MATHEMATICAL MODELS OF MEANING

A Dynamic Systems Approach to Possible World Semiotics

Paul Kockelman

The MIT Press
Cambridge, Massachusetts
London, England

The MIT Press
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02139
mitpress.mit.edu

This book was set in Times New Roman by Paul Kockelman. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Kockelman, Paul, author.
Title: Mathematical models of meaning : a dynamic systems approach to possible world semiotics / Paul Kockelman.
Description: Cambridge, Massachusetts : The MIT Press, 2025. | Includes bibliographical references and index.
Identifiers: LCCN 2024038530 (print) | LCCN 2024038531 (ebook) | ISBN 9780262552684 (paperback) | ISBN 9780262383486 (pdf) | ISBN 9780262383493 (epub)
Subjects: LCSH: Semiotics. | Mathematical linguistics. | System theory.
Classification: LCC P99.4.M63 K63 2025 (print) | LCC P99.4.M63 (ebook) | DDC 410.1/51–dc23/eng/20241214
LC record available at https://lccn.loc.gov/2024038530
LC ebook record available at https://lccn.loc.gov/2024038531

10 9 8 7 6 5 4 3 2 1

EU product safety and compliance information contact is: mitp-eu-gpsr@mit.edu

For Mia, Zeno, and Lara

Contents

List of Figures xi
Preface xv
Acknowledgments xvii

1 Introduction 1
1.1 The Core Components 1
1.2 The Main Equation 6
1.3 Agents That Evolve and Learn 10
1.4 Signers, Interpreters, and Signals 12
1.5 Meaning, Information, and Value 14
1.6 Overview of the Chapters 20
1.7 Relevant Literature 24

I AGENTS THAT THINK

2 Symptoms and Sickness 31
2.1 Grounding the Scenario 31
2.2 Finding the Posteriors 35
2.3 The Value of the Interpretants 38
2.4 Calculating the Critical Price 41
2.5 The Value of a Sign 43
2.6 The Information in a Sign 46
2.7 Better and Worse Grounds 50

3 Predators and Prey 57
3.1 The Grounds of Predation 57
3.2 From the Perspective of the Prey 64
3.3 Predation as Conversation 65
3.4 Two-Dimensional Dynamics 69
3.5 Landscape as Ally and Antagonist 72
3.6 From Hyperbolic Valley to Rolling Hills 76

II AGENTS THAT EVOLVE

4 Biosemiotic Agents 83
4.1 Evolution and the Difference Equation 83
4.2 The Relative Fitness of Biosemiotic Agents 85
4.3 Dynamic Organism, Fixed Environment 87
4.4 Dynamic Organism, Dynamic Environment 91
4.5 Semiosis and Symbiosis 99
4.6 Meaning in Biosemiotic Systems 112

5 Hawks, Doves, and Mutants 115
5.1 The Classic Scenario 115
5.2 Biosemiotic Mutations 120
5.3 Hawklike Mutants 126
5.4 Dovelike Mutants 130
5.5 The Cost of a Semiotic Capacity 132
5.6 The Fixed Points of Costly Agents 134

III AGENTS THAT LEARN

6 Reinforcement Learning 141
6.1 The Basic System 141
6.2 From Individual to Ensemble 146
6.3 Establishing a Convention 148
6.4 Motivating Conventions 154
6.5 Parasites 158
6.6 Niche Disruption 164

7 Machine Semiosis 169
7.1 Machine Learning 169
7.2 Neural Networks 171
7.3 Establishing a Code 173
7.4 The Time It Takes to Establish a Code 180
7.5 Machine Learning as Meta-Semiosis 184
7.6 Language Models 187

IV PRESUPPOSITIONS AND EXTENSIONS

8 Possible World Semiotics 191
8.1 Introduction to Possible Worlds 191
8.2 Possible World Semantics 194
8.3 Probabilities of Possible Worlds 196
8.4 Bayesian Networks and Expected Value 199
8.5 Conversational Backgrounds 202
8.6 Worlds, Times, and Scales 206
8.7 Cutting the Universe Down to Size 208

9 Meta-Semiotic Processes 211
9.1 Variations on the Main Equation 211
9.2 Fluid Grounds 216
9.3 Self-Updating Semiotic Agents 219
9.4 Modeling Another Agent's Model 223
9.5 Culture as Shared Interpretive Grounds 229
9.6 Enemies, Parasites, and Infrastructure 232

10 Conclusion 235
10.1 Synopsis of Arguments 235
10.2 Revelation and Confrontation 236

A Energy and Entropy, Information and Value 237

B Complexity, Organization, and Constraint 247

C The Utility of a Sign 255

References 259
Index 265

List of Figures

1.1	Signs, Objects, and Interpretants	2
1.2	Relations between Relations	4
1.3	Signaling Agents	13
1.4	Meaning of Signs as Change in the Probabilities of Objects	16
1.5	Object and Sign as Sets of Possible Worlds	16
2.1	Posterior Probabilities	37
2.2	Meaning as the Change in Probability of an Object, Given the Sign	39
2.3	Conditional Values	41
2.4	Critical Price Versus Prior Probability	42
2.5	The Value of a Sign (in Two Dimensions)	45
2.6	The Value of a Sign (in Three Dimensions)	46
2.7	The Informativeness of a Sign	49
3.1	Predator's Evaluative Landscape (in One Dimension)	63
3.2	Prey's Evaluative Landscape (in One Dimension)	65
3.3	World Lines of Predator and Prey (in One Dimension)	68
3.4	World Lines of One Predator and Two Prey (in One Dimension)	69
3.5	The Evaluative Landscape of a Predator (in Two Dimensions)	72
3.6	The Evaluative Landscape of a Prey (in Two Dimensions)	73
3.7	World Lines of Predator-Prey Interactions (in Two Dimensions)	74
3.8	World Lines of Predator and Prey in a Hyperbolic Valley	77
3.9	World Lines of Predator and Prey in Rolling Hills	79
4.1	Frequency of Alleles over Generations	90
4.2	Evolution of System ($\alpha \geq a_1, a_2$ and $\beta \geq b_1, b_2$)	97

4.3 Phase Portrait 98
4.4 Key Features of Phase Portraits 106
4.5 Four Phase Portraits Compared 108
4.6 Attractiveness of Language 1 110
4.7 Behavior near Kinks 112
5.1 Temporal Dynamics of the Classic Hawk-Dove Scenario 119
5.2 Fixed Points of the Classic Hawk-Dove Scenario 120
5.3 Fixed Points of Hawklike Mutants 128
5.4 Fixed Points of Dovelike Mutants 132
5.5 Fixed Points of Costly Hawklike Mutants 135
5.6 Fixed Points of Costly Dovelike Mutants 136
6.1 Expected Value of Individuals over Time 146
6.2 Expected Value of Ensemble over Time 148
6.3 Establishing a Code 151
6.4 Converging on a Common Code 153
6.5 Predictable Objects 154
6.6 Relative Predictability 155
6.7 Motivated Conventions and Time-to-Code 158
6.8 Motivated Conventions and Costly Signs 159
6.9 Expected Values, $f < .5$ 164
6.10 Semiotic Strategies, $f < .5$ 165
6.11 Expected Values, $f > .5$ 166
6.12 Semiotic Strategies, $f > .5$ 167
7.1 A Simple Model of a Single Neuron 171
7.2 The Sigmoid Function, $\sigma(\cdot)$ 172
7.3 A Neural Network with Two Hidden Layers 173
7.4 Signer and Interpreter as Coupled Neural Networks 174
7.5 Signer and Interpreter Establishing a Code with N Symbols 180
7.6 Forward Propagation as a Semiotic Process 185
7.7 Backpropagation as a Meta-Semiotic Process 186
7.8 Next-Word Prediction as a Semiotic Process 188
8.1 Causal Graph (with Salient Extensions) 200

8.2 Proposition p is simple f-necessary in w, where $f(w) = \{p_1, p_2, p_3\}$ 203
8.3 Proposition p is simple f-possible in w, where $f(w) = \{p_1, p_2, p_3\}$ 203
9.1 Incorporated and Incorporating Diagrams 212
9.2 Self-Reflexive Semiotic Agent 217
9.3 Modeling Another Agent's Model 224
9.4 Enemies, Parasites, and Infrastructure 232
A.1 A Simple Thermodynamic System 238
A.2 The Boltzmann Distribution 239
A.3 Entropy of the At-Equilibrium System 240
A.4 Energy and Free Energy of the At-Equilibrium System 241
A.5 The Information of a Sign 244
A.6 The Value of a Sign 245
B.1 Organization of the Universe as a Function of p 251
B.2 Valley of Indecision 253
C.1 The Utility of a Sign 256
C.2 The Value of a Sign (Revisited) 257
C.3 Convergence of m to $\mathcal{V}(S)$ 258

Preface

This book offers a mathematical model of meaning, and thereby provides answers to the following kinds of questions: What is meaning? What is the relation between meaning, information, value, and purpose? What ingredients are necessary for a system to exhibit meaning? What behaviors, and capacities for behavior, are particular to meaning-oriented agents? Is there a relatively simple mathematical model that can adequately capture the dynamics—and diversity—of meaning-oriented agents? How do we best bridge the divide between interpretive paradigms that are qualitative and context-rich and formal methods that are quantitative and domain general?

At the center of this model is a distributed agent that can sense and instigate relatively immediate events and, through these, project and effect relatively mediate events, in reference to a dynamic set of commitments and values (understood as an interpretive ground), and by means of a double integration over past and future worlds.

This book argues that interpretive grounds are central to meaningful processes. It argues that such grounds can be embedded in environments no less than enminded in organisms, and hence turn on relatively objective patterns and resources no less than relatively subjective commitments and values. And it shows that such grounds function as dynamic variables: at once shaped by meaningful processes and shaping of meaningful processes. As will be seen, such a dynamic coupling between figures and grounds, qua interactional practices and interpretive resources, makes this mathematical model of meaning particularly rich and revealing.

In offering such an analysis, this book brings together the objects of signs and the ends of agents, and hence motivation as much as meaning. It connects agents that can select (insofar as they can choose different courses of action in real time) and agents that are selected (such that they can evolve over generational time). It accounts for the behavior of agents that are oriented to diverse kinds of value: from expected utility to free energy, from biological fitness

to social status. It shows the connection between the possible worlds of formal semantics and the microstates of statistical physics. And it puts the fixed points of dynamic systems theory into relation with the hermeneutic circles of critical social theory.

While the model incorporates core ideas from a pragmatist tradition, it weaves together a range of powerful ideas from other paradigms, including Bayesian inference, statistical mechanics, decision theory, mathematical biology, evolutionary game theory, possible world semantics, machine learning, linguistics, and anthropology. Its analytic framework thereby provides a relatively seamless integration of distinct methods and theories.

After introducing the model, and reviewing its core assumptions, chapters 2 through 9 explore the entailments of the model, and assess its merits, by using it to analyze a variety of increasingly complex scenarios. As will be seen, the math is done in a complete, but conversational way. And the formalism begins simply and ramps up slowly, such that a wide range of readers will be able to understand the concepts, follow the arguments, imagine novel scenarios, and extend the analysis themselves.

Acknowledgments

Greg Urban and Liam Taylor read a draft of this manuscript and gave me incredibly helpful suggestions. Philip Laughlin, my editor at MIT Press, was also very supportive. Jitendra Kumar, the senior project manager, was extremely helpful and patient. Thank you.

A range of friends and interlocutors transformed my thinking about various parts of this argument. Many thanks to members of the Semiotics Working Group at the Center for Transcultural Studies, especially Benjamin Lee, Terra Edwards, Andrew Carruthers, Katie Stewart, Miyako Inoue, Bill Hanks, Greg Urban, Xochitl Marsilli-Vargas, and Kamala Russell. I also found conversations with Siri Lamoureaux, Robert Meister, Mike Cepek, Nick Enfield, Gary Tomlinson, Jonathan Beller, and Julia Elyachar to be very stimulating.

Elijan Mastnak radically altered my relation to Arch Linux and LaTeX.

Support for this project was provided by the MacMillan Center for International and Area Studies at Yale University. Time for writing was provided by a generous research leave from Yale University. Many thanks to the Department of Anthropology, with a loud shout out to Doug Rogers, Erik Harms, and Marleen Cullen.

Special thanks to Erik Thogersen, Jessica Strick, Zoe Zoe Zoe, and everyone that Grannie Valerie brought into being.

1 Introduction

Overview

Section 1.1 summarizes some relevant ideas from Charles Sanders Peirce's theory of meaning, and highlights the ways that this book both builds on them and goes beyond them. It describes the core components of this model of meaning in a non-mathematical way. Section 1.2 projects a mathematical framework onto these components. It motivates and explicates the main equation of the model and highlights some of its key assumptions. Section 1.3 shows how the main equation can be applied to agents capable of evolving on phylogenetic time scales, as well as agents capable of learning on interactional time scales. Section 1.4 focuses on signaling processes and the division of semiotic labor. Section 1.5 distinguishes between the meaning, value, and information of a sign, and offers a characteristic measure for each of these three properties. Section 1.6 summarizes the contents and arguments of the chapters that follow, and section 1.7 surveys the relevant literature.

1.1 The Core Components

Charles Sanders Peirce (1839–1914) was an American philosopher and logician, famous not only for his contributions to logic, mathematics, and meaning, but also for being the foundational theorist of pragmatism. In the rest of this section, we describe this model of meaning in relatively qualitative terms, focusing on the ways it resonates with—but also routes around—some of Peirce's core assumptions.

In a Peircean view of meaning, a *semiotic process* involves three interrelated components: a *sign* is whatever stands for something else; an *object* is whatever is stood for by a sign; and an *interpretant* is whatever a sign creates, insofar as it is taken to stand for an object. For example, a father points (sign) to a bird (object) and his daughter turns to look (interpretant). A gold or copper coin is pulled from an urn (sign), indicating a particular composition of

coins within the urn (object), and the urn is then accepted or rejected (interpretant). Finally, a student raises their hand (sign), indicating their desire to ask a question (object), and the teacher calls on them (interpretant).

Unlike many other accounts of meaning, which foreground a single relation of standing for (e.g., signifier-signified, expression-referent, signal-message, and so forth), Peirce's account foregrounds a relation between two such relations. Loosely speaking, *a sign stands for its object on the one hand, and its interpretant on the other, in such a way as to make the interpretant stand in relation to the object corresponding to its own relation to the object*. These three components, as well as this relation between relations, are shown in figure 1.1.

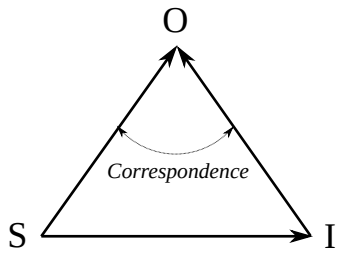

Figure 1.1: Signs, Objects, and Interpretants

While there is some slippage in Peirce's work between the interpretant as an agent (or 'mind') that interprets the sign, and the interpretant as another sign (brought into being by the agent's interpretation of the first sign), we will be careful to distinguish between the two senses. In the scenario just sketched, for example, the father is the signifying agent, or *signer*, and his pointing gesture is the sign; the daughter, in contrast, is the interpreting agent, or *interpreter*, and her turning to look is the *interpretant*. The sign-interpretant relation is thus both mediated by, and mediating of, this signer-interpreter relation. Phrased another way, a relation between two entities (the sign and the interpretant) mediates a relation between two agents (the signer and the interpreter). As will be seen, such social relations—and relations between relations more generally—are often the impetus for and outcome of semiotic processes.

In contrast to Peirce, who tended to focus on *significance* (or meaning, in the sense of 'standing for'), we will also focus on *selection* (qua motivation, or meaning in the sense of 'striving for'). This model therefore includes not only the objects that signs stand for (and the interpretants that signs create), but also the values that agents strive for and/or the functions that agents serve.

Phrased another way, semiosis, and hence both signification and interpretation, are mediated not just by teleological, teleonomic, and teleomatic processes (broadly construed), but also by 'teleognomic' processes (Enfield and Kockelman 2017). Therefore, while only a few such agents are capable of selecting (in the sense of self-consciously choosing, as stereotypically understood), almost all such agents are capable of being selected (whether via natural selection, reinforcement learning, algorithmic sieving, design, manufacture, discipline, regimentation, framing, and the like).

In short, semiotic agents are not just the instigators, but also the outcomes, of various modes of selection. And so aside from their shared capacity to signify and interpret, such agents are radically heterogeneous. As will be seen, this model of meaning can be used to analyze the behavior of a wide range of agents, some of which are self-conscious, well-informed, technologically sophisticated, and far-seeing, and some of which are at the limit of what counts as life.

In contrast to Peirce, who foregrounded three components in his analysis of semiotic processes (sign, object, and interpretant), we will sometimes include not just the agent, but also a fourth component: the *consequent*. Just as we distinguish between the sign (as whatever is sensed by the agent) and the object (as whatever is mediated by the sign), we also distinguish between the interpretant (as whatever is instigated by the agent) and the consequent (as whatever is mediated by the interpretant). In the first scenario to be examined, the sign will be a symptom, the object will be an illness, the interpretant will be the action of taking (or not taking) a medication, and the consequent will be the effect, and/or side effect, that the medication has when taken. While both kinds of entities (objects and consequents) are 'absent' (in the sense of not being directly sensed or instigated by the agent), they are absolutely 'present' (in the sense of being indirectly sensed and instigated by the agent through such mediating signs and interpretants, whatever the degree of remove).

Phrased another way, just as there is a kind of *experiential slash* between the sign (which the agent can directly sense) and what it stands for (the object), there is also a kind of *agentive slash* between the interpretant (which the agent can directly instigate) and what it results in (the consequent).

All the foregoing components may be put into relation, and thereby diagrammed and described, as follows: *The relation between what the agent (A) instigates (I) and what it senses (S) makes sense in relation to the relation between what the agent effects (C) and what it projects (O), given the relation between that relation (O/C) and the agent's identity, origins, and/or interests*. See figure 1.2. A key argument of this book is that meaning resides in such interrelations.

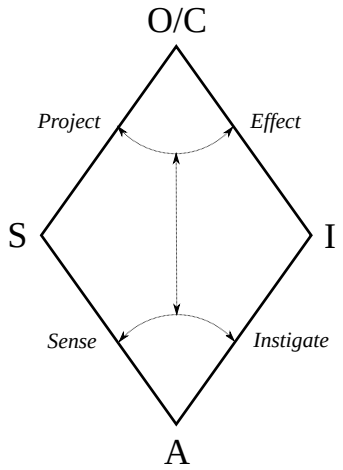

Figure 1.2: Relations between Relations

As will be seen, where we—as analysts—draw the boundary between object and sign or interpretant and consequent, and hence how we determine where sensation ends and projection begins, or where instigation ends and effecting begins, is often a frame-dependent and/or scale-specific decision (Kockelman 2011). For example, and more generally, the interpretant of a prior semiotic process may constitute the sign in a subsequent semiotic process; the sign of a lower-order semiotic process may constitute the object of a higher-order semiotic process; the difference between interpretants and consequents, or between objects and signs, may sometimes be collapsed; affects, inferences, outputs, and traits may be framed as interpretants no less than actions, habits, utterances, and responses; a single, relatively distributed agent may incorporate several smaller agents; most selecting agents were themselves selected; agents that select often employ agents that were selected; semiosis may mediate relatively private cognitive processes no less than relatively public communicative practices; thinking, evolving, and learning are irreducibly coupled; and so on, perhaps indefinitely.

While Peirce—and certainly Peirceans—tend to focus on sign-object relations through the frame of his most famous trichotomy (icons, indices, and symbols), not much will be made of these categories here. Instead, we will focus on the sensibilities and assumptions that agents have regarding not only sign-object relations, but also interpretant-object and consequent-interpretant relations. When dealing with human-like agents, such sensibilities and assumptions will often be framed as propensities and probabilities, and hence as

intensities of desire (or affect more generally) and degrees of commitment (belief, doubt, and/or uncertainty). Such sensibilities and assumptions may be referred to as the *grounds*, or 'guiding principles', of semiotic processes. As will be seen, not only can such interpretive grounds constitute the objects of meta-semiotic processes (and thereby become 'figures'), they can also be transformed by semiotic processes. They thereby constitute a relatively fluid, as opposed to a fixed, semiotic resource.

Such grounds might thereby be likened to the metric in general relativity, insofar as they constitute dynamic variables that enable agents to measure quantities, or at least register intensities, in ways that take into account invariance no less than variation. And, even more crucially, just as the metric is both determined by the distribution of matter and determining of the flow of matter, interpretive grounds are at once shaped by semiotic processes and shaping of semiotic processes. As will be seen, and as might be expected, such a coupling can lead to somewhat complicated equations of semiosis, with rich and revealing solutions.

In short, it is assumed that (humanlike) semiotic agents can engage in relatively immediate processes of sensation and instigation, meaning that they can take differences as inputs (and thereby be affected) and make differences as outputs (and thereby be effective), where a difference is some kind of differentiable event. It is also assumed that agents can engage in relatively mediate processes, such that they are sensitive to whatever causes their sensations and whatever is caused by their instigations. Not only can such agents relate to the mediate through the immediate (in two directions), they also have a model of such relations (however embodied or engenomed, erroneous or simplistic, innate or unconscious). In particular, they have a sense of the organization and intensity with which the immediate and mediate are coupled: what kinds of objects lead to what kinds of signs, and what kinds of consequents follow from what kinds of interpretants, and with what degree of strength, likelihood, or frequency. Finally, they have a sense—however unconscious, instinctual, misguided, or limited—of the relative value of various relations among such events; and hence a sense of which worlds they would rather reside in, and well as which actions they should undertake, in which contexts, to better realize such worlds.

Crucially, saying that such agents have a 'sense' of such interrelations does not mean that they must consciously register them in any way. It only means that analysts must take into account such relations if they are to make sense of the behavior, identity, and/or origins of such agents. Such a strong, and perhaps strange claim will become clearer in the sections that follow.

While Peirce's account of semiosis is frequently used to conduct relatively qualitative studies of real-time semiotic processes as they unfold in 'the real world' (by anthropologists, linguists, historians, discourse analysts, critical theorists, and like-minded scholars), we use it—or at least our extension and transformation of it—to offer a mathematical model of meaning. Peirce himself, of course, made great contributions to logic, statistics, and inference; and so he would surely be comfortable with the quantitative and logical direction that his usually qualitatively deployed ideas will be taken. (However trenchant his critique of the limits of this particular approach might have been.) The strategy used in this book, however, is relational as opposed to quantitative. Rather than crunching numbers, proving theorems, or running computer simulations, we will elucidate the *critical points* of relatively abstract scenarios, which are chosen to foreground key aspects of the model of meaning being offered in relation to classic topics from a variety of literatures. In this way, we usually strive to find analytic solutions to illuminating scenarios, along with their fixed points, singularities, and limits.

Moreover, this book is not just meant to offer a mathematical model of meaning; it is also meant to put quantitative approaches to meaning in dialogue with qualitative approaches, as well as to put naturalist approaches in dialogue with critical and hermeneutic approaches. This is not just done as a bridge-building enterprise, such that scholars and scientists of different persuasions can be in conversation. It is also done to offer qualitative theorists (in history, anthropology, media studies, critical theory, and the like) a framework for analyzing the history and practice of mathematical modes of meaning, many of which constitute key agents and intermediaries in the infrastructure of modern life.

The next section lays out some of the key features of this mathematical framing in a more formal way, and with a focus on relatively rational agents. As will be seen, it assumes that readers have a basic understanding of the mathematics of probability, expected value, and related notions. Readers desiring a relatively simple and concrete scenario, with each mathematical term explained in detail and each calculation carried out in full can look to chapter 2, which provides a relatively easy transition to some of these technicalities.

1.2 The Main Equation

Building on the foregoing conceptual model, an *interpretive ground* may be more formally represented as an eight-tuple $\langle S, I, O, C, P, L, E, V \rangle$. The first four elements in this ordered list constitute an *ontology*, understood as the following four core variables (along with their possible values):

- A set of *signs*, $S = \{S_1, S_2, S_3, \dots\}$, understood as whatever an agent can sense or register more generally (where $|S|$ will denote the number of signs in this set, and similarly for the other three variables that follow);
- A set of *interpretants*, $I = \{I_1, I_2, I_3, \dots\}$, understood as whatever an agent can instigate, update, or otherwise directly do;
- A set of *objects*, $O = \{O_1, O_2, O_3, \dots\}$, understood as whatever an agent can project, infer, learn about, or come to know, given what it senses;
- A set of *consequents*, $C = \{C_1, C_2, C_3, \dots\}$, understood as whatever an agent can effect, produce, accomplish, modify, maintain, or otherwise indirectly bring about given what it instigates.

The next three elements in the tuple constitute *commitments*, which may be modeled as conditional probability distributions (along with a presupposed causal structure) as follows:

- The *priors*, $P(O_i)$, where index i ranges over the objects in the ontology such that the priors may usually be represented as an $|O| \times 1$ matrix;
- The *likelihoods*, $P(S_i/O_j)$, where the indices range over the signs and objects in the ontology such that the likelihoods may usually be represented as an $|S| \times |O|$ matrix;
- The *eventualities*, $P(C_i/I_j \wedge O_k)$, where the indices may range over the consequents, interpretants, and objects in the ontology such that the eventualities may usually be represented as a $|C| \times |I| \times |O|$ matrix.

The last element in the tuple consists of *values*, $V(C_i \wedge I_j \wedge O_k)$, where the indices usually range over the consequents, interpretants, and objects in the ontology such that the values may be represented as a $|C| \times |I| \times |O|$ matrix.

We will say that such an interpretive ground is relevant to a relatively rational agent in a particular context insofar as the agent uses the ground to evaluate its instigations given its sensations in that context. And we will assume that the mode of evaluation may itself be modeled using some variant of the following equation:

$$V(I_i/S_j) = \sum_k \sum_l V(C_k \wedge I_i \wedge O_l) \cdot P(C_k/I_i \wedge O_l) \cdot P(O_l/S_j)\,, \tag{1.1}$$

where the last term in the equation may be expanded using Bayes' theorem:

$$P(O_l/S_j) = \frac{P(S_j/O_l) \cdot P(O_l)}{\sum_m P(S_j/O_m) \cdot P(O_m)} = \frac{P(S_j/O_l) \cdot P(O_l)}{P(S_j)}\,. \tag{1.2}$$

Given a suitable specification of an agent's ground, equation (1.1) allows us to calculate the value of various interpretants (I_i), given various signs (S_j), as a function of the objects (O_l) that are stood for by those signs, as well as

the consequents (C_k) that are created by those interpretants. As may be seen, its terms are fully specified by the elements in the interpretive ground, and hence by the agent's ontology, commitments, and values. Assuming that the agent instigates the most valuable interpretant that it can, given the sign that it presently senses, this equation thereby specifies the behavior of the agent in all relevant situations: how it will most likely interpret each of the possible signs it might encounter insofar as it is beholden to such a ground. The agent's sign-dependent interpretants—and thus their actions, inferences, interrelations, and affect—are thereby guided by such a ground.

Given its importance to the analysis, equation (1.1) will be referred to as the "main equation" in the chapters that follow. As may be seen, it involves a double summation. First, there is a sum over all the possible objects (as those entities that are stood for by signs); second, there is a sum over all the possible consequents (as those entities that are created by interpretants). In so doing, the agent is, in effect, summing over the product of values, eventualities, likelihoods, and priors for the objects and consequents in question. This double sum (in the discrete case), or double integration (in the continuous case to be discussed later), is our way of handling the two kinds of value that are essential to meaning: what signs stand for and what agents strive for.

It should be noted that we did not derive the main equation from first principles, or anything else so sophisticated. Rather, using classic notions like conditional probabilities, expected values, and Bayes' theorem, we simply expressed the foregoing conceptional framework (consisting of signs, interpretants, objects, consequents, agents, and grounds) in the simplest mathematical formulation we could imagine. Peirce himself would have had access to, and thus could have employed, the same resources for his own writings. There are, to be sure, many other ways that we might mathematically realize these relations, using more recent developments like neural networks, expected utility, and prospect theory, inter alia. We will look at some of these alternative formulations, which are arguably more complicated, and possibly more realistic, in later chapters. Phrased another way, this mathematical model is designed to be as simple as possible while remaining useful, plausible, transposable, and generalizable.

While the main equation can be used for any ontology of any size, in the particularly simple scenario that will be analyzed first, there will be two signs (S_1 and S_2), two objects (O_1 and O_2), two interpretants (I_1 and I_2), and two consequents (C_1 and C_2). In effect, then, we will be dealing with a universe in which there are 16 'possible worlds'—understood as the $2 \times 2 \times 2 \times 2$ possible ways that such a scenario-specific universe might be realized. We will later introduce a continuous form of this equation, in which there might be an

infinity of possible signs, objects, interpretants, and consequents (and hence, loosely speaking, an infinity of possible worlds).

As will be discussed at length in later chapters, and as presupposed by our specification of the agent's commitments—$P(O)$, $P(S/O)$, and $P(C/I \wedge O)$—we typically assume that (the agent tacitly assumes that) objects causally influence signs, interpretants and objects causally influence consequents, and signs only causally influence interpretants through agents. Signs typically provide evidence of objects; they are auspicious. Interpretants, in the context of objects, typically create consequents; they are efficacious. This is what is meant when we say that an agent's ground takes for granted a certain causal structure (perhaps erroneously).

Just as grounds may be framed relatively subjectively (as an agent's understanding of the causal patterns and valuable resources present in an environment, insofar as such patterns and resources are relevant to the agent's semiotic processes), they may also be framed relatively objectively (as the casual patterns and valuable resources that are actually found in an environment, insofar as such patterns and resources influenced the semiotic processes of such agents and/or gave rise to them as agents). In other words, for certain kinds of agents, in certain contexts, we make sense of their semiotic processes by reference to their *subjective grounds* (such as their beliefs, desires, habits, or instincts); for other kinds of agents, in other contexts, we make sense of their semiotic processes by reference to the *objective grounds* of their environments (such as the conditions under which they evolved, or past regimes in which their behavior was socially regimented, governed, or otherwise punished and rewarded). As will be seen, to make sense of the behavior of most kinds of agents, both kinds of grounds must be considered, as well as the discrepancies and overlaps between them.

All this is another way of saying that analysis must take into account such interpretive grounds, be they relatively objective or subjective, in order to make sense of semiotic processes and the origins of the agents involved in them. Indeed, it is usually much more complicated than this, for the consequential environments of most semiotic agents are precisely the semiotic processes, and hence semiotic grounds, of other agents, which are themselves often temporally dependent and hence in transition. Thus there is often nothing like an objective context, or fixed environment, for agents to accommodate to—at least on longer time scales. As will be seen, the analysis is precisely designed to handle such complicated agentive couplings, and the complex time-dependent interactions that result.

It should be emphasized that the main equation is only interesting insofar as the scenarios that it is used to model are interesting. Indeed, it would be

a simple thing to specify the ground and then solve for the best interpretant, given each sign. Except in the case of the first scenario analyzed, we will not engage in such calculations. We will rather examine solutions to this equation for various scenarios as particular variables change to highlight various critical points, or thresholds.

It should also be emphasized that the main equation, as it stands, is meant to be flexible, dynamic, and portable. Many scenarios will turn on relatively reduced variants of it (e.g., contexts in which consequents can be ignored, such that the main equation only makes reference to signs, objects, and interpretants). Most of the scenarios examined will involve more than one agent, such as a signer interacting with an interpreter. In such cases, the equations governing each agent's behavior may make reference to the equations governing the behavior of the other agents. Finally, insofar as the grounds of agents often transform over time, the equations will often be functions of time. In short, the main equation will often be reduced to simplified variants, coupled with complementary equivalents, and/or dynamically iterated.

Moreover, such equations will usually be put into relation with other mathematical approaches. In so doing, it will be seen that the values in question can range from economic values (like price and utility) and existential values (like prudence or honor) to biological values (like fitness), affective values (like pleasure and pain), and thermodynamic values (like free energy). We will thereby put this equation in dialogue with various insights and formulations from mathematical biology, statistical mechanics, information theory, political economy, machine learning, and anthropology. Indeed, as the title of this book suggests, and as later chapters will make clear, this model of meaning is really composed of a family of closely related models.

The next section highlights two of these alternative approaches, the first involving agents that evolve, and the second involving agents that learn. This will be our first foray into the dynamic coupling of semiotic grounds and semiotic processes, whereby each enables and constrains the other, leading to agents that alter their behavior, and thus their interactions with each other, through experience over time.

1.3 Agents That Evolve and Learn

While the main equation, as just described, is meant to model the behavior of relatively mindful agents, such as humans and animals, that *select* particular interpretant-sign relations given the contents of their grounds, a variant of it will also be used to model the effects of relatively mindless agents, such as alleles, which *are selected* as a function of the degree to which

the semiotic processes that they engender are relatively value-generating, or fitness-promoting, in a given environment.

More precisely, we will examine alleles A_i that cause their bearers to cause ($\rightarrow$) particular events (qua interpretants) in the context of particular events (qua signs) with certain conditional probabilities:

$$A_i \rightarrow P_{A_i}(I/S)\,, \tag{1.3}$$

where the relative fitness of such alleles may be determined using the following equation:

$$F(A_i) = \sum_{jk} P(S_j) \cdot P_{A_i}(I_k/S_j) \cdot V(I_k/S_j)\,. \tag{1.4}$$

Here, the third term on the right-hand side is simply the main equation, whose components will be suitably reframed (as discussed previously) to describe relatively objective patterns and resources within an environment (as opposed to agent-specific, and hence often relatively erroneous and/or subjective, understandings of such patterns and resources). And the first term on the right-hand side, understood as the probability of a triggering event, qua sign, may depend on the features of a relatively stable environment or the behaviors of other evolving agents, inter alia.

This measure of fitness will be used to track the changing frequency of such alleles over generations within a population, such that the evolution of biosemiotic agents may be studied. In particular, it is argued that the genomes of organisms embody interpretive grounds that transform on phylogenetic time scales, whereby the behavioral phenotypes generated by such grounds become better adapted to particular environments (however fleetingly), and where such environments can include the semiotic processes and interpretive grounds of other agents.

The main equation can also be used to study semiotic agents that learn through experience via positive and negative regimentation, reinforcement, or discipline. Such agents register the results of past experience in their habits, as opposed to their genomes. To see how, let the term *practice* denote the instigation of an interpretant in the context of a sign. And let the term *register* (e.g., a 'habitus' in the tradition of Marcel Mauss, or simply a 'habit' in the tradition of Peirce) denote an array of weights indexed to particular practices $W(I/S)$. The weights in an agent's register, or habitus, are not just determined by its past experience; they are also determining of its future practices. In particular, each time an agent engages in a practice (instigating some interpretant in the context of some sign), the value that it receives is added to the appropriate weight in its register; and the greater the weight of a practice in the agent's register, the

more likely the agent is to instigate that particular interpretant in the context of that particular sign in future interactions.

As will be more carefully shown later, if each of the weights in an agent's register at some slice of time is known, $W(I_i/S_j)$, the expected weight of each of the behaviors in their register during the next slice of time, $W'(I_i/S_j)$, can be determined using the following difference equation:

$$W'(I_i/S_j) = W(I_i/S_j) \cdot e^{-\alpha} + P(S_j) \cdot P(I_i/S_j) \cdot V(I_i/S_j)\,. \tag{1.5}$$

Here, α is a parameter that determines how rapidly the agent 'forgets' prior experience (by discounting weights that were registered earlier); $P(S_j)$ is the probability that a particular sign occurs in some environment, which is analogous to the same term in equation (1.4); and $V(I_i/S_j)$ is again the main equation. The term $P(I_i/S_i)$, meaning the probability that the agent instigates I_i in the context of S_j, may be determined by the current weights in the agent's register:

$$P(I_i/S_j) = \text{softmax}\left(W(I_i/S_j)\right) = \frac{e^{\beta \cdot W(I_i/S_j)}}{\sum_k e^{\beta \cdot W(I_k/S_j)}}\,. \tag{1.6}$$

Here β is a parameter that determines how concentrated the probabilities are around the highest weights in the agent's register, and the softmax function takes in weights and turns out probabilities (by exponentiating all the weights in the register and then normalizing the weights in each column). In short, such agents internalize the effects of prior interactions, and the habits thereby engendered guide their future practices.

We now turn to coupled semiotic agents that come to communicate using conventional signals.

1.4 Signers, Interpreters, and Signals

Both of the foregoing kinds of agents—those that evolve and those that learn—often arise in contexts that involve multiple agents, each of which plays a distinct and complementary role in a division of semiotic labor. Figure 1.3 shows one such scenario.

As may be seen, there are two agents mediated by a single semiotic process, consisting of an object (O), a sign (S), and an interpretant (I). The signer, A_1, senses an object and instigates a sign. The interpreter, A_2, senses the sign and instigates an interpretant. And the relation between the object and interpretant yields a value that benefits both signer and interpreter alike.

Just as the interpreter's access to the object is mediated by the signer, so is the signer's access to the interpretant mediated by the interpreter. Each agent, as it were, has direct access to only two-thirds of the interaction; so

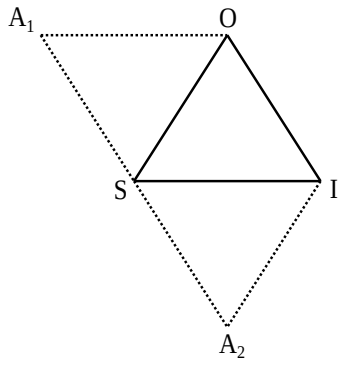

Figure 1.3: Signaling Agents

the behavior of each agent makes sense only in reference to the behavior of the other.

While such a scenario may be imagined in many different ways, for present purposes the object should be understood as the location of a source of food that the signer can sense but not obtain (and the interpreter can obtain but not sense); the sign as some sound or gesture that the signer can instigate and the interpreter can sense; and the interpretant as the action of going to one location or another to obtain the food (which will then be shared by the two agents).

Note that in such a simple scenario, there is no need of consequents. In effect, it is assumed that what the agent instigates (going to some location) is what the agent achieves (arriving at that location).

To flesh out the interpretive grounds of the two agents and make the scenario relatively general (and symmetric), we might assume that there are N objects, N signs, and N interpretants. We might assume that each of the N objects is equally likely, such that $P(O) = 1/N$. We might assume that the semiotic strategies of the signer and interpreter are $N \times N$ matrices, understood as time-dependent, conditional probabilities, which may be denoted as $P_t(S/O)$ and $P_t(I/S)$, respectively. And we might assume that the value received by each agent, as a function of their signifying and interpreting practices, is

$$V(I_i \wedge O_j) = \frac{v}{2} \cdot \delta_{ij}\,, \tag{1.7}$$

where v is the value of the food and δ_{ij} is the Kronecker delta function (or an $N \times N$ identity matrix), such that δ_{ij} is 1 when $i = j$ and 0 otherwise. In other words, so long as the interpreting agent goes to the right location, regardless of the sign that is employed by the signifying agent, both agents benefit equally.

Given such assumptions, the main equation becomes two coupled, time-dependent equations, which describe the values received by the agents as a function of their semiotic strategies:

$$V_t(S_i/O_j) = \sum_k V(I_k \wedge O_j) \cdot P(O_j) \cdot P_t(S_i/O_j) \cdot P_t(I_k/S_i)\,, \tag{1.8a}$$

$$V_t(I_k/S_i) = \sum_j V(I_k \wedge O_j) \cdot P(O_j) \cdot P_t(S_i/O_j) \cdot P_t(I_k/S_i)\,. \tag{1.8b}$$

As may be seen, for the scenario in question, we assume that the values and priors are independent of time, and that consequents can be ignored. As may also be seen, the value of each agent's strategy depends on the other agent's strategy. Indeed, given the way that $V(I \wedge O)$ was defined, the expressions being summed over in $V_t(S/O)$ and $V_t(I/S)$ are equivalent. This should be unsurprising: the fates of such semiotically coupled agents rise and fall together.

To be sure, this is one of the simplest kinds of semiotic systems: two agents learning, or evolving, to employ relatively conventional (or 'arbitrary') signals. If plugged into the kinds of selection processes described in section 1.3 (which themselves are expanded to encompass signers in addition to interpreters), the two agents will evolve, or learn, to employ one of the $N!$ possible codes to communicate. It is introduced here to show how the main equation, or rather one of its simplified variants, can capture one of the most celebrated of semiotic processes.

We have so far offered a relatively capacious definition of meaningful interrelations, shown how such interrelations can be studied mathematically (by projecting a certain causal, statistical, and evaluative structure onto them), highlighted some of the presuppositions that go into such a projection, contrasted three kinds of prototypic agents (those that reason, those that evolve, and those that learn), and sketched some of the ways that the main equation can be used to capture the division of semiotic labor and processes like signaling.

The next section examines the relation between three tightly coupled, and easily conflated, concepts.

1.5 Meaning, Information, and Value

Terms like *meaning*, *information*, and *value* are somewhat slippery, sometimes referring to what signs stand for, and sometimes referring to what agents strive for. We now sketch some of the ways that these terms will be used, and mathematically realized, in the chapters that follow. To keep things simple, we will focus on relatively humanlike agents, and let later chapters generalize such

claims to a wider range of agents. But be warned: as was the case in the last two sections, a few parts of this discussion may seem relatively elliptic at first glance, insofar as the concepts and measures in question will not be fully explained until certain ideas and definitions are in place.

In a very narrow sense, the meaning of a sign is the object that it stands for. To paraphrase Peirce, the object of a sign is what a sign, so far as it is known, and known as a sign, allows one to know. But, as seen in the foregoing sections, this term can also be expanded to include the motivation of the signer (for expressing the sign in the first place), the motivation of the interpreter (for offering the particular interpretant that they do), the interpretant per se, the consequent that the interpretant effects, relations between relations more generally (recall figure 1.2), and much else besides. Focusing on the narrow sense of this term, we will see that signs rarely have a single well-defined meaning. What they *do*, rather, is change an agent's commitments regarding the probabilities of objects (and thereby change the likelihood of particular interpretants). Moreover, objects themselves are often best understood not as particular entities or events, but as sets of possible worlds.

The foregoing points might be formalized by thinking of the meaning of any sign, S_i, as a row of vectors (one for each object, O_j, in the agent's ontology), where the tail of a vector is located at $P(O_j)$, understood as the prior probability of O_j, and the head of the vector is located at $P(O_j/S_i)$, understood as the posterior probability of O_j (where the transition from prior to posterior probability is conditioned on the occurrence of the sign; see figure 1.4). In effect, such sign-specific sets of vectors, which are thus tensor-like, represent the change in the probabilities of objects given signs (at certain times). A sign not only may make a particular object more or less probable (to an agent) than before, it also may do so to a greater or lesser degree. In this way, the meaning of a sign is not only agent specific and time dependent (insofar as it turns on the present interpretive ground of a particular agent), but also graduated and modal (insofar as it turns on relative intensities of various propensities).

Crucially, and keeping with caveats introduced in earlier sections, the foregoing points are not meant to imply that any agent is necessarily, or even usually, conscious of such changes. It is simply the case that it is necessary to take into account such changes to make sense of semiotic processes insofar as the grounds of agents and environments depend on them.

For the purposes of this discussion, a possible world is any conjunction of an object, sign, interpretant, and consequent that is available in an agent's ontology. Let w_{ijkl} denote the world in which the object is O_i, the sign is S_j, the interpretant is I_k, and the consequent is C_l. If a universe, denoted as Ω, is a complete set of possible worlds, then the size of this set, its cardinality, is simply

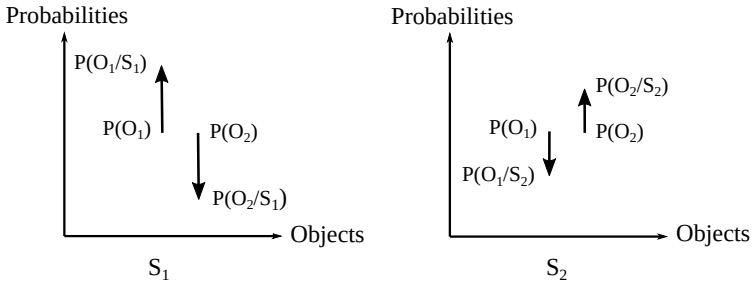

Figure 1.4: Meaning of Signs as Change in the Probabilities of Objects

$|O| \times |S| \times |I| \times |C|$, where $|O|$ is the number of objects in the agent's ontology (and similarly for the other three variables; see figure 1.5). The agent's ontology is really a cosmology—not so much a view of the world (or *Weltanschauung*) as a stance towards a universe (in the context of a scenario). And a possible world is simply one way in which such a universe might unfold, and thereby become fully realized.

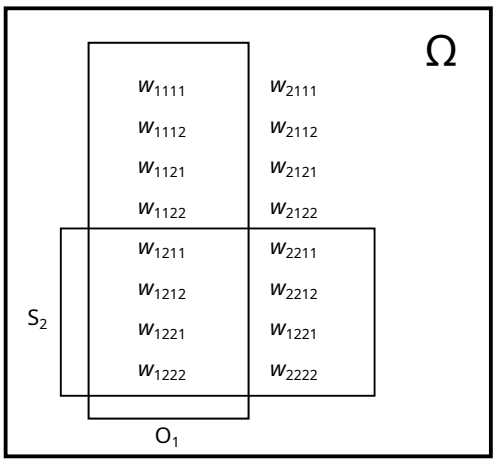

Figure 1.5: Object and Sign as Sets of Possible Worlds

In saying that an object is a set of possible worlds, we mean that it is the set of worlds compatible with the index of the object in question. It is thus all the worlds in which the object obtains (and hence certain ontology-specific, and hence ground-specific, conditions are met), however else those worlds may otherwise differ. For example, O_1 is the set of worlds whose first index is fixed

at 1 (and whose other indices are free to vary). Crucially, if an object is a set of worlds, so are signs, interpretants, and consequents. For example, S_2 is the set of worlds in which the second index of w_{ijkl} is fixed at 2 (while all the other indices are free to vary). To return to the first point, learning that one is in a world in which S_2 obtains changes one's estimation of the probability that one is in a world in which O_1 obtains. And this, in turn, changes how one acts (thereby changing the probability that a particular interpretant obtains) which, in conjunction with the object, changes the probability that a particular consequent obtains.

What signs do, in this framing, is change the probability that the actual world in which an agent resides belongs to one subset of the universe or another. Phrased in another way, insofar as a sign is meaningful for an agent, in the narrow sense, by knowing something about the second index of w_{ijkl}, that agent can come to know something about the first index (which thereby constrains the other two indices as well). Loosely, meaningfulness allows an agent to skip across slashes, now framed as gaps between available indices on possible worlds.

As will be seen, one kind of interpretant enabled by the meaning of a sign—perhaps the *ultimate* interpretant—is precisely a transformation of the agent's ground, which may include a change in their ontology, and hence their beliefs and values, as well as their habits more generally. It is, in effect, a new stance towards a universe (be it embodied or enminded), and all the worlds that it may include or omit, reject or embrace. We borrow Peirce's terminology, however different our theory is, because such a change in habit or belief, as engendered by a sign, may potentially affect all future semiotic processes in which the agent participates, and hence all future worlds that the agent may one day inhabit.

Within this framing, a somewhat subtle point is that signs and objects, as well as interpretants and consequents, are not usually entities or events per se (such as a symptom or a sickness); they are, rather, interrelated entities and events such as 'the person evinces the symptom' (as a sign) or 'the person is sick' (as an object). In effect, such variables are full-fledged propositions. In one sense, then, they are sets of worlds; in another, equivalent sense, they are functions from worlds (and indices more generally) to truth values. That is, they return the value 'true' (as output), if the world that they are given (as input) is a member of the set that they constitute. Moreover, as propositions, they are infinitely rich in how they might be rendered—and so they may involve complicated relations not just among entities, events, qualities, and the like, but also among times and worlds per se. Just as anything sensible can be a sign and anything instigatable can be an interpretant, just about anything imaginable

can be an object or a consequent. At least *in potentia*, to a particular kind of agent, with a large-enough universe (given its ontology, and semiotic capacities more generally).

We now move from the meaning of a sign (in a narrow sense) to the information of a sign, which is to be understood as the change that the sign makes in an agent's uncertainty regarding which object—or which world more generally—obtains. In one important sense, which goes back to Claude Shannon, the informativeness of a sign, S_i, is its surprise value, understood as the logarithm of its improbability: $-\log_2 P(S_i)$. The informativeness, or 'entropy' (denoted as H), of a set of signs is then the average surprise value of all the signs in the set: $H(S) = -\sum_j P(S_j) \cdot \log P(S_j)$. Rather than focus on the surprise value of a set of signs, however, we are interested in the expected *change* in the surprise value of a set of objects due to the acquisition of a sign (for an agent, given its ground, on average). Somewhat loosely, this may be understood as the average number of questions that a sign saves an agent from having to ask (to learn the identity of the object), and hence the change in the agent's uncertainty regarding which object (or set of worlds) obtains:

$$\mathcal{I}(S) = \sum_j P(S_j) \cdot \left[H(O) - H(O/S_j) \right] . \tag{1.9}$$

Here,

$$\begin{aligned} H(O) - H(O/S_j) = &- \sum_i P(O_i) \cdot \log_2 P(O_i) \\ &+ \sum_i P(O_i/S_j) \cdot \log_2 P(O_i/S_j) \end{aligned} \tag{1.10}$$

is the difference between the prior and posterior uncertainties of the agent regarding which object obtains. (The posteriors, $P(O/S)$, may be found from the priors and likelihoods using Bayes' theorem.) In effect, one simply calculates the average surprise value of all the objects in the set (prior to the acquisition of a sign) and subtracts from this the average surprise value of all the objects in the set conditioned on the occurrence of a sign, which itself is averaged over all possible signs that could occur. It can be readily shown that this is equivalent to the relative entropy of $P(O/S_j)$ with respect to $P(O)$, as averaged over all possible signs.

A related measure may be used to calculate how the presence of an agent contributes to the change in Shannon entropy (or 'complexity') of a universe, in the sense that their presence makes certain modes of patterning more or less likely. This is because information is closely related to entropy, just as organization, and hence the reduction of entropy—understood as the creation of

patterning—is closely related to work. This is not work in the sense of giving form to substance for the sake of function, but work in the sense of organizing complexity for the sake of predictability. Indeed, the possible worlds of formal semantics are closely related to the microstates of statistical mechanics, and this fact allows for a relatively seamless connection not only between meaning and information, but also between Boltzmann entropy, organization, work, and constraint. Phrased in another way, besides thinking about the meaning, information, and value of a sign (from the standpoint of an interpreting agent), one may also consider the entropy (complexity or unpredictability) of a scenario—itself filled with one or more interpreting agents—from the standpoint of an analyst.

Having distinguished between the meaning and information of a sign, in a relatively specific sense, we may now take up the value of a sign, which may be understood as how much an agent would be willing to sacrifice to obtain the sign, insofar as its meaning allows that agent to better guide its actions (justify its reasons, express its feelings, and/or modify its social relations).

If the interpretive ground of an agent is known, the main equation allows one to calculate the value of an interpretant given a sign, $V(I_i/S_j)$. As was shown earlier, this equation takes into account not only the object that the sign stands for, but also the consequent that the interpretant leads to. Indeed, using the agent's ground, one can even calculate $V(I_i)$, which may be understood as the value of an interpretant in the absence of any information provided by a sign. (Simply replace the posteriors, $P(O/S)$, in equation (1.1) with the priors, $P(O)$.) Assuming that the agent always instigates the most valuable interpretant that it can (given its ground), whether it has received a sign or not, one can calculate the change in expected value that any particular sign provides, $\mathcal{V}(S_j)$. Averaging this value across all possible signs, one can calculate the expected value of receiving a sign per se:

$$\begin{aligned}\mathcal{V}(S) &= \sum_i P(S_i) \cdot \mathcal{V}(S_i) \\ &= \sum_i P(S_i) \cdot \left[\max\left(V(I/S_i)\right) - \max\left(V(I)\right)\right].\end{aligned} \tag{1.11}$$

Here, the term $\max\left(V(I)\right)$ denotes the expected value of the most valuable interpretant (in the absence of any sign), whereas the term $\max\left(V(I/S_i)\right)$ denotes the expected value of the most valuable interpretant given some particular sign. As may be seen, $\mathcal{V}(S_i)$ is just the difference in expected value that sign S_i makes (insofar as it is meaningful to an agent). And $\mathcal{V}(S)$ is just the average difference in expected value that signs provide an agent, insofar

as they allow that agent to better relate to its environment, and thereby better direct its efforts.

Note that, in contrast to the information of a sign, $\mathcal{I}(S)$, which only takes into account the relation between objects and signs (in the guise of priors and likelihoods), the value of a sign, $\mathcal{V}(S)$, takes into account all aspects of an agent's ground (and hence all relations among objects, signs, interpretants, and consequents). Indeed, a key argument of chapter 2 will be that standard measures of information—like relative entropy or Kullback–Leibler divergence—are relatively uninformative.

As will be seen, one can even calculate the value of an agent's interpretive ground, $\mathcal{V}(G)$, given the actual causal patterns and valuable resources available in an environment. Somewhat loosely, this is a measure of the goodness-of-fit between the agent's stance towards a universe and the universe per se. Such a measure allows one to assign a relative value to better and worse representations of reality, and hence to weigh the costs and benefits—whatever the currency, be it euros, pleasure, or free energy—of having a particular worldview.

If the use-value of a sign may be understood as the function that it serves in a communicative encounter, then three additional and closely related varieties of value have just been provided: truth value (qua meaning), surprise value (qua information), and exchange value (qua price). As should be clear, while such terms have very different measures, they all turn on relatively quantitative changes that are induced by the presence of signs in relation to agents and their interpretive grounds in certain worlds, at certain times. Each of these concepts will be expanded and explicated, as well as qualified and critiqued, in the chapters that follow.

The next section surveys the arguments and architecture of this book.

1.6 Overview of the Chapters

In broad strokes, all the chapters that follow make two kinds of contributions. At one level, they offer case studies of relatively simple, adjustable, and generalizable scenarios. In so doing, they provide a concrete exploration of the entailments of this model, walk readers through the mathematical details, and provide lively demonstrations of interesting systems that may be easily adapted to wider concerns. At another level, they flesh out the presumptions of the model or show how it can be extended to understand more complicated processes. In so doing, they not only explicate, but also criticize, complicate, and reformulate the conceptual underpinnings of the model and its mathematical realization.

Aside from this introduction and a brief conclusion, this book is composed of four parts, each of which consists of two chapters. The chapters in Part I analyze the behavior of relatively mindful agents that can flexibly determine their own behavior (within certain limits).

Chapter 2 analyzes the relation between sickness, symptoms, treatments, and outcomes. It models the behavior of agents that need to decide whether they should take a medication, given the results of a diagnostic and the possibility of side effects. It also calculates the value of a sign by determining how much such agents would be willing to pay for the diagnostic, given their grounds. It argues that standard measures of information, such as Shannon entropy (which only takes into account the distribution of signs) and Kullback–Leibler divergence (which only takes into account the distribution of signs and objects), are wholly inadequate for understanding meaning. In their stead, it offers a way to measure the value of meaningful interrelations: one that takes into account signs, objects, interpretants, and consequents; and not just the ways that such components are causally interrelated, but also the ways that such interrelations are valued by an agent. And it argues that a good interpretive ground—one that adequately represents the patterns and values in an environment—is one of the most valuable goods. This chapter thereby retheorizes seemingly microeconomic processes in meaningful terms.

Chapter 3 studies the real-time dynamics of predators chasing prey in an open environment. In contrast to chapter 2, which turns on the discrete version of the main equation, this chapter uses the continuous version of the main equation: in effect, the coupled movements of predators and prey are simultaneously signs for each other to interpret, and interpretants of each other's signs. It analyzes the interactions of such agents in spaces of various dimensions and environments with various terrains. In so doing, it shows how this framework can be used to study real-time, coupled, and continuous interactions between two or more semiotic agents, as they take turns inhabiting the roles of signer and interpreter. It thereby studies the dynamics of relatively simple conversations between agents who are trying to either capture or elude each other. The main equation, in its continuous form, thereby functions as a kind of kernel that generates complex world lines, understood as the coupled movements of signers and interpreters through space-time.

Unlike the chapters in Part I, which focus on relatively mindful agents, and hence agents who can choose their own actions, as well as reason more generally, the chapters in Part II focus on agents that evolve on evolutionary time scales.

Chapter 4 introduces readers to the mathematical machinery of difference equations and critical points, as a key means for tracking the evolutionary

dynamics of such biosemiotic agents. After working through several simple scenarios, whereby agents evolve in relatively fixed environments, or evolve in dynamic environments constituted by other evolving agents, it works through the details of an extended case study. In particular, it analyzes a scenario in which two populations of agents inhabiting a relatively noisy, but value-rich, environment come to share a code. It measures the value of particular codes insofar as such codes better enable their users to communicate about salient features of their environments to cooperatively secure the values that are afforded by those environments. And it shows how different kinds of codes are more or less likely to emerge in different kinds of environments.

Chapter 5 returns to a classic scenario from evolutionary game theory: pairwise competitions over a valuable resource between relatively aggressive hawks and relatively passive doves. It shows what would happen to an otherwise stable population of hawks and doves if a particular kind of mutant is introduced: an agent that can semiotically ascertain—if only slightly better than chance—the identity of its competitor (hawk versus dove) and alter its behavior accordingly. It studies the conditions in which such a mutant is sieved out of the population completely, drives out one or both of its competitors, or comes to stably exist with a certain frequency in the population. In effect, this chapter studies the evolutionarily stable strategies of such social parasites as a function of their semiotic capacities.

Unlike the chapters in Part II, which focus on the evolution of semiotic agents, the chapters in Part III focus on semiotic agents that can learn on interactional time scales.

Chapter 6 models the behavior of agents that can improve the fit between their interpretive grounds and the environments they inhabit by internalizing the consequences of their past practices and adjusting their present practices accordingly. It reframes reinforcement learning and Edward Thorndike's law of effect in terms of the main equation. And it uses this framing to analyze several key scenarios: when agents must learn a set of purely arbitrary conventions to communicate with each other; when the conventions learned are relatively motivated, such that the code privileges certain signs over others; and when a parasite (third party or interloper) benefits from such conventions, such that the agents in question must repeatedly adjust their code to thwart it. It thereby explores the relation between nature and convention (or the motivated and the arbitrary) and the relation between codes and ciphers (or messages and secrets).

Chapter 7 models semiotic agents as neural networks that are able to learn from experience. It shows that the architectures of such networks, along with the values of the parameters they contain, are equivalent to interpretive

grounds. It thereby puts this model of meaning in dialogue with classic ideas from machine learning. After introducing readers to artificial neural networks and machine learning techniques, it uses those techniques to model a scenario in which relatively simple agents establish a shared code consisting of N arbitrary conventions. It argues that backpropagation—the key algorithm used to establish parameter values in a neural network—is a mode of meta-semiosis. And it shows the relation between language models (in a narrow sense) and models of meaning that are theorized in relation to models of reality. Loosely, if the former focus on word-word relations (in the guise of next-word prediction), the latter focus on word-world relations (in the guise of sign-object and interpretant-consequent mediation). It argues that the key capacity of language models (next-word prediction) is their main limitation: *worldlessness*.

Unlike the chapters in Parts I, II, and III of this book, which focus on case studies of revealing scenarios, whether they involve selecting agents or selected agents, the chapters in Part IV explore some of the deeper presuppositions and farther-reaching entailments of the model.

Chapter 8 introduces readers to the technical details of possible worlds, an analytic approach that is often used to model the semantics of natural languages. It introduces the idea of a possible world using standard conventions, and then it connects possible worlds to formal semantics, with a particular focus on modal operators (such as necessity and possibility). It relates possible worlds to this mathematical model of meaning, with a particular focus on Bayesian networks (or causal graphs) and expected value. In so doing, this chapter shows precisely what kind of causal graph was presupposed by our model, and it derives the main equation from first principles. Finally, this chapter shows the relation between times and worlds and uses this relation to understand the changing grounds of agents as they come to know—and transform—the worlds around them. In some sense, then, it offers an account of possible world semiotics, which is meant to encompass not only possible world semantics, but also the pragmatics of actual worlds.

Chapter 9 uses the main equation to model meta-semiotic processes, whereby the grounds of semiotic agents are figured, and thereby become the objects and consequents, and hence topics and ends, of semiotic practices. It shows how agents may update their interpretive grounds by changing not just their commitments and values, but also their ontologies and models, as a function of the feedback that they receive regarding the fit between their interpretive ground and the environment they inhabit. It also shows the ways that agents may come to model the interpretive grounds of other agents, as well as have their own interpretive grounds modeled in turn, to higher and higher levels of

embedding. It argues that culture may be usefully understood as an intersubjectively held interpretive ground, along with the semiotic processes, qua figured practices, that constitute its roots and fruits. And it further explores the range of relatively antagonistic agents—such as enemies, parasites, and noise—that exploit and degrade the otherwise shared interpretive grounds of others.

Finally, a brief conclusion lays out the relation between interpretive grounds and Max Weber's notion of an ideal type. It analyzes the relation between the interpretive grounds of analysts (who seek to model particular scenarios) and the interpretive grounds of the agents in the scenarios so modeled. Also, it highlights the productive tension between revelation and confrontation.

1.7 Relevant Literature

As discussed earlier in this chapter, this formulation of semiotic processes takes inspiration from Peirce (1984 [1867], 1986 [1877], 1986 [1878]). We borrow his notions of sign (or representatum), object, and interpretant; and add to them the notions of agent and consequent. And rather than treat interpretive grounds in terms of iconic, indexical, and symbolic relations, they are framed in terms of commitments, values, and ontologies. In this sense, we build on Peirce's notion of a guiding principle, while framing it in a way that explicitly takes quantities, or differential intensities, into account. Finally, as will become clear in the rest of this section, while Peirce had a deep understanding of core topics in probability and statistics, we make extensive use of mathematical ideas that either preceded him (like Bayes' theorem and expected value) or came after him (such as Bayesian networks and game theory).

The body of secondary literature on Peirce is enormous. Three particularly helpful guides to his arguments and interpretations of his ideas may be found in Colapietro (1988), Parmentier (1994), and Lee (1997). These authors also develop important connections between Peirce's ideas and classic works in linguistics, anthropology, and philosophy.

Kockelman (2013a) offers a more detailed presentation of the framework outlined in section 1.1. Kockelman (2017) extends this framework to think about Bayesian reasoning, Kockelman (2020a) uses it to analyze machine learning, and Kockelman (2024) uses it to understand large language models and artificial intelligence. In contrast to the analysis offered in this book, however, these works are more directly addressed to anthropologists, critical theorists, and scholars of science and technology.

For more on the slash, as framed through metaphors like bridges, horizons, and fog, see Kockelman (2010, 2016, 2024). For classic work on similar themes, see Whitehead (1920) on jagged edges (coming out of Peirce and Hume), Bateson (2000) on slash marks (coming out of Claude Shannon's

notion of redundancy), Roman Jakobson on the backslash, E^N/E^S, which relates the narrated event and the speech event (coming out of Otto Jesperson's notion of shifters), Thomas Bayes on the probability of a hypothesis conditioned on evidence, $P(H/E)$, and of course folks like Albert Einstein and Stephen Hawking on the contrast between the horizons of subjects and the world lines of objects.

The mathematical model offered here, as sketched in section 1.2, incorporates two relatively simple, yet powerful, mathematical ideas: Bayes' theorem (Bayes 1763; Laplace 1951; Jeffreys 1998) and expected value (Laplace 1951; Huygens 1714). Indeed, both ideas were already well developed by Pierre-Simon Laplace, whose celebrated text *Théorie analytique des probabilités* was published in 1814 (and thus about 25 years before Peirce was born). Hacking (2001, 2006) offers a particularly clear and inspired introduction to the philosophy of probability, the rise of statistics, and Peirce's relation to it.

While the main equation was introduced using expected value, it will be extended to encompass expected utility. And while our framing of certainty is left unspecified, we often presume subjective probabilities (as least when dealing with subjective grounds), following scholars like Ramsey (2016), de Finetti (2017), and Savage (1972). Savage's classic work, *The Foundations of Statistics*, which puts together Bayesian inference and expected utility and even offers a prescient account of the utility of information, is particularly relevant.

Other approaches to value, often highly critical of expected utility, include prospect theory (Kahneman and Tversky 2013; Levy 1992; Barberis 2013), Veblen-inspired theories of distinction (Veblen 2017; Bourdieu 1987, 1977), and ideas like strong evaluation (Taylor 1992), value rationality (Weber 2019), and higher-order desire (Frankfurt 1971). While our approach to value, at least in chapter 2, is relatively microeconomic, the framework is compatible with core ideas in substantivist economics (Malinowski 2013; Polanyi 2001; Sahlins 2013), as well as economic anthropology (Maurer 2011; Lee and LiPuma 2004; Lee 2020; Guyer 2004; Peebles 2010) and economic sociology (Lépinay 2011; MacKenzie, Muniesa, and Siu 2007). Kockelman (2007a, 2007b, 2015, 2020a) links the Peircean framework, as laid out in section 1.1, to many of these traditions.

This mathematical model is tightly coupled to diagrammatic representations, as intimated in the foregoing sections and as will be more fully fleshed out in chapter 9. For more on causal graphs, in relation to probabilistic reasoning, see the classic works of Sewall Wright (1921, 1934). Bayesian networks, as developed in chapter 8, are introduced and theorized in Pearl (1988)

and Neapolitan (1990). On auspiciousness versus efficatiousness, see Weinrich (2020) and the literature that he cites (Lewis 1981; Skyrms 1982). For a survey and synthesis of the large literature on probabilistic causality, see Hitchcock (2021); for an early classic work, see Reichenbach (1956). A standard work on agents, artificial intelligence, probabilistic reasoning, and decision theory is Russell and Norvig (2002). For an engaging introduction to the history of ideas in these traditions, see Pearl and Mackenzie (2018).

This approach to information, entropy, and organization is grounded in key ideas from thermodynamics and statistical mechanics (Reif 1965), as well as classic ideas in information theory from works like Shannon (1948), Kullback and Leibler (1951), and Lindley (1956). We also employ the definition of organization in relation to constraints that was offered by Brooks and Wiley (1988). Kockelman (2013b) compares and contrasts classic theories of information through a pragmatist lens.

Our model of natural selection, via game theory, has its origins in Smith (1982). See Nowak (2006) for a more recent survey and synthesis of the field. For thinking about difference equations, and their fixed points, Strogatz (2018) is a key resource. And the model of organism-environment interactions offered here is resonant with Lewontin's classic text (1983) and allied work, such as the essays collected in Oyama (2003).

Reinforcement learning goes back to Thorndike (1898), as developed by scholars like Bush and Mosteller (1955), Roth and Erev (1995), and Herrnstein (1970). Skyrms (2010) helpfully summarizes high points in the history of reinforcement learning, and deploys those ideas in his account of signals. Our approach to machine learning is grounded in the classic paper by Rumelhart, Hinton, and Williams (1986). Nielsen (2015) offers an elegant and approachable introduction to this fascinating topic.

Our account of possible world semiotics builds on the possible worlds literature (Lewis 1986; Carnap 1988), especially as elaborated by formal semanticists (Kratzer and Heim 1998; Von Fintel and Heim 2011). The work of Angelika Kratzer (2012) is particularly relevant. Arguments offered in chapter 8 build on her notions of necessity, possibility, and conversational backgrounds. The concept of a possible world goes back at least as far as Gottfried Wilhelm Leibniz, as described in his *Theodicy*, and later in his *Monadology*. For more on his ideas, in relation to the world in which they germinated, see Borowski (2024) and Weatherby (2016).

Biosemiotics is a thriving discipline. The essays collected in Favareau (2010) nicely showcase the key texts and rich conceptual field of its proponents and originators, as does the book by Barbieri (2008) and the programmatic essay by Kull, Deacon, Emmeche, Hoffmeyer, and Stjernfelt (2009). That said, while

the approach taken here is loosely aligned with these authors, insofar as it also takes Peirce, as well as von Uexküll (2013), as sources of inspiration, and insofar as it takes up the relation between semiotic processes and natural selection, the framing is mathematical rather than biological; and our overall conceptual approach is different as well. Moreover, we stick close to modern developments in mathematical biology, dynamic systems theory, ethology, linguistics, anthropology, cognitive science, and allied fields. Finally, as should be clear from this introduction, the model offered here applies to much more than biosemiotic processes.

For more on the difference between teleology, teleonomy, and teleomaticity, see Mayr (1974, 1992) and allied work (Enfield and Kockelman 2017). For classic accounts of animal signaling practices in relation to selecting processes, see Smith and Harper (2003) and the large discussion that it generated (Stegmann 2005). Millikan (2004) offers a particularly clear and influential discussion of meaning in relation to selection and the nature of function (or purpose) more generally.

Recent books by Deacon (2011) and Tomlinson (2023) are particularly resonant with this one, insofar as they approach meaning through a partially Peircean lens. They are, however, oriented toward the emergence of meaningful behavior in an extended historical sense (e.g., how mind might have emerged from matter, and/or what does and does not count as symbolic behavior as we look across all life forms). And they are qualitative accounts of such origins as opposed to mathematical models. Tomlinson offers a sharp critique of teleosemantic theories of meaning (such as those put forth by Millikan), teleodynamic theories of meaning (such as those put forth by Deacon), and relatively simplistic post-humanist theories of meaning (which currently abound in the social sciences and humanities). He also offers a very interesting distinction between meaning and information, and a detailed explication of the abstract machines that underlie the emergence of life. And both theorists utilize Peirce's trichotomy of icon, index, and symbol, which plays only a small role in this work. That said, this work is in agreement with Deacon, and especially thinkers like Kauffman (2003), Brooks and Wiley (1988), and Nelson (2004), that constraints are fundamental for thinking about entropy, information, and work. See, for example, Kockelman (2009). And it is allied with Tomlinson insofar as it is in dialogue with approaches in continental philosophy and critical theory, not just analytic philosophy and evolutionary biology. See, in particular, Kockelman (2011).

Other mathematical approaches to meaning, particularly resonant with this work, include Lewis (2008), Dretske (1981), and Skyrms (2010). Lewis offers a seminal account of the origins of conventions using game theory. Dretske

synthesizes information theory and core concerns in analytic philosophy, focusing on flows of information in relation to knowledge. Skyrms builds on both these thinkers by working out a sophisticated account of signaling processes in relation to information content, reinforcement learning, game theory, and social networks. Unlike Lewis and Dretske, however, he also engages in extended mathematical modeling, summarizing work that was undertaken by himself, his collaborators (Hofbauer and Huttegger 2008; Huttegger 2007; Huttegger et al. 2014), and other theorists.

While this book builds on these three theorists, as well as the ideas of Smith and Millikan, it differs in several important ways. First, it is grounded in pragmatism, anthropology, and critical theory as much as analytic philosophy, game theory, and evolutionary biology. In contrast to Skyrms, it focuses on rational agents and evolving agents as much as on learning agents. It takes into account value as much as meaning, and entropy in addition to information. Unlike Skyrms, who summarizes work done elsewhere, this work walks readers through the mathematical model and its results in detail. Its earlier chapters are designed to get readers up to speed on key ideas and techniques from game theory, possible world semantics, mathematical biology, and information theory. And finally, the actual model that it offers, both conceptually and mathematically, is different from their models (although, as will be seen, it can account for the same dynamics that interested them).

While this book does not offer a mathematical model of culture per se, it shows how its account of meaning can be used to operationalize culture, understood as relatively shared interpretive grounds (plus the semiotic processes, qua meaningful public interactions and social relations, that are both mediated by and mediating of such grounds). It thereby resonates with the classic work of Urban on metaculture (2001) and the motion of culture (2010), however different its conceptual framework and mathematical formalism are. And it is thus quite distinct from other mathematical models of culture, such as those put forth by Boyd and Richerson (2005) and McElreath and Boyd (2008). That said, it is partially grounded in similar tools as these last two works: game theory, evolutionary dynamics, difference equations, fixed points, and the like. And it should appeal to similar kinds of readers: those wanting a naturalistic and mathematical understanding of the emergence of conventions, values, social relations, and signs.

In short, by incorporating key insights and critiquing particular commitments across this wide range of literature, this model provides a relatively seamless integration of otherwise disparate topics, methods, and paradigms. The chapters that follow will apply this model to a series of distinct case studies as a means to explore its entailments and assess its merits.

I AGENTS THAT THINK