PART I

BREAKDOWNS OF WILL: THE PUZZLE OF *AKRASIA*

CHAPTER 1

INTRODUCTION

There have been plenty of books and articles that describe how irrational we are – in consuming drugs and alcohol and cigarettes, in gambling, in forming destructive relationships, in failing to carry out our own plans, even in boring ourselves and procrastinating. The paradoxes of how people knowingly choose things they'll regret don't need rehashing. Examples of self-defeating behaviors abound. Theories about how this could be are almost as plentiful, with every discipline that studies the problem represented by several. However, the proliferation of theories in psychology, philosophy, economics, and the other behavioral sciences is best understood as a sign that no one has gotten to the heart of the matter.

These theories almost never mention failures of will.[1] This is just not a concept that behavioral scientists used much in the twentieth century. Some writers have even proposed that there's no such thing as a "will," that the word refers only to someone's disposition to choose. Still, the word crops up a lot in everyday speech, especially as part of "willpower," something that people still buy books to increase.

It's widely perceived that some factor transforms motivation from a simple reflection of the incentives we face to a process that is somehow *ours*, that perhaps even becomes *us* – some factor that lies at the very core of choice-making. We often refer to it as our will, the faculty by which we impose some overriding value of ours on the array of pressures and temptations that seem extrinsic. People usually ascribe control of temptation to the power of will and the unpredictability of this control to the freedom of will. Unfortunately, there has been no way to talk about such a faculty in the language of science, that is, in a way that relates it to simpler or better-understood elements. Without addressing this factor, science paints a stilted picture of human experience in general. However, quantitative motivational research has produced a distinctly new finding that promises to account for the phenomenon of will – with elements that are already familiar to behavioral science. That, in a sentence, is the topic of this book.

1.1 A BRIEF HISTORY OF SELF-DEFEATING BEHAVIOR

A lot has been said about the will since the classical Greeks wrote about why people don't – or shouldn't – follow their spontaneous inclinations. Plato quoted Socrates describing what can go wrong when people weigh their future options:

> Do not the same magnitudes appear larger to your sight when near, and smaller when at a distance? . . . Is not [the power of appearance] that deceiving art which makes us wander up and down and take the things at one time of which we repent at another? . . . Men err in their choice of pleasures and pains, that is, in their choice of good and evil, from defect of . . . that particular knowledge that is called measuring.

Aristotle gave this disorder a name, *akrasia,* "weakness of will."[2] Thus a human faculty, not called will until later, was defined by the situation in which it failed.

Normally, a person was said to follow "reason," to weigh her options in proportion to their real importance; but sometimes an option seemed to loom too large, a process called "passion." Passion was the enemy of reason. As this dichotomy evolved, it began to define a functional anatomy of the self. Reason was the major part of your real identity; passion was something that *came over you* – the term was often contrasted with "action," something you *do.*[3]

The self used reason to defend itself from passions and, if successful, developed a "disposition" to behave temperately. Reason and a temperate disposition were the good guys; passion and *akrasia* were the bad guys, perhaps the *other* guys. The Roman physician Galen said that their relationship was that of a man to an animal: "Irascible" passions could be tamed, but "concupiscible" passions (appetites, like sex and gluttony) were too wild and could be controlled only by starving them.[4]

The Judeo-Christian theological view of "weakness of the flesh" developed in parallel with the Greek rationalist one. A noteworthy difference was that the theological view made reason somewhat external to the self, and passion more internal. Reason was the word of God, and a function called will was, to a large extent, supplied by God's grace. Passion was sin, a relentless part of man's identity since Adam's fall; but passion was sometimes augmented by external possession in the form of demons. The self swayed between reason and passion, hoping, in its reflective moments at least, that God would win:

> I do not even acknowledge my own actions as mine, for what I do is not what I want to do, but what I detest. But if what I do is against my will, it means that I agree with the law and hold it to be admirable. But as things are, it is no longer I who perform the action, but sin that lodges in me . . . the good which I want to do, I fail to do; but what I do is the wrong which is against my will; and if what I do is against my will, clearly it is no longer I who am the agent, but sin that has its lodging in me. I discover this principle, then: that when I want to do the right, only the wrong is within my reach. In my inmost self I delight in the law of God, but perceive that there is in my bodily members a different law, fighting against the law that my reason approves and making me a prisoner under the law that is in my members, the law of sin.[5]

The assertion that the individual will had somewhat more power than this, and thus might not depend on the grace of God, was rejected as one of the great heresies, Pelagianism.[6]

Other philosophies and religions have all included major analyses of the passions. They also discuss how to avoid them. Buddhism, for instance, concerns itself with emancipation from "the bond of worldly passions" and describes five strategies of purification, essentially: having clear ideas, avoiding sensual desires by mind control, restricting objects to their natural uses, "endurance," and watching out for temptations in advance.[7] However, the ways that non-Western religions enumerate causes of and solutions to self-defeating behaviors seem a jumble from any operational viewpoint of trying to maximize a good.

Despite all the attention paid, not many really new ideas about self-control have appeared over the years, even in the great cultural exchanges that brought the whole world into communication. One significant advance was Francis Bacon's realization that reason didn't have its own force, but had to get its way by playing one passion against another: It had to

> set affection against affection and to master one by another: even as we use to hunt beast with beast. . . . For as in the government of states it is

> sometimes necessary to bridle one faction with another, so it is in the government within.[8]

The implication was that passion and reason might be just different patterns in the same system. Furthermore, they might be connected not by cognition but by some internal economic process, in which reason had to find the wherewithal to motivate its plans.

Another new idea was the Victorian discovery that the will could be analyzed into specific properties that might respond to strengthening exercises. We'll look at these in detail later (Section 5.1.4).

Even as some nineteenth-century authors were dissecting the will, others began to get suspicious of it. Observers had long known that the will could get bogged down in minutiae, a problem that medieval scholastics called a "scrupulous conscience."[9] In early Victorian times Soren Kierkegaard warned of a more general but insidious affliction that seemed to come from the very success of willpower in controlling passion – a loss of what the existential school of philosophy, Kierkegaard's heirs, came to call "authenticity." The existentialists said that authenticity comes from a responsiveness to the immediacy of experience, a responsiveness that is lost when people govern themselves according to preconceived "cognitive maps."[10]

At the turn of the twentieth century, Freud described a division of motivational processes into those that serve long-range goals (the "reality principle") and those that serve short-range ones (the "pleasure principle"). But the long-term processes are always distorted by an alien influence, "introjected" from parents, making them rigid. Freud rarely used the word "will," and used it trivially when he did; but his farsighted processes and the "superego" that made them rigid would have been recognizable to his audience as components of will and willpower.[11]

Interest in the will grew steadily until about the time of World War I. After that the concept of will suddenly became highly unfashionable, even distasteful – as if people blamed it for their countries' steadfastness in commanding millions of soldiers to face murderous fire and perhaps for the fortitude that led the soldiers to obey.[12] Whatever the reason, the twentieth century saw our concepts of impulsiveness and self-control become diffuse. We continued to analyze reason in the form of utility theory, which defined that perfect rationalist, Economic Man. Passion and *akrasia*, however, are another story entirely, as are any devices that might be needed to overcome them. Explanations of them are ad hoc and higgledy-piggledy.

Willpower had become a popular Victorian virtue without any examination of where it came from. When it became tainted there was no agreed-upon way to analyze what was wrong, or what alternatives there might be, or even precisely what function it was supposed to perform.

1.2. HOW TO STUDY SELF-DEFEATING BEHAVIOR

Something is obviously wrong or at least incomplete about the way we've understood *akrasia* and self-control. I believe that new findings make it possible to say a lot about the will and the reasons why it succeeds and fails where it does; but first, we have to look at what's already been said. Behavioral scientists still study weakness and strength of will, although usually without those specific concepts in their minds – sometimes without even the concept of motivation. But these scientists don't talk to most of their colleagues. Like so many fields where people are probing a mystery, decision science has split into schools whose members agree within their groups on certain assumptions and ways of doing research. Reading other schools' writings means forgoing the shorthand you've become adept at in your own school, not to mention the confidence that what you write yourself will have a willing audience. Mostly, we don't bother.

But these schools have separately discovered many different tools to work on the will problem. Before we start work, we need to look at the available methods. Here's an informal list of the schools that have studied will-related decisions:

Behaviorism is the school that has designed most of the systematic experiments on utility theory. The behaviorists have made especially good use of animal models. Lower animals are different from people, of course, but their subcortical brain structures are similar, including the systems that govern motivation, and this similarity is reflected in a similar response to most (but not all) schedules of reward. For instance, animals can become addicted to all the substances that affect people. Based on their ability to judge how rich different sources of reward are, animals often seem to be more rational than people.[13]

The neurologist Paul MacLean once observed that the human cortex rides on lower brain functions like a man riding a horse. Although we can't use animals to study some higher functions – wit, irony, or self-consciousness, for instance – we can use them to study the horse we all ride. And when a mental process can be demonstrated in animals – like a conflict between motives at successive times – it spares us speculation about subtle causes like quirks of culture.

However, the careful experiments that the behaviorists do have been overshadowed by their righteousness about method. To the average educated person, a behaviorist is somebody who believes that the mind doesn't exist, and that people's behavior can be accounted for entirely by the observable stimuli that impinge on them. Even the academic community tired of this brand of logical positivism and stripped the behavioral school of most of its glory. As a source of carefully controlled data, however, it remains unsurpassed, and its data are the starting place of this book.

Cognitive psychology, often as applied to social psychology, is currently the most widespread approach to both research and theory dealing with irrational behavior. It generally has high standards of experimental proof and has described many examples of maladaptive behavior. However, its theorists seem to have gone out of their way to avoid dealing with the process of motivation, seeing it as at most some kind of internal communication that a higher judge – the irreducible person – can and often should disregard. Thus its theories of irrationality have been restricted to finding errors of perception or logic.

Economics is the field that deals with rational decision-making in the real world. In modern times it has embraced the assumptions of utility theory, as characterized by Paul Samuelson: "The view that consumers maximize utility is not merely a law of economics, it is a law of logic itself." Gary Becker showed that economic concepts could handle even nonmonetary incentives like drug highs and the risk of jail.[14]

However, economists have made some unrealistic assumptions about decisions: that they're all deliberate, that they're based only on external goods (as opposed to rewards that you might generate in your own head), and that they're naturally stable in the absence of new information. Since this stability should make decisions consistent, economic theories have attributed irrationality only to inadequate information or steep discounting of the future, explanations that are both inadequate, as we'll see.

Philosophy of mind looks at model-making itself, and has pioneered thought experiments whereby every reader can test a particular theory.[15]

However, it has stayed within the conventional assumptions of a unitary self – unitary in the sense of not housing contradictory or unconscious elements. If anything should allow exploration of a more molecular model of the self, it should be thought experiments; but the seeming paradoxes that some have demonstrated have not led analysis beyond standard utility theory. They remain paradoxical.

Psychoanalysis was the first major attempt to confront self-contradictory behaviors with utility analysis. As an explorer of scientifically virgin territory, Freud sketched out several different models – one based on motivation ("libido"), one based on consciousness, one based on organization ("id," "ego," "superego"), and so on. But he didn't work out how the various models got along with each other. Without the discipline of either controlled observation or conceptual parsimony, psychoanalysis grew overinclusive, until it resembled the polytheisms from which it drew some of its observations.

Oversold in the middle third of the twentieth century, psychoanalysis has lately been the target of vigorous attacks aimed at its standards of observation and proof. The essayist Frederick Crews concluded that

> the designer of psychoanalysis was at bottom a visionary but endlessly calculating artist, engaged in casting himself as the hero of a multivolume fictional opus that is part epic, part detective story, and part satire on human self-interestedness and animality.[16]

It hasn't been fashionable to ask whether even a fictional opus that once had such immense popularity among intelligent people may offer insights worth keeping.

Actually, Freud brought together a lot of previous work that describes disunity of the self, and this has gone into limbo with him.[17] Worse, people who have found his answers wrong or incomplete have stopped asking his questions, and these questions have to be in the forefront of any attempt to explain impulsiveness and impulse control: Is all behavior motivated? How can someone obey internally contradictory motives? How can you hide information from yourself? How can self-control sometimes make you worse off? On many questions I'll start with Freud's ideas – because, in my view, after modern criticism tackled the ball carrier, no one ever picked up the ball.

Bargaining research, a new discipline, has used elementary games to see how small groups of competing agents can reach stable relationships. It is especially suggestive when it shows how such a group can reach stable decisions that are not in all or any member's best interest. However, until now, bargaining research has not seemed applicable to conflict within the individual because of the supposed unity of the person. Given a rationale for disunity, we'll find it useful.

Chaos theory, an even newer theory of analysis, has been applied to other subjects – the weather, for instance – to explore how outcomes may depend on a recursive feedback system. It has also shown how such a system may lead to similar patterns at different levels of magnification and even to the growth of the different levels themselves. So far chaos theory has lacked any important motivational example. However, the fundamental unpredictability of the human will, which has defied attempts to explain it by antecedent causes, makes it look like some of the natural phenomena where the chaos approach has proven useful. As we find recursive processes in the will, chaos theory will become relevant.

Sociobiology has studied competition among populations of reward-seeking organisms, so it has developed concepts that might be useful for populations of behaviors – the range of behaviors that an organism tries out – as well. Behaviorists have proposed that reinforcement acts on behaviors the way natural selective factors act on organisms.[18] This suggests some way that sociobiological theory might apply to conflicting motives.[19]

Neurophysiology has produced increasingly precise findings on brain mechanisms, including those that create motivation. It's possible to see, for instance, exactly where and by what neurotransmitters cocaine rewards the behaviors that obtain it;[20] but pinpointing the transmitters doesn't explain how a conflict between alternative rewards gets resolved or why it fails to get resolved in some cases. It may be, for instance, that some alcoholics have inherited settings in their reward mechanisms that make alcohol more rewarding for them than for most people; but this doesn't tell us why many alcoholics are conflicted about their drinking – why they often decide not to drink despite the intensity of this reward and, having decided this, why they sometimes fail to carry out their own decision. Neurobiology will be useful here mainly as a check on reality, as a body of findings with which any motivational theory must at least be consistent.

Theology shouldn't be disregarded. It has studied a part of our decision-making experience that seems to lie outside the will and has been least influenced by the lure of utility theory. Despite its own theory that its insights come mystically, by faith, revelation, or some such nonempirical route, theology actually demands that its tenets ring true to experience. Sin, for instance, seems synonymous with the self-defeating behaviors that the more scientific disciplines have talked about; the debates that have occurred over the power of the individual will to overcome sin have appealed to what is, in effect, clinical experience. But what this inspirational approach has gained in sensitivity it has lost in testability, and it becomes arbitrary when it tries to nail down its insights in a systematic way. Like psychoanalysis, it will be a source more of questions than of answers. But the questions are important ones.

Finally, any explanation of *akrasia* has to be at least compatible with *subjective experience* and might well find evidence there. Some behavioral scientists sniff at experiential evidence as "folk psychology" and warn of the days when psychologists tried to gather data using trained introspectors. While common sense is suggestible at best and, as theory, almost always inconsistent and ad hoc, it is by far the largest body of human observations. Useful samples of common experience appear in the writings of the preexperimental (Victorian) psychologists and of later clinicians who have interviewed patients, as well as in those works of fiction that have rung true with generations of readers. Jon Elster has been especially insightful in sorting the pieces of our written heritage by their motivational implications.[21]

1.2.1 My Approach to the Problem

So how should we assemble a working tool kit from all of these methods? I'll suggest one way, obviously not the only one possible. But as far as I can tell, it's the only proposal so far that reconciles the familiar paradoxes of motivation with basic research.

I warn the reader in advance that this approach is *reductionistic.* That is, I assume that every change in thinking, feeling, wanting, planning, and so on, has a physical basis in the nerve cells of the brain, which in turn depend on chemical changes within the cells, and so on. I'm not saying that thinking and feeling are best studied by studying the chemistry of cells – only that all explanations of behavior should at least be consistent with what's known in the physical and biological sciences.

Nonreductionistic (and antireductionistic) theories have been created for a reason, of course. In the past, reductionistic theories ignored causes that were hard to observe or to imagine – that is, too hidden or complex