CHAPTER 1

INTRODUCTION
What Are Models For?

Wherein the aims and goals of ascriptive science—the discipline of making valid ascriptions of mental states to humans on the basis of observing their behavior—are introduced and illuminated against the current intellectual landscape of normative, descriptive, and prescriptive approaches to the human sciences and against the background of strands of hermeneutics, the philosophy of language, and the theory of rational choice inherited from a century and a half of writing and thinking about the subject.

WHAT FOLLOWS is an attempt to build a rigorous science of human action from a repertoire of moves and conceptual structures provided by decision and rational choice theory, classical epistemology, artificial intelligence, the philosophy of language, and the experimental methods and results of cognitive and social psychology. The resulting science expressly and explicitly distances the modeler of human behavior and thought from the subject of his or her models. In building this science I make deliberate and frequent use of mathematical and logical models of thinking, believing, emoting, and behaving, and I use them to the explicit end of helping modelers achieve distance from the behavior being modeled with the aim of increasing the precision with which one can represent what and how humans do, think, and feel.

By abducting the essence of the models away from the often lukewarm and fuzzy innards of common English language usage, I hope to accomplish two objectives. The first is to increase the precision with which we can formulate propositions about thinking and behavior and design tests of those propositions. Mathematical representations and first-order logic help greatly with the project of turning quality into quantity or scale, which is important for ascertaining progress in a field of inquiry—even if not always for progress itself: "Better" is made more precise when it is interpreted as "more accurate" or "more valid." If we want "better" models and can agree on measures of validity and accuracy, then we will be able to know and tell which way we are going when we model. It is one thing to say that John is "poor at reining in his appetite for whipped cream" but quite another to characterize the rate at which he trades off a certain amount of guaranteed whipped cream consumption for other entities he values—including the value he derives from the sustained validity of his self-concept as a being capable of self-control—as a function of various visceral states (such as satiation, hypoglycemia, level of sexual arousal) and of various prototypical social situations in which he finds himself (work-related meeting with bosses, work-related meeting with subordinates, stroll with friends). By showing the payoff that decision science brings to action science in terms of precision, I hope to exculpate the former from the often valid accusation that it is a "toy science"—a banal endeavor that is good enough at making all-things-considered ensemble predictions about the behavior of consumers but that cannot and should not be deployed when things begin to matter: in *this* particular case, in the high-stakes decision scenario, in the one-off interaction that "does one or foredoes one quite." I want, then, to create a *decision-science-for-when-it-matters*.

The second objective is to capitalize on the distancing effect that is produced when we talk about people as agents or decision agents (or TOTREPs, "trade-off-talking rational economic persons" [Kreps, 1988]) and attempt to measure various quantities that are relevant to our models of these agents—just as when, in a study of "animal learning," we would measure the proclivity of a rat in a maze to exhibit a modification of its behavior in response to a repeated set of pain/reward-mediated stimuli. Indeed, the distancing effects that characterize the modeling approach of traditional rational choice theory and the experimental approach of pre– and post–Cognitive Revolution psychology are, I predict, among the most valuable contributions these fields will be deemed by future historians of ideas to have made to the understanding of human behavior. Universal models—such as those provided by rational choice and decision theory—will be used to create an emotional *distoscope*, which functions (conversely to what one would expect of an emotional *micro*scope) to produce emotional distance between the modeler and the "modelee"—a move that is particularly helpful when those we wish to model are either ourselves or other "emotionally close" individuals. Thereby, "action science" will become more science-like even as it remains focused on action.

Achieving these goals hinges delicately on what I mean by "models" and what I intend to do with them; delicately because, if misunderstood, the new action science I am aiming for quickly becomes another "discipline"—which I would consider an unfortunate outcome—rather than "a way of living" for those interested in the competent prediction and intelligent production of behavior. So, on to *models*, then, and their uses.

1. WHAT MODELS ARE FOR: REPRESENTATIONAL AND PERFORMATIVE DIMENSIONS

We have inherited the following picture of models in science: They are representations of "reality," of behavior or thought, that can be used to take us from a set of observable or known quantities or variables (past choices, past measured features of thinking) to a set of predictions of future—or otherwise unobserved, and thus unknown—quantities or variables (future behavior, hidden and private features of thinking). Models embed within them algorithms and formulas for predicting the evolution of observables. Thus, a simple answer can be given to the question of "why model?": *to make inferences about what we do not know on the basis of what we do know*. If I observe Mathilda choose white bread from a bread-stand that offers both white and wheat bread, then I can model Mathilda as an individual who instantiates—through her behavior—a set of preferences (of white over wheat bread, in this example) and use it to infer that she will choose white bread over wheat bread the next time she has a choice between these two options. The *representation* of Mathilda's behavior as the instantiation of a set of preferences (which are hidden to me, the modeler) is critical to the model of Mathilda used to predict her future behavior; and it is the representation that we use to *model* Mathilda that constrains the kinds of questions we can hope to answer by making use of the model. Had we modeled Mathilda as an automaton that reliably produces certain kinds of behavior ("buy a loaf of bread and devour it") in response to a particular stimulus ("abusive behavior by her lover"), then we could make no prediction about Mathilda's selection of one kind of bread over another but we could still make predictions about Mathilda's behavior after certain life events.

The link between the nature and structure of models and the nature of the questions that we can pose on the basis of those models suggests that the representational function of models is incomplete and, in fact, unreasonably *benign*. In particular, the modeler can (A) *create* models that answer certain kinds of questions that she is interested in, (B) *interact* with the objects of her models, and even (C) *force them* into answering certain kinds of questions that are based on a particular representation or (D) induce them to *interpret themselves* through the lens of the proposed model, thereby altering their behavior.

Example of (A). I am interested in understanding how you act when you are simultaneously faced with (a) an impulse to produce destructive behavior (an unrestrained temper tantrum) in the context of (b) a situation in which acting on such an impulse has a high social cost to you (loss of face and reputation). I create a model of you in which the two conflicting impulses or motives (for self-expression and self-censure) can be co-present and capable of interacting, and I then posit various interaction mechanisms (winner-take-all, compromise, competitive equilibrium) between the impulses that can be used to explain certain kinds of observable behavior (temporary loss of attention or of eloquence, sudden lapses into sullenness) that you may produce when I intentionally (but covertly or deniably) behave so as to irritate you in a public setting that is meaningful to you.

Example of (B). The question, "Why did you choose to arrive five minutes late to this meeting?" forces upon the subject a particular model of behavior and thought (conscious, knowing, and intentional choice of a set of actions that have led to the tardiness) that has a *lensing* effect: The representation becomes a lens through which behavior is seen for the purpose of the interaction. Of course, the subject may reject the representation of his person and behavior implicit in the question (*Because the elevator took a long time to arrive*), but the lens can often be reestablished (*Why did you* choose *not to leave a few minutes earlier in order to arrive on time, given that you knew or had reason to know there was some uncertainty in the timeliness of the elevator service?*).

In each case, models emerge as interventional devices and not merely representational ones. They are used to *intervene by representing*, and these two functions are closely interconnected. In what follows I will draw on the significant representational power of models generated in various branches of decision theory (from the phenomenological through the economic and the psychological) and focus on the *performative dimension* of these models, comprising that which one does to oneself and to the modeled by the act of modeling. This move commits me to a view of the science of human behavior that is based on an ongoing interaction between modelers and modelees. What does this science look like?

2. THE NEW ACTION SCIENCE OF HUMAN WAYS-OF-BEING: FROM NORMATIVE-DESCRIPTIVE-PRESCRIPTIVE SCIENCE TO ASCRIPTIVE SCIENCE

We are accustomed to following Howard Raiffa's distinctions and to speak about descriptive (what is the case?), normative (what should be the case?), and prescriptive (what should be done, given what we know to be the case?) approaches to the science of behavior and thought. Thus, *descriptive* behavioral science informs us of how people *do* respond to certain stimuli; *normative* rational choice theory tells us about what a logically and informationally omniscient decision maker *would* do *if* she had certain preferences; and *prescriptive* behavioral decision theory tells us how people *should* try to think and act, given what we know or what we think they should know about their own and others' deviations from the model of the ideal decision maker posited by rational choice theory. In each of the three cases, felicitous (and often unwitting) use is made of two elements: the *ideal type* and the *statistical ensemble* (or the "average man," if you believe in turning statistics into probabilities). The ideal type is to normative science what the average man is to descriptive and prescriptive science: a stylized, tractable reduction of the all-important ideal agent that enables simple inferences from the unknown to the known. If Gary models Pauline as the omniscient, coherent ideal type of rational choice theory and deduces Pauline's preferences from her observed choice behavior, then he is on his way to building a model that can be used to predict Pauline's future behavior. If Amelie models Gary as an "average man" whose patterns of thinking and behavior—based on tests performed by Amelie and her friend Dan in their labs—differ from those legislated by the ideal model in reliable ways, then she can produce predictions of Gary's future behavior based on the same kind of inferential device that Gary used to make predictions about Pauline. Thus proceeds the conventional approach to the science of human behavior, which resonates greatly with the representational use of models.

Now enters Chris, who cares not only about understanding what Pauline will do *given* that she is a rational or super-rational being but also about whether or not Pauline (*this* individual *here and now*) *is*, in fact, a super-rational being; and not only what Gary will do given that he is *an average guy* (as defined by Dan and Amelie's tests) but also whether or not Gary *is* actually an average guy in the precise Amelie-and-Dan sense of averageness. Chris is in a position familiar to most humans (including Gary, Amelie, and Dan themselves, when they leave their offices), who must deal with individuals here and now and whose here-and-now predictions matter (and matter a lot) *here* and *now*. Chris no longer has the luxury of being able to "explain away" a particular behavior either as "irrational" (an explanation that is available to normative and prescriptive theorists) or as "noise" (available to descriptive and prescriptive theorists).

In exchange for the additional complication, Chris has the opportunity to *interact* with the subject of his models; thus his science taps into the performative dimension of models. Of course, Chris never quite leaves the representational dimension of models. He starts out with his own models of Pauline (perhaps the same as Gary's) and of Gary (perhaps those of Amelie and Dan). In fact, Chris needs a certain minimal model of Gary just to begin talking to him (for, in talking, he is *saying* and in saying he makes truth claims and in so doing he assumes that Gary understands them, which means that Gary has the sophisticated and never-before-seen-in-the-animal-kingdom capacity called "linguistic understanding," and so forth). Yet Chris understands that the models themselves shape the interaction: They supply or constrain the questions that Chris can ask and also co-opt certain kinds of behavior from Pauline and Gary and, in turn, from Chris himself. The models are causally influential within the interaction and can lead to the production of novel behaviors from all three protagonists.

Such an interaction places Chris's science outside of the usual space spanned by the descriptive, normative, and prescriptive dimensions of the "normal science" of humans. It is a different kind of science, which another Chris (unrelated to our hero) calls "action science" [Argyris, 1993] and I shall call *ascriptive science*—not because I feel the need to be different (as evidenced by my appropriation of this term from Itzhak Gilboa [1991]) but rather because I require a meaning that is different from the one that "action science" has taken on through repeated usage. "Ascriptive" science, as the name suggests, is the science of making valid ascriptions: of ascribing entities (such as rationality and intentionality) to oneself and others on the basis of structured interactions that aim both to discover and to educate.

Understanding ascriptive science is most easily accomplished by asking how it overlaps and interacts with normative, descriptive, and prescriptive science. In order to do this, it is important to understand not only the different kinds of science of behavior but also the predispositions, activities, and dominant concerns of the scientists who practice them. For this purpose I draw on Ian Hacking's [1983] categorization of natural scientists as speculators, calculators, or experimenters. Hacking argues that natural science "comes together" as the result of the conjoint efforts of the activities of three kinds of people: Speculators (e.g., Galileo, Newton, Einstein) generate new distinctions and introduce new concepts that are useful in describing, manipulating, and/or creating phenomena; calculators (Laplace, Penrose, Thorne) use the basic schemata articulated by speculators and perform the necessary logical and computational work that takes us from concepts to testable models and useful algorithms; and experimenters (Kepler, Eddington, Penzias) create effects in the laboratory and the field that are based on distinctions articulated by the speculators and sharpened by the calculators—"effects" that historians of science, as well as philosophers of science bent upon rational reconstructions of scientific activity, then portray as being "tests" of one theory (or model) or another. Seen through this behavioral categorization pattern, normative scientists speculate and calculate (Bayes, Ramsey, de Finetti, Schelling, Nash, Aumann); descriptive scientists calculate and experiment (Allais, Ellsberg, Tversky); and prescriptive scientists do some of each (in addition to pontificating, which is a function that I shall ignore). The ascriptive scientist draws on the core skills of each of the speculators, calculators, and experimenters who have conspired to engage in normative, descriptive, and prescriptive behavioral science. He speculates insofar as he uses and refines the basic distinctions made by normative science (rationality/irrationality, knowledge/belief/ignorance) to put together structured models of behavior and thought, which he uses calculatively to make predictions about a subject's behavior and thinking, which he tests by designing real-time experiments whose results he then uses to modify or fortify his speculative models of the subject's behavior and thought. But unlike the normative, descriptive, or prescriptive scientist, the ascriptive scientist is guided in his inquiry and behavior by pragmatic questions (*How can I induce X to do Y? What will X do or say if I do or say Y?*) that are typical of everyday human interactions and ways-of-being-toward-one-another of humans rather than by the questions that concern typical behavioral scientists vis-à-vis their subjects (*How can effect X be instantiated at level of reliability Y in subject population S given temporal and material budget constraint C?*).

The claim that I shall defend throughout the book is that, in spite of the significant difference in the types of questions that preoccupy ascriptive and "normal" scientists, there is great value to the ascriptive scientist in appropriating both the discipline and the skill sets of the speculators, calculators, and experimenters whose work has produced the normative, descriptive, and prescriptive human sciences. "Shall" entails that this defense is "about to happen," so the reader is asked to withhold judgment on my claim until a substantial part of this "ascriptive" science has been developed.

2.1. *Precursors I: Oneself as Another—Nietzsche's Overman*

Perhaps the earliest precursor of *ascriptive* science appears in the documented thought of Friedrich Nietzsche. In *Also Sprach Zarathustra* [Thus Spake Zarathustra; Nietzsche, 1892/2007] he introduced the concept of the *Ubermensch*—the Overman, sometimes inappropriately (notably by Goebbels) translated as the Superman—to represent the individual who can hold himself out as an object of inquiry, analysis, and experimentation. The distinguishing feature of the Overman is the ability to distance himself from his own raw feels, emotions, perceptions, thoughts, intentions, and other first-person internal psychological states to the point where these states (and the attending behaviors) can be beheld *as if they were another's*. What is gained thereby—the argument goes, and I also contend—is a capability for a more accurate representation of one's own states of mind and body and the relationships between internal psychological states and behaviors. If one can model oneself while maintaining the same state of mind as when one models a light switch or a rat in a maze, then one can more accurately and validly see one's own behavior and intervene to ameliorate that behavior—along with the thoughts and feelings associated with it—by using the same representation-guided tinkering used to repair the light switch or to change the path of the rat; and one can also hope to produce both internal states and external behaviors in ways akin to those by which one designs different mazes and reward conditions to influence an experimental rat's behavior. The act (and attendant benefits) of such dispassionate self-beholding is, of course, far more easily recognized than produced, and the intention of ascriptive science is to facilitate the *production* of the states that characterize the Overman. On this view, the Overman is neither a personality type nor a character in the "hard-wired" sense of those terms; rather, it is a state of being that can be achieved through training.

Of what could this training consist? The argument I shall develop claims that the basic tool kit of the speculators, experimenters, and calculators that produced the contemporary human sciences circa 2010 CE can be interpreted in a way that provides a valuable set of distancing mechanisms whose value ranges far beyond the purely descriptive use to which they have been put so far. Understanding oneself (or a close friend or coworker) as an agent equipped with a computationally potent (but not omnipotent) reasoning faculty aimed at producing optimal or maximal outcomes based on informationally "scient" (but not omniscient) perceptual inputs allows one to relate the mental and verbal behavior of the "creature" being modeled in a way that is far closer to how one understands a light switch or the behavior of a rat in a maze than to the everyday (fuzzy and forgiving) ways in which humans usually "give accounts" and "produce narratives" about themselves and others.