has to do with the way the system's "perceptions" feed back (so to speak) into its behavior. When the system always moves towards a certain state, we see that state as the system's "goal". It is the self-monitoring, self-controlling nature of such a system that tempts us to use teleological language.

But what kinds of systems have feedback, have goals, have desires? Does a soccer ball rolling down a grassy hill "want" to get to the bottom? Most of us, reflexively recoiling at such a primitive Aristotelian conception of why things move, would answer *no* without hesitation. But let's modify the situation just a tiny amount and ask the question again.

What about a soccer ball zipping down a long, narrow roadside gutter having a U-shaped cross-section — is it seeking any goal? Such a ball, as it speeds along, will first roll up one side of the gutter and then fall back to the center, cross it and then roll up the other side, then again back down, and so forth, gradually converging from a sinusoidal pathway wavering about the gutter's central groove to a straight pathway at the bottom of the gutter. Is there "feedback" here or not? Is this soccer ball "seeking" the gutter's mid-line? Does it "want" to be rolling along the gutter's valley? Well, as this example and the previous one of the ball rolling down a hill show, the presence or absence of feedback, goals, or desires is not a black-and-white matter; such things are

Fallacy the First

The primary fallacy in this scenario is that we have not taken into account the actual device carrying out the exponential process — the sound system itself, and in particular the amplifier. To make my point in the most blatant manner, I need merely remind you that the moment the auditorium's roof collapsed, it would land on the amplifier and smash it to bits, thus bringing the out-of-control feedback loop to a swift halt. The little system contains the seeds of its own destruction!

But there is something specious about this scenario, too, because as we all know, things never get that far. The auditorium never collapses, nor are the audience members deafened by the din. Something slows down the runaway process far earlier. What is that thing?

Fallacy the Second

The other fallacy in our reasoning also involves a type of self-destruction of the sound system, but it is subtler than being smashed to smithereens. It is that as the sound gets louder and louder, the amplifier stops amplifying with that constant factor of *k.* At a certain level it starts to fail. Just as a floored car will not continue accelerating at a constant rate (reaching 100 miles per hour, then 200, 300, 400, soon breaking the sound barrier, etc.) but eventually levels out at some peak velocity (which is a function of road friction, air resistance, the motor's internal limits, and so forth), so an amplifier will not uniformly amplify sounds of any volume but will eventually saturate, giving less and less amplification until at some volume level the output sound has the same volume as the input sound, and that is where things stabilize. The volume at which the amplification factor becomes equal to 1 is that of the familiar screech that drives you mad but doesn't deafen you, much less brings the auditorium crashing down on your head.

And why does it always give off that same high-pitched screeching sound? Why not a low roar? Why not the sound of a waterfall or a jet engine or long low thunder? This has to do with the natural resonance frequency of the system — an acoustic analogue of the natural oscillation frequency of a playground swing, roughly once every couple of seconds. An amplifier's feedback loop has a natural oscillation frequency, too, and for reasons that need not concern us, it usually has a pitch close to that of a high-frequency scream. However, the system does not instantly settle down precisely on its

final pitch. If you could drastically slow down the process, you would hear it homing in on that squealing pitch much as the rolling soccer ball seeks the bottom of the gutter — namely, by means of a very rapid series of back-and-forth swings in frequency, almost as if it "wanted" to reach that natural spot in the sonic spectrum.

What we have seen here is that even the simplest imaginable feedback loop has levels of subtlety and complexity that are seldom given any thought, but that turn out to be rich and full of surprise. Imagine, then, what happens in the case of more complex feedback loops.

Feedback and Its Bad Rap

The first time my parents wanted to buy a video camera, sometime in the 1970's, I went to the store with them and we asked to see what they had. We were escorted to an area of the store that had several TV screens on a shelf, and a video camera was plugged into the back of one of them, thus allowing us to see what the camera was looking at and to gauge its color accuracy and such things. I took the camera and pointed it at my father, and we saw his amused smile jump right up onto the screen. Next I pointed the camera at my own face and presto, there was I, up on the screen, replacing my father. But then, inevitably, I felt compelled to try pointing the camera at the TV screen itself.

Now comes the really curious fact, which I will forever remember with some degree of shame: I was *hesitant* to close the loop! Instead of just going ahead and doing it, I balked and timidly asked the salesperson for *permission* to do so. Now why on earth would I have done such a thing? Well, perhaps it will help if I relate how he replied to my request. What he said was this: "No, no, *no*! Don't do *that* — you'll break the camera!"

And how did I react to his sudden panic? With scorn? With laughter? Did I just go ahead and follow my whim anyway? No. The truth is, I wasn't quite sure of myself, and his panicky outburst reinforced my vague uneasiness, so I held my desire in check and didn't do it. Later, though, as we were driving home with our brand-new video camera, I reflected carefully on the matter, and I just couldn't see where in the world there would have been any danger to the system — either to the camera or to the TV — if I had closed the loop (though *a priori* either one of them would seem vulnerable to a meltdown). And so when we got home, I gingerly tried pointing the camera at the screen and, *mirabile dictu,* nothing terrible happened at all.

The danger I suppose one could fear is something analogous to audio feedback: perhaps one particular spot on the screen (the spot the camera is pointing straight at, of course) would grow brighter and brighter and brighter, and soon the screen would melt down right there. But why might this happen? As in audio feedback, it would have to come from some kind of amplification of the light's intensity; however, we know that video cameras are not designed to *amplify* an image in any way, but simply to *transmit* it to a different place. Just as I had figured out in the calm of the drive home, there is no danger at all in standard video feedback (by the way, I don't know when the term "video feedback" was invented, nor by whom; certainly I had never heard it back then). But danger or no danger, I remember well my hesitation at the store, and so I can easily imagine the salesperson's panic, irrational though it was. Feedback — making a system turn back or twist back on itself, thus forming some kind of mystically taboo loop — seems to be dangerous, seems to be tempting fate, perhaps even to be intrinsically *wrong,* whatever that might mean.

These are primal, irrational intuitions, and who knows where they come from. One might speculate that fear of any kind of feedback is just a simple, natural generalization from one's experience with audio feedback, but I somehow doubt that the explanation is that simple. We all know that some tribes are fearful of mirrors, many societies are suspicious of cameras, certain religions prohibit making drawings of people, and so forth. Making representations of one's own self is seen as suspicious, weird, and perhaps ultimately fatal. This suspicion of loops just runs in our human grain, it would seem. However, as with many daring activities such as hang-gliding or parachute jumping, some of us are powerfully drawn to it, while others are frightened to death by the mere thought of it.

God, Gödel, Umlauts, and Mystery

When I was fourteen years old, browsing in a bookstore, I stumbled upon a little paperback entitled "Gödel's Proof". I had no idea who this Gödel person was or what he (I'm sure I didn't think "he or she" at that early age and stage of my life) might have proven, but the idea of a whole book about just one mathematical proof — any mathematical proof — intrigued me. I must also confess that what doubtlessly added a dash of spice to the dish was the word "God" blatantly lurking inside "Gödel", as well as the mysterious-looking umlaut perched atop the center of "God". My brain's molecules, having been tickled in

the proper fashion, sent signals down to my arms and fingers, and accordingly I picked up the umlaut-decorated book, flipped through its pages, and saw tantalizing words like "meta-mathematics", "meta-language", and "undecidability". And then, to my delight, I saw that this book discussed paradoxical self-referential sentences like "I am lying" and more complicated cousins. I could see that whatever Gödel had proved wasn't focused on numbers *per se,* but on reasoning itself, and that, most amazingly, *numbers* were being put to use in reasoning about the nature of mathematics.

Although to some readers this next may sound implausible, I remember being particularly drawn in by a long footnote about the proper use of quotation marks to distinguish between use and mention. The authors — Ernest Nagel and James R. Newman — took the two sentences "Chicago is a populous city" and "Chicago is trisyllabic" and asserted that the former is true but the latter is false, explaining that if one wishes to talk about properties of a *word,* one must use its *name,* which is the expression resulting from putting it inside quotes. Thus, the sentence " 'Chicago' is trisyllabic" does not concern a city but its name, and states a truth. The authors went on to talk about the necessity of taking great care in making such distinctions inside formal reasoning, and pointed out that names themselves have names (made using quote marks), and so on, *ad infinitum.* So here was a book talking about how language can talk about itself talking about itself (etc.), and about how reasoning can reason about itself (etc.). I was hooked! I still didn't have a clue what Gödel's theorem was, but I knew I had to read this book. The molecules constituting the book had managed to get the molecules in my head to get the molecules in my hands to get the molecules in my wallet to… Well, you get the idea.

Savoring Circularity and Self-application

What seemed to me most magical, as I read through Nagel and Newman's compelling booklet, was the way in which mathematics seemed to be doubling back on itself, engulfing itself, twisting itself up inside itself. I had always been powerfully drawn to loopy phenomena of this sort. For instance, from early childhood, I had loved the idea of closing a cardboard box by tucking its four flaps over each other in a kind of "circular" fashion — A on top of B, B on top of C, C on top of D, and then D on top of A. Such grazing of paradoxicality enchanted and fascinated me.