Models of Philosophy

Marcus Pivato

May 14, 2003

Contents

1 Introduction

(i) Reconceiving Epistemology

Philosophy is traditionally defined as 'The Search for Knowledge.' There is some contention about what exactly 'Knowledge' *is*, but it is generally agreed that one essential characteristic of Knowledge is *certainty*. Knowledge is not contingent; it is not provisional. You 'know' that a certain statement is true if it is *simply impossible for that statement to be false*. You 'know' things are a certain way if they *simply could not be any other way*.

The trouble is: all certainty is ultimately based upon an act of faith. To establish the certainty of assertion **X**, you really only have two options.

(1) Assert that **X** is 'self-evident'.

(2) Logically deduce **X** from other certainties

The problem with strategy **(2)** is obvious: the 'certainty' of **X** depends on the pre-existing 'certainty' of other statements. In trying to build an edifice of certainties, we inevitably either:

- Fall into an infinite regress of antecedent assumptions,
- Enter a self-justifying system of circular logic,

  or

- Ultimately invoke certain truths as 'self-evident' —ie. resort to strategy **(1)**.

Most philosophers ultimately resort to **(1)**, whether they invoke the 'natural light of Reason', refer to *a priori* truths, or are honest about it and simply admit to making a 'leap of faith'.

The justifications for the so-called 'self-evidence' of truths are not particularly compelling. They usually fall into two categories:

Appeal to Common Sense: Assert that 'Any fool can see **X** is true'.

Rejection of the Unimaginable: Assert that the negation of **X** is 'inconceivable'.

The *Appeal to Common Sense* fails because there is always at least one 'fool' who insists that, actually, he *can't* see that **X** is obviously true. There is actually little consensus in 'common sense'. What consensus exists is usually a product of culture; many of Plato's 'common sense' truths may strike modern readers as somewhat absurd.

The *Rejection of the Unimaginable* is simply a failure of the imagination. It is, in fact, a monumental hubris to believe that, simply because imagining something exceeds the limits

of our own puny cognitive faculties, it *simply cannot be*. Many formerly 'unimaginable' things often end up being true: the roundness of the earth, the nonexistence of an *entelechy* in living things, the physical basis for consciousness, the unity of space and time, and the curvature of both.

I don't believe anything can be known for certain. This may seem absurd, since many things seem 'certain': the reality of everyday events, the empirically verified facts of natural science, the truths of mathematics, the self-evident validity of logic, or the incontrovertible physicality of our own existence. 'How can you argue,' you may ask, 'That I do *not* know for certain that two plus two equals four?'

I do not argue it, however. I do not argue *against* someone's assertion of certainty —I simply do not believe it. It is absurd to demand that I must justify my *lack* of belief; the burden of proof is on the *maker* of the assertion, not upon the skeptic.

Of course, I am not seriously entertaining the possibility that $2+2=5$, or that my entire life is a hallucination. I am quite willing to *believe* in mathematics, science, and the evidence of my sense —but on a *provisional* basis. While not 'certain', these beliefs are definitely *useful*. They give meaning and structure to my existence, and provide a basis for decision-making; a basis which, so far, usually yields desirable results. This fact alone justifies my continuous, tentative acceptance of these beliefs.

Nor am I actively asserting, 'Nothing is certain, nor can anything be certain.' It would be self-contradictory to assert, with 'certainty', that nothing could be certain. Perhaps the best formulation is, 'I am not certain of anything, even this statement.' Perhaps it would even be responsible to add the word 'yet'.

This is a *pragmatic* epistemology. Constructing an edifice of certainty is (probably) impossible, but it is also *unnecessary*. It is sufficient to build a structure of beliefs which we *provisionally* accept as true, and which help us answer the questions and make the decisions which confront us.

(ii) Philosophy and Models

The most vociferous defense of the idea of certainty comes from the philosophers themselves. After all, if Philosophy is the search for Knowledge, and Knowledge is defined by Certainty, how can there by Philosophy if there is no Certainty? A rejection of the possibility of certainty seems to undermine the *raison d'etre* of philosophy.

However, rejection of certainty does not jeopardize the entire philosophical program. It only requires a slight reconception of philosophy. Philosophy still has at least three crucial roles:

To challenge and question: Incisive philosophical questions reveal our ignorance of what we thought we knew, and expose the incoherence or inconsistency of our beliefs. Even

(indeed *especially*) when they are unanswered, these questions clarify our thinking, and identify the limits of our understanding.

To clarify communication: A philosophical conversation may never resolve the question which started it. But it forces its participants to clearly express their ideas and precisely define their terms. By doing so, they achieve a deeper insight into the question, and come to better understand not only each other's beliefs, but their own.

To build models: Rather than peddling false assurances of certainty through 'answers' to philosophical questions, philosophy can offer *models*. A model is *not* a certainty; it is instead a well-defined, formal theoretical framework, which accords with our intuitions and real-world experience (as much as possible), and which generates (provisional) answers to philosophical questions.

It is this notion of a *philosophical model* which I will be primarily concerned with. A good analogy can be made with *scientific* models. Scientists long ago accepted the provisional nature scientific knowledge. Scientific theories are *temporary* constructions. They are mental approximations of reality, built to match the real thing as well as possible, but built always with the awareness that they may someday become obsolete. Scientific theories are predictively powerful, metaphysically comfortable, and pragmatically useful. They are also fundamentally disposable.

The job of a scientific theory is not to provide us with absolute Knowledge of the workings of nature. Instead, science serves several pragmatic purposes:

- Science structures our understanding of nature, providing us for a framework within which to build new theories, interpret data, ask questions, and design experiments.

- Science allows us to predict future behaviour of natural systems based on their present state –indeed, the basic test of any theory is how accurate its predictions are.

- Science makes possible the design and construction of technology, and the application of rationality to real-world decision making, thus providing us with practical ways of improving our lives.

- Science plays a metaphysical role for many people, by providing us with a sense of intrinsic order and structure –even *meaning* –in the cosmos.

Likewise, we should conceive of philosophy as the construction of *models* —models which are temporary, tentative, contingent, provisional, and disposable. A philosophical model should yield (provisional) answers to philosophical questions, but to be satisfactory, it should meet several other criteria:

Logical consistency: It must be impossible to deduce contradictory conclusions from the model.

Well-definedness: The terms of the model must be clearly and precisely defined, so there is no ambiguity in their interpretation.

Phenomenological compatibility: As much as possible, a model must be in accord with our everyday personal experience. When it deviates from this, it must have good reason. For example, Dennet's [10] 'Multiple drafts' model of consciousness is radically different than the 'Cartesian Theater' suggested by our subjective experience. But Dennet converges upon this unintuitive model only after he concludes that the Cartesian Theater is unsatisfactory for several reasons, and he shows how 'Multiple Drafts' can coherently account for what he calls 'user illusion' —our subjective experience of a single stream of consciousness.

Intuitive Plausibility: The model should agree with our intuitions. However, since our intuitions have internal contradictions, it will be impossible to perfectly match them with any logically consistent theory. Philosophical model-building will inevitably challenge our cherished intuitions; in the process of integrating them into some logical framework, we may be forced to revise them, restructure them, or even reject them. A good philosophical thought-experiment often confronts us with a fundamental incoherency in intuitions we previously took for granted.

Structuring Understanding: A good model provides an intellectual framework which clearly define the limits of our (provisional) knowledge. It tells us what questions we should be asking, and how we might go about answering them. By forcing us to structure our ideas in a formal context, the model clarifies the logical and cultural relationships between ideas, exposing inconsistencies and logical dependencies, and revealing patterns which may eventually become paradigms.

Theoretical Fecundity: The model should provides a language and/or methodology for the development of further theories. The model may ultimately be superseded, but perhaps its greatest contribution can be in helping formulate its own successors. For example, physicists depend heavily on the language and intuitions of classical mechanics when formulating quantum mechanics.

(iii) Working Assumptions

One advantage of this 'modeling' approach to philosophy is that it obviates many traditional epistemological problems. Rather than labouring in futility to lay a rock-solid epistemological foundation, we'll simply (provisionally) endorse an epistemological framework

which is minimal, practical, and seems reasonable, and then go from there. Let's make some working assumptions concerning the *a priori*, physical reality, and other minds.

The *a priori*: Let's accept the truth-preserving nature of logical deduction. Next, let's endorse some axiomatization of mathematics, and assume the applicability of this mathematics in describing the empirically observed patterns of personal experience. We use this framework because it seems to work. If it stops working, we may stop using it.

Physical Reality: I choose to believe that my sensory experience is the consequence of *some* independent physical reality, and not just a dream. This is a purely practical decision. I can't *prove* that an independent physical reality exists. However, even if I am dreaming, this dream behaves *as if* it had an independent reality (for example, I can't predict/control the unfolding of the dream), so I might as well treat it *as if* it was real. Also, frankly, the alternative (to dismiss life as a hallucination) is boring and depressing. Unless I had a compelling reason to decide I was dreaming (for example, persistent inconsistencies and discontinuities in my experience, or an inexplicable degree of control over my own reality), I would reject it on purely aesthetic/emotional grounds.

Notice that I make no assertions about the *nature* of this independent reality —for example, whether it is made from atoms and molecules, or is just a 'virtual reality' within a computer core. It is irrelevant (and unknowable) to me whether I am a 'brain in a body' or a 'brain in a vat'. It suffices to observe that the world behaves *as if* it was made out of atoms, and I experience it *as if* I was a brain in a body —thus, I choose to believe this to be the case. I can't prove it, but I don't need to.

Other minds: I believe in the existence of other minds because my sense-data suggests the existence of other beings with mental processes similar to my own. For example, I (seem to) have conversations with people which surprise and enlighten me. I often find myself thinking, 'I would never have thought of that.' I can't *prove* that the (apparent) participants in these conversations have minds like me own, but this hypothesis is an excellent *model* of their behaviour —certainly far better than any of the alternatives. For example, the hypothesis that other people have conscious minds like mine —with similar emotions and cognitive limitations —yields a surprising amount of success in predicting their behaviour in certain situations.

Clearly, these other minds are not identical to my own. However, they more similar to me than to any other phenomenon I experience. Modeling *people* as conscious entities akin to myself yields predictive success, whereas modeling (say) storm systems, automobiles, or trees as conscious entities does not.

(iv) Domains of Discourse

Philosophy can be conceived as 'personal science'. Science is a public endeavour which seeks to recognize and codify regularities and patterns in our (collective) experience of the world. Philosophy is a private endeavour, whereby I seek to to recognize and codify the regularities in my (personal) experience of reality.

Different scientific theories have different *domains of discourse*. Quantum theory describes the interaction of microscopic systems; classical mechanics applies to (low-energy) interactions of macroscopic systems; special relativity describes high-energy interactions in weak gravitational fields; general relativity, in strong fields. At the borderlands between these theories, they should agree; for example, the laws of classical mechanics can be deduced from quantum theory by taking the 'macroscopic limit'. However, things don't always work out: neither quantum theory nor relativity gives a good description of microscopic systems in strong gravitational fields. This means that the 'coverage' of our physical theories is *incomplete*, but it does not invalidate each theory within its own domain. Ultimately, of course, we want an 'umbrella' theory, which subsumes both quantum and relativistic physics. But even a fragmented theoretical edifice is useful.

This *demarcation of domains* is important, because it again allows independent development in different domains. It is not necessary to develop quantum and gravitational physics synchronously, and constantly ensure compatibility between the two.

In the same way, different philosophical models can address questions in different domains. Ideally, these models should agree at the common boundaries of their domains. If they disagree, it means that our theoretical edifice is imperfect; however, each model may still be valuable within its domain.

For example, since *language* is about communication between *minds*, a theory of linguistic semantics will be related to a theory of mental representation. However, we may find that our best model of linguistic semantics which is incoherent with our best model of mental representation. I will develop two different theories of linguistic representation in Chapters 2 and 5, which, though similar, are not identical. Nevertheless, each is useful within its own domain.

(v) Neural, Linguistic, and Mathematical models

A *model* is any mental representation of a pattern or regularity in our experience. In this sense, all of us unconsciously and continually construct and employ thousands of personal models of the world around us. You can walk only because, implicitly, your cerebellum contains a sophisticated unconscious model of the mechanics of bipedal locomotion. When your eyes follow a moving object across your visual field, they are employing a model of ballistic movement to instantaneously *predict* the location of the object in the very near future, and then *direct* the next saccade to fixate the eye on that point. When you intuit

that you have said something to upset your friend, you are employing the extremely complex, subtle, and almost entirely unconscious model of human psychology we call *empathy*.

These models are coded in the neural structures of the brain; they are unavailable to conscious introspection, and impossible to communicate to others. We could call them "prelinguistic", but this seems pejorative, suggesting that it would somehow be "better" if they were linguistic in nature. I will call them **neural models**. Other examples include reflexes and instincts, learned physical skills (ie. playing the piano, ballet-dancing), geometric/physical intuition (ie. a carpenter's intuition that a structure is stable, or a painter's ability to create the illusion of depth in a picture through the suggestion of perspective).

Neural models can be quite powerful, but they have disadvantages.

- They are unavailable to introspective examination. We can't understand how they work, so we can't understand why they fail. We can't intelligently correct or improve them, but instead must simply trust in the natural, unconscious learning mechanisms of the brain.

- Since their mechanism is mysterious, neural models do not structure our understanding or provide a basis for further theory generation.

- Neural models can't be communicated to other people. This makes it impossible to share insights, and difficult to reach consensus when the models disagree.

For this reason, we prefer **linguistic models** —that is, models which can be formulated in words, diagrams, tables, etc., and thereby communicated to other people. Indeed, it is fair to say that the recurring theme in the intellectual development of human civilization has been the replacement of inarticulate, unconscious, private *neural* models, with consciously articulated, publically accessible *linguistic* ones. When Socrates asks, 'What is Justice?', he is asking his friends to transmute their neural models of justice into linguistic ones. The response, 'I don't know what it is, but I know it when I see it,' is essentially an admission of failure in this endeavour.

Linguistic models transcend some of the limitations of neural models, but they have disadvantages of their own:

- The absence of a precise language for theory-specification often makes it difficult or impossible to precisely communicate the model to other people.

- Different people thus end up with different versions of the model, and thus, derive differing conclusions. Hence, initially unanimous 'schools of thought' inevitable schism into conflicting factions. In matters of theology or ideology, this often leads to war.

- Since we must reason 'linguistically' about linguistic model (ie. employ verbal dialogue or monologue as a problem-solving methodology), linguistic ambiguity can be magnified

into outright fallacy. The history of philosophy is filled with examples of this. One will suffice:

> *God is (by definition) the perfect being. Nonexistence is a form of imperfection, and God is perfect, so God cannot nonexist. Hence, God exists.*

This spurious ontological 'proof' exploits the ambiguity of the word 'perfect'.

Early science began with purely linguistic models (eg. the physics of Aristotle). However, to address the problem of linguistic ambiguity, scientists began to employ *mathematical models*. The best-known ancient example was Archimedes; the first modern example was probably Galileo.

A **mathematical model** is a linguistic model where the meanings of all the relevant terms are defined with perfect precision. This has several advantages:

Precise Communication: Mathematical language allows precise specification of the model, so it can be communicated with perfect fidelity.

Deductive Clarity: In a mathematical model, consequences can be deduced using purely logical arguments. All consequences of the theory must follow *tautologically* from the original premises. This renders unambiguous the validity or fallacy of any chain of reasoning within the model.

Metatheoretical: We can apply rigorous and powerful methods to analyze a mathematical model, to discover its flaws and limitations.

Predictive: Mathematical models yield precise, quantitative predictions. This facilitates the design of technology and the formulation of policy. Furthermore, these predictions are *falsifiable* in the sense that they can be unambiguously tested. A falsified prediction refutes the theory.

Virtually all modern science deals in mathematical models. Theories originally formulated in nonmathematical terms (eg. Darwin's theory of natural selection) are soon reformulated in mathematical ones (eg. Fischer's statistical population genetics). We'll look at mathematical models again in Chapter 7.

(vi) Mathematical Models in Philosophy

If philosophy, like science, is about constructing plausible models rather than uncovering incontrovertible proofs, then, like science, philosophy could benefit from the use of mathematics. Philosophy obviously does not seek to produce empirically testable predictions, but mathematical models have many other advantages, as mentioned above

Mathematics has already made some appearances in the philosophical discourse:

Bayesian Theory Confirmation uses Bayes' Theorem[1] to formulate a precise and compelling (though by no means perfect) account of how it is that an experimental result can 'confirm' the validity of a scientific theory.

The Probabilistic Account of Causality developed by Patrick Suppes [40].

Metalogic [25, 16, 41] is a branch of mathematics concerned with studying the limitations of mathematical reasoning itself, and yields results with consequences for epistemology, ontology, and the philosophy of language. Some results show that certain truths are forever beyond the ken of any deductive reasoning system:

Gödel's First Incompleteness Theorem [14] states that, for any (consistent) axiomatization of mathematics, there are *true* statements which are *unprovable* within that axiom framework.

Gödel's Second Incompleteness Theorem states that it is impossible to determine whether a particular axiomatization of mathematics even *is* consistent.

Formal Undecidability: After Gödel, many mathematical problems were shown to be *formally undecidable* —that is, unanswerable within the deductive framework of (standard) mathematics. One of the most well-known concerns mathematical models of computation called *Turing machines*. The nominal purpose of such machines is to carry out complex computations to answer certain questions. However, Turing showed that it is undecidable whether a particular machine will ever finish it's computation, or just keep grinding away forever [18].

Other results demonstrate the inescapable expressive limitations of any human language, by showing the existence of mathematical objects so 'big' or 'complex' that they can never be explicitly described.

[1]Bayes' Theorem says this: Suppose $\mathcal{A}$ and $\mathcal{B}$ are propositions whose truth or falsehood is unknown. Let $\mathbf{P}(\mathcal{A})$ and $\mathbf{P}(\mathcal{B})$ be their 'prior' probabilities of being true (ie. given no information). Let $\mathbf{P}(\mathcal{A}|\mathcal{B})$ be the probability of $\mathcal{A}$ being true, *given* knowledge that $\mathcal{B}$ is already true. Similarly, let $\mathbf{P}(\mathcal{B}|\mathcal{A})$ be the probability of $\mathcal{B}$, *given* $\mathcal{A}$. Then:

$$\mathbf{P}(\mathcal{A}|\mathcal{B}) = \frac{\mathbf{P}(\mathcal{B}|\mathcal{A})\mathbf{P}(\mathcal{A})}{\mathbf{P}(\mathcal{B})}.$$

The interpretation in the Philosophy of Science is this: suppose $\mathcal{A}$ is the truth of some theory, and $\mathcal{B}$ is some prediction made by that theory. Then, the probability of theory $\mathcal{A}$ being true, given that prediction $\mathcal{B}$ was valid, is $\mathbf{P}(\mathcal{A}|\mathcal{B})$. Since theory $\mathcal{A}$ implies prediction $\mathcal{B}$, we have $\mathbf{P}(\mathcal{B}|\mathcal{A}) = 1$. Thus, we can rewrite the above equation:

$$\mathbf{P}(\mathcal{A}|\mathcal{B}) = \mathbf{P}(\mathcal{A})/\mathbf{P}(\mathcal{B}).$$

Thus, if prediction $\mathcal{B}$ is, *a priori* a very unlikely event, but theory $\mathcal{A}$ predicts it anyways, then confirmation of prediction $\mathcal{B}$ will be strongly 'corroborate' theory $\mathcal{A}$, in the sense that the posterior probability $\mathbf{P}(\mathcal{A}|\mathcal{B})$, will be significantly greater than the prior probability $\mathbf{P}(\mathcal{A})$.

Russel's Paradox demonstrates that the *domain of discourse* of any logically consistent mathematical theory must be a strict *subset* of the 'set of all mathematical objects'. Any attempt to formally define the 'set of all sets' must result in contradiction.

Incomputable Numbers are real numbers which cannot be expressed or described using any finite string of symbols. I can't *give* you an example of such a number, because if I could *communicate* it to you in any way (or if I could even *conceive* of it mentally), then it would be *computable*, by definition. Nonetheless, Allan Turing demonstrated that such numbers not only exist, but actually form the *majority* of the real numbers.

Unreachable Cardinal Numbers are infinite quantities so large that it would take an infinite amount of time just to describe how large they are [25].

Social Consensus Theory [21, 34] addresses political and ethical questions using quantitative, analytic methods. The strategy is to formulate ethical/political issues as 'optimization' problems (akin to economics). One can then prove theorems about what solutions exist or don't exist. Two famous results:

Arrow's Impossibility Theorem [1] says (loosely) that there is no rational political system which is *Paretian* [2], and *independent of irrelevant alternatives*[3] which is not *dictatorial*, in the sense that a single person dictates all decisions.

Sen's Impossibility Theorem [39, 38] says (loosely) there there is no rational political system which is Paretian and provides absolute protection of individual rights (in the sense that there are certain decisions —involving, say, your person or property —over which you have absolute control).

Of course, these results depend upon very specific mathematical formulations of notions like 'rational political system', 'democracy', and 'individual rights', and these formulations are open to challenge, and a different formulation may yield different conclusions. Nevertheless, these results are exciting because they provide a precise formulation and rigorous justification of political assertions which have been debated for centuries.

Ethical Game Theory seeks to rationally justify altruism through Game Theory. Gauthier [13] has disputed the classical game theoretic conclusion that, in situations such as the 'Prisoner's Dilemma' (see Table 8.3 on page 90 of Chapter 8§(iii)), the most 'rational' strategy is that of amoral selfishness. Axelrod [3] and Danielson [9] have used

[2] This means that, if everyone prefers $\mathcal{A}$ to $\mathcal{B}$, then the system will always choose $\mathcal{A}$ over $\mathcal{B}$.

[3] The relative ranking of some third option $\mathcal{C}$ has no effect on the system's choice of $\mathcal{A}$ vs. $\mathcal{B}$

computer simulations of the 'iterated Prisoner's Dilemma' to empirically demonstrate that 'moral' players actually fare better than immoral players, in the long run.

Cognitive Science develops mathematical models of human cognition, and eventually, perhaps, of consciousness. A mathematical model is subject to mathematical critique. For example, the 'Strong AI' paradigm asserts, essentially, that 'Consciousness is computation' —that is, that the human mind *is* a Turing machine or similar computational device. This position has been attacked by Lucas [26] and Penrose [29] using the concept of *formal undecidability* discussed above, although this argument has been convincingly refuted by, for example, Hofstadter [17]

My goal in this book is to develop mathematical models to address some contemporary philosophical problems.

(vii) Organization and Overview

The chapters which follow are basically independent of one another, and may be read in any order, although it may be helpful to read Chapter 2 before Chapter 3. I have tried to keep the mathematical prerequisites to a minimum, but it has often been clearly advantageous, or even necessary, to employ certain mathematical concepts and terms. Because of this, I've included appendices explaining all the relevant terminology. My advice is to read these appendices on a 'need to know' basis.

In Chapter 2, I propose a model of lingistic and mental representation, and use this model to examine issues like the intersubjectivity of knowledge, the incommensurability of language, the reification of culture, and the 'ineffable'.

In Chapter 3, I argue that 'order' and 'disorder' are purely subjective concepts, which arise from our (necessarily) incomplete information about complex systems and our (necessarily) limited cognitive resources. Using a model of 'representation' similar to that of Chapter 2, I discuss the Second Law of Thermodynamics, the difference between energy as 'work' and energy as 'heat', and the relationship between (subjective) disorder and (subjective) complexity.

In Chapter 4, I develop model of language as a set of (probabilistic) constraints on the arrangement of symbols. I use this to examine whether language constrains our private thought and our collective cultural development. I conceive of computational *abstraction levels* as a form of language, and use this to critique the design of information technology.

In Chapter 5 I develop a model of linguistic/mental representation in terms of probabilistic correlations. I use this to examine the semantics of language and the intentionality of mental representations.

In Chapter 6 I examine the issue of *continuity of identity*, at both a personal and cultural level.

In Chapter 7 I develop a mathematical description of the nature of scientific models and scientific theories. I use this to discuss the nature of scientific inference, and distinguish between 'description', 'explanation', and 'prediction'. I then examine the empirical verification of theories, and discuss 'Occam's Razor'.

A careful examination of scientific practice is crucial if we conceive of philosophy as a 'quasiscientific' activity, which aspires to construct 'philosophical models' analogous to scientific models. This chapter begins such an examination, but leaves many questions unanswered.

In Chapter 8, I develop a game-theoretic model of human social, economic, and political interactions. I use this to examine concepts such as power, justice, freedom, and the stability of sociopolitical systems.

An interesting feature of this analysis is the importance of *semiotics* in understanding how the 'players' interpret each other's actions, and how they project their power through (nonverbal) communication. I sketch how the 'Social Game' model can accomodate real-world phenomena like advertising, propaganda, and diplomatic posturing (phenomena which are usually neglected in simpler models of politico-economic interactions between 'rational maximisers').

2 Representation, Perception, and Speech

> *In the long run, the most productive kinds of thought are not the methods with which we solve particular problems, but those that lead us to formulating useful new kinds of description.*
>
> —Marvin Minsky, *The Society of Mind*

(i) Representations as functions

How does perception work? A state of the physical world (the 'perceived') induces a mental state (the 'percept') in a person (the 'perceiver'). We are inclined to regard this percept as her 'mental representation' of the perceived worldstate.

How does speech[1] work? The speaker encodes her mental state ('intent') in some arrangement of 'signifiers' (verbal utterances, written symbols, pictures, gestures, facial expressions,

[1]...or any other form of communication.

movements of gaming tokens, etc.), which act as a 'linguistic representation' of her intent. This linguistic representation, when perceived by the audience, induces a percept: a 'mental representation' of the speaker's intent.

But what is *representation*? In the previous examples, representation seems to be the process whereby an element of one domain (eg. a physical state, a mental intent, etc.) is transformed into an element in another domain (a mental percept, an arrangement of signifiers, etc.). Mathematically speaking, a representation is a thus a *function*

$$f : \mathcal{X} \longrightarrow \mathcal{Y}.$$

Here, $\mathcal{X}$ is the space of things to be *represented*, and $\mathcal{Y}$ is space in which we *represent* them. For example, *perception* is a function

$$p : \mathcal{W} \longrightarrow \mathcal{M} \tag{2.1}$$

where $\mathcal{W}$ is the space of world-states, and $\mathcal{M}$ is the space of mental states of the perceiver. *Language* takes the form of a pair of functions:

$$\mathcal{M}_1 \xrightarrow{s} \mathcal{L} \xrightarrow{p} \mathcal{M}_2. \tag{2.2}$$

Here, $\mathcal{M}_1$ is the mental statespace of the speaker, and the function s represents the speech act. $\mathcal{L}$ is the space of all possible signifiers (eg. sentences, pictures, etc.), $\mathcal{M}_2$ the mental statespace of the audience, and p represents the process of linguistic comprehension

We can also compose these functions. For example, if Byron tells Catherine, 'There's an Airplane!', we can combine diagrams (2.1) and (2.2) to get:

$$\mathcal{W}_A \xrightarrow{p} \mathcal{M}_B \xrightarrow{s} \mathcal{L} \xrightarrow{p} \mathcal{M}_C.$$

where $\mathcal{W}_A$ represents the space of the world (or at least, the airplane), $\mathcal{M}_B$ is the mind of Byron, and $\mathcal{M}_C$ that of Catherine.

This crude model of representation suffers from two limitations:

- It is a model of the semantics of *complete assertions* (eg. 'It is raining') or *complete perceptions* (ie. a percept of a rain cloud). It does *not* address the semantics of individual words (eg. 'rain') or 'thought-fragments' (the 'idea of a rain cloud'). [2]

[2]Indeed, the 'meaning' of the word 'rain' is actually much more complex object than the meaning of a sentence like 'It is raining'. The meaning of 'rain' is entirely context-dependent (eg. 'rain or shine' vs. 'rain on your parade' vs. 'desert rain' vs. 'rain of fire', etc.) Arguably, the word 'rain' has *no* meaning, except in context. More precisely, the *meaning* of 'rain' is somehow an ensemble of *all* the meanings the word assumes in various contexts.

- This model does not explain *how* a sentence represents a mental state. To put it another way, it does not tell us *why* we are justified in asserting that the sentence, 'I am sad', represents a mental state of sadness (as opposed to the mental idea of formal undecidability, or the melody of Bach's Passacaglia and Fugue in C Major).

I will address these issues in Chapter 5. The goal of the present chapter is to apply the above crude 'function' model of representation to explicate issues such as intersubjectivity, ineffability, and incommensurability.

Representation vs. Causality In the example of perception and speech, the worldstate w 'causes' the percept $p(w)$, and mindstate m 'causes' the speech act $s(m)$. However, representation is not always causal in nature. For example, consider the following kinds of mental representation:

Recollection of past worldstates ($\mathcal{W}_{\text{past}} \longrightarrow \mathcal{M}_{\text{now}}$).

Reverie over past mindstates ($\mathcal{M}_{\text{past}} \longrightarrow \mathcal{M}_{\text{now}}$).

Prediction of future worldstates ($\mathcal{W}_{\text{fut}} \longrightarrow \mathcal{M}_{\text{now}}$).

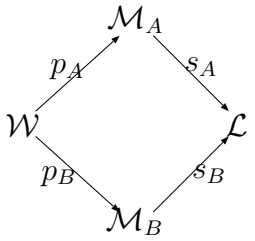
Anticipation of future mindstates ($\mathcal{M}_{\text{fut}} \longrightarrow \mathcal{M}_{\text{now}}$).

Empathy ($\mathcal{M}_{\text{other}} \longrightarrow \mathcal{M}_{\text{self}}$).

In these cases, the representation function cannot be so clearly identified with a causal process.

(ii) Translation and Intersubjectivity

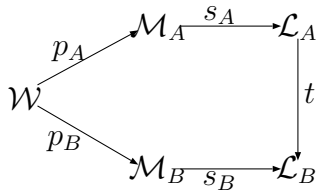
Each person speaks a slightly different 'dialect' of English. The discrepancies between these dialects often result in misunderstandings, even when communication seems clear. Aletheia and Byron can be said to speak 'exactly the same' English dialect if the following diagram commutes:



In other words, the same worldstate $w \in \mathcal{W}$, perceived by both Aletheia and Byron, may lead to different mental percepts $p_A(w)$ and $p_B(w)$, but ultimately yields the same speech act:

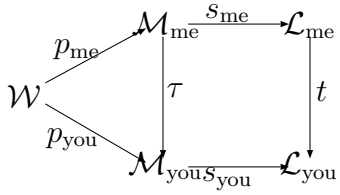
$$s_A\Big(p_A(w)\Big) \quad = \quad s_B\Big(p_B(w)\Big)\,.$$

Even if Aletheia and Byron speak different dialects, we should be able to translate between them. A *perfect translator* is a bijection $t : \mathcal{L}_A \longrightarrow \mathcal{L}_B$ so that the following diagram commutes:



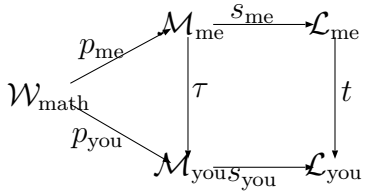
It is not clear, *a priori*, that such a translator exists, or, if it does, that it can be identified. Even if a perfect translator exists, it only tells us that Aletheia and Byron make intertranslatable declarations about the world. It does not tell us that they actually *think alike*.

All human minds are different. But many philosophers assume that, underneath superficial differences, we are all pretty much the same. This is the assumption of *intersubjectivity* —that my mental states are, in principle, translatable into your mental states. We can represent it with a commuting diagram:



where $\tau : \mathcal{M}_{\text{me}} \longrightarrow \mathcal{M}_{\text{you}}$ is the translator of mental states. This does *not* mean that you and I are the same person, or that we hold the same beliefs; it just means that, in principle, the 'mental vocabulary' within which I represent my thoughts can be perfectly translated into your 'mental vocabulary'.

Intersubjectivity seems like a reasonable assumption. It is certainly the experience of mathematicians; indeed, the strongest argument for the 'independent existence of mathematical objects' is the commuting diagram experienced by every mathematician who has ever communicated with a colleague:



(To avoid issues concerning the ontology of mathematical entities, let's interpret $\mathcal{W}_{\text{math}}$ as being the 'world of mathematical literature', rather than some metaphysical 'world of mathematical objects'. Hence, this diagram should be interpreted as, for example, two mathematicians inspecting the same (written) proof, and independently reaching the same conclusion about its correctness.)

Different mathematicians may think in different ways, and perceive the 'mathematical universe' differently, and express their ideas using different notations, but ultimately, mathematical statements and ideas —if true —are intertranslatable. Similarly, the collective endeavour of any scientific community is based upon the presumed intersubjectivity of the perception of physical reality.

Because of these successes, we are inclined to carry the assumption of intersubjectivity into the philosophical domain. A philosopher theorizes based upon her personal experience, but she presumes that the self-evident truth of her observations and conclusions will be clear to anyone who hears and understands her. When her ideas are rejected, it must be due to communication breakdown.

However, it is also possible that, when we depart from the 'objective' realm of mathematics or physical science, intersubjectivity breaks down. Perhaps different people have fundamentally different experiences of their own consciousness, or time, or sense-data. This is a fact which philosophers should keep in mind when we debate the nature of consciousness, etc. There may not be a single right answer.

(iii) Expressive Domains: Boolean Algebras

I have spoken of $\mathcal{W}$ as the 'space of worldstates'. A point $w \in \mathcal{W}$ is thus a 'complete specification' of a worldstate, down to the position and motion of ever atom and molecule[3]. When you describe a 'state of the world', however, you never speak with such exactitude. When you believe, 'It is raining in Toronto', you are in fact mentally referring to a very large *set* of worldstates; the set of all $w \in \mathcal{W}$ such that, in the worldstate w, rain is falling on Toronto.

[3]I am being deliberately vague here. The exact meaning of points in $\mathcal{W}$ depends upon a physics model. In a classical model, a point in $\mathcal{W}$ exactly specifies the position and momentum of every particle in the universe. Thus, in a universe of N particles, $\mathcal{W}_{\text{class}} = \mathbb{R}^{6N}$. In a quantum model, elements of $\mathcal{W}$ would be *wavefunctions*: complex-valued 'probability distributions' over all possible classical states. Thus, $\mathcal{W}_{\text{quant}} = \mathbf{L}^2(\mathcal{W}_{\text{class}}; \mathbb{C})$. However, the detailed nature of $\mathcal{W}$ is irrelevant for this discussion.

Similarly, the elements of $\mathcal{M}$ correspond to exhaustive descriptions of mental states, down to the level of every calcium ion in every neuron[4]. A 'mental state' in everyday parlance, such as, 'I feel sad,' does not point to a single point in $\mathcal{M}$, but instead, a large subset of $\mathcal{M}$.

In short, the *referent* of a mental representations is not an *element* of $\mathcal{W}$, but a *subset*. Likewise, the referent of a linguistic representation is a subset, not an element, of $\mathcal{M}$. This is inevitable. The world is much more complex than our minds can appreciate. And the subtlety of our thoughts often eludes our crude language. In representing the world in our thoughts, or our thoughts via language, we gloss over distinctions and obliterate information. For example: the space $\mathcal{W}$ is probably uncountable, whereas $\mathcal{L}$, being a collection of finite sequences in some finite alphabet, is necessarily countable. The 'size' of $\mathcal{M}$ is hard to judge, but it is probably much smaller than $\mathcal{W}$, and larger than $\mathcal{L}$. Hence, the 'perception' function $p : \mathcal{W} \longrightarrow \mathcal{M}$ and the 'speech' function $s : \mathcal{M} \longrightarrow \mathcal{L}$ must necessarily be many-to-one.

The *meaning* of a speech act $\ell \in \mathcal{L}$ is the preimage set $s^{-1}\{\ell\} := \{m \in \mathcal{M} \; ; \; s(m) = \ell\}$. Thus, really, the language $\mathcal{L}$ does not allow us to precisely specify *points* in $\mathcal{M}$ (ie. complete mental states) but only *subsets* of $\mathcal{M}$. The collection of all subsets we can describe using $\mathcal{L}$ is the **expressive domain** of the language. Let us call this collection $\mathcal{D}$.

Next, let's suppose that the language $\mathcal{L}$ possesses the capability to express logical *conjunction*, *disjunction*, and *negation*. Thus, if λ and ℓ are both elements of $\mathcal{L}$, then so are "λ **and** ℓ", "λ **or** ℓ", and "**not** λ".

The proposition "λ **and** ℓ" obviously corresponds to the set $s^{-1}\{\lambda\} \cap s^{-1}\{\ell\}$; the proposition "$\lambda$ **or** ℓ", to $s^{-1}\{\lambda\} \cup s^{-1}\{\ell\}$; and "**not** λ", to $\mathcal{M} \setminus s^{-1}\{\lambda\}$. If $\mathcal{D}$ is a collection of subsets of $\mathcal{M}$ closed under the operations of intersection, union, and complementation, then $\mathcal{D}$ is called a **Boolean algebra**[5]. The elements of $\mathcal{D}$ are **linguistic categories**; they are the collections of mental states which one can specify within the framework of $\mathcal{L}$.

In the same way, the perception function $p : \mathcal{W} \longrightarrow \mathcal{M}$ yields a Boolean algebra $\mathcal{D} \subset \mathcal{P}(\mathcal{W})$ of subsets of $\mathcal{W}$. These are **mental categories**, the collections of worldstates which one can mentally represent.

(iv) Sense vs. Referent

In his analysis of linguistic semantics, Frege [12], distinguished between **sense** and **reference**. He used the example of *the Morning Star and the Evening Star*. We know that the terms 'Morning star' and 'Evening star' *refer* to the same object: the planet Venus. However, the *sense* of the two terms is different, since they describe different *subjective experiences*. The first describes an experience which takes place at dawn, while facing east; the

[4]I am being deliberately vague here. The precise meaning of points in $\mathcal{M}$ depends upon a model of mentation. For example, in a dualist model, elements of $\mathcal{M}$ would not correspond to physical states at all. Again, the detailed nature of $\mathcal{M}$ is irrelevant for us.

[5]See Appendix D.

other, an experience at dusk, facing west. This experiential difference historically engendered the belief that the Morning Star and the Evening Star were in fact two different entities.

In terms of the formulation I've presented here, the *sense* of a speech act $\ell \in \mathcal{L}$ is its preimage $s^{-1}\{\ell\}$, a subset of $\mathcal{M}$. In other words, if ℓ is the sentence, 'I see the morning star', then the *sense* of ℓ is the set of all *mental* states represented by these words, which, I expect, involve subjective experiences of early dawn, looking east, etc.

The *referent* of ℓ is the double preimage $p^{-1}s^{-1}\{\ell\}$, a subset of $\mathcal{W}$. In other words, the *referent* of 'I see the morning star' is the set of physical states where the speaker and the planet Venus are in a certain relative position, it is early dawn, the sky is not overcast, etc.[6]

(v) Ineffability

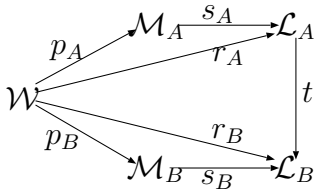
What are we to make of the claim that some ideas are simply **ineffable**, or incapable of linguistic expression?

If $\mathcal{D}$ contains *all* subsets of $\mathcal{M}$ —ie. $\mathcal{D} = \mathcal{P}(\mathcal{M})$, the *power set* of $\mathcal{M}$ —then it is possible, in principle, to specify *any* subset of $\mathcal{M}$ using $\mathcal{L}$. In other words, it is possible to articulate any mental state, or set of mental states, in the vocabulary of the language. However, it is more likely that $\mathcal{D}$ is only a small subset of $\mathcal{P}(\mathcal{M})$. This means that there are mental states which simply *cannot* be precisely specified in the language $\mathcal{L}$. Such states might be called "ineffable".

Artists and mystics report this phenomenon all the time, but the mental states they report are transcendental ecstacies which a logical positivist might dismiss as 'contentless'. But perhaps what is true in art is also true in philosophy. Perhaps it is possible to develop a mode of understanding the world which it is simply impossible to convey in words. Because of this, we should be wary of *logocentrism*, the assumption that all truths can be conveyed using language.

(vi) Incommensurability

Suppose Aletheia and Byron are scientists arguing about physical reality. Their different perceptions and terminologies may inhibit communication —especially if they have different areas of expertise. We seek a translator t yielding a commuting diagram:



[6]Note that I've described the sense and referent of an entire *sentence*, 'I see the morning star', rather than the sense and referent of the *word*, 'morning star', which was Frege's original concern. Again, it's easier to pin down the meaning of an entire *sentence* than of an isolated *word*. See note 2.

Here, $r_A = s_A \circ p_A$ and $r_B = s_B \circ p_B$. Let $\mathcal{D}_A$ and $\mathcal{D}_B$ be the Boolean algebras induced in $\mathcal{W}$ by r_A and r_B respectively.

If $\mathcal{D}_A = \mathcal{D}_B$, then there is no real difficulty. Once Aletheia and Byron discover that they are simply using the same words to mean different things, they can find some translation scheme t, and achieve clear communication.

However, what if $\mathcal{D}_A \neq \mathcal{D}_B$? Then there are certain things that Aletheia can express in her language which *simply cannot be translated* into Byron's language. This is the phenomenon of **incommensurability** between languages. Aletheia's language and Byron's are *incommensurable*, because ideas expressible in one are not expressible in the other, and vice versa. This creates major difficulties, if, for example, Aletheia is a proponent of one scientific theory, Byron a proponent of a rival theory, and they are trying to discuss the relative merits of their theories. This incommensurability between scientific languages, according to Kuhn[24], can only be resolved through a scientific revolution

Linguistic incommensurability separates academics in different fields, making it hard for specialists to appreciate the value —or even the coherence —of research outside their field of expertise. The gulf is especially wide between the sciences and the humanities. Linguistic incommensurability also thwarts resolution of ideological debates. It is easier to demonize one's opponents then make the radical mental shift necessary to appreciate the internal logic of their strange ideas.

(vii) Culture and Reification

Mathematician and playwright John Mighton[27] once facetiously suggested a 'Meaning Decay Coefficient': a mathematical measure of the rate at which a new word, introduced into the popular lexicon, semantically decays over time. From initial precision, the word's meaning inevitably dissolves into vagueness and ambiguity, finally becoming devoid of semantic content. When new ideas —artistic, philosophical or political —are digested by mass culture, they rarely remain in their original form. They are misconstrued, reinterpreted, *reified*. Often, an idea intended to *subvert* status quo ideologies ends up being co-opted to support them. For example:

- During political revolutions (eg. France, Russia), ideals of 'liberty' and 'equality' are used to justify the creation of a postrevolutionary tyranny even more oppressive and totalitarian than the *ancien régime* it supplanted.

- Initially 'anti-Establishment' musical/cultural movements (rock, punk, alternative) soon become the 'Establishment'. An 'anti-conformist' subculture engenders a cultural norm to which people must *conform* to attain social acceptance. 'Anti-consumerist' music is commodified by music companies, and sold as records, posters, and souvenir tee-shirts. 'Anti-fashion' becomes fashionable.

- A Westernized pastiche of foreign culture (eg. Chinese cuisine, Arabic music, Zen Buddhism) often attains far greater dissemination and public recognition than the authentic original. For example, Said[35] critiques the Western European construction of 'the Orient', a construction which is more familiar and far more 'real' to most Westerners than the actual cultures it represents.

This phenomenon of reification can be explained in terms of incommensurability.

The commensurability between the languages of two people depends upon the similarity of their values and worldviews, which, in turn, is a function of cultural background. Indeed, commensurability partially *characterizes* culture: people belonging to the 'same culture' generally have commensurable languages. The more incommensurable their languages become, the greater the 'cultural differences' between them.

Most members of a given culture, then, have roughly the same way of representing the world, both linguistically and mentally. What happens if a radical new idea is introduced into this culture? How will these individuals assimilate this idea? How will the culture assimilate it?

Suppose you are trying to explain a radical new idea to me, one which is *ineffable* within my mental representation system. If I am mentally flexible, and I *want* to understand, then perhaps I can gradually evolve my mental representation system to accommodate your radical idea, inventing what might be called a new "mental category" to represent it. I might start by trying to *approximate* this new category using some collection of existing mental categories. At first, this approximation will be crude, but over time, as my mind adapts (ie. as my Boolean algebra of mental categories grows or changes) I will (hopefully) develop in my mind a good approximation of the idea you tried to convey. This is called *learning*.

Learning requires active mental effort, intelligence and mental flexibility, and finally, it requires *time*. But suppose that, through impatience, laziness, or stupidity, I instead develop a gross misconception of your radical idea. I have reformulated it in terms which *I* understand, but have bastardized it in the process. I have approximated an element of $\mathcal{D}_{\text{you}}$ by some element of $\mathcal{D}_{\text{me}}$, and the approximation is not a good one.

Now, I go away and try to explain this idea to other people of a similar cultural background. They find my bastardized version easy to internalize in their Boolean algebra of mental categories —much easier than your original radical idea.

In articulating your idea to me, you have lost control of it. I and others are free to misconstrue and misrepresent what you have said. Furthermore, since 'similar culture' means 'similar mental representation scheme', different members of the same cultural group are likely to misconstrue the idea in roughly the same way, and then, in discussions with one another, *reinforce* their common misunderstanding. Over time, these misconceptions often converge upon a sort of equilibrium, an 'attractor' in the space of mental categories. This attractor represents mass culture's attempt to approximate the radical idea using a conceptual vocabulary which is simply inequipped to properly express it.

This is **reification**: the process whereby a culture appropriates and bastardizes a radical idea, and reduces it to something which is recognizable, but facile and inauthentic. Reification is particularly problematic in an information society, where ideas is are transmitted and retransmitted far more rapidly than they can be properly assimilated. This distortion can affect the meaning of a single word, a complex idea, a piece of art, or an entire cultural movement.

The problem is exacerbated by the deliberate efforts of commercial advertising and political propaganda to manipulate mass culture. These 'cultural engineers' often attempt to transform the meaning of a word like 'freedom' in order to manipulate consumer demand or political opinion.

Notes

The model of representation in this chapter seems similar to the 'tower bridge' model of Egan et al. [8, 11]. The differences are twofold:

1. In the 'tower bridge' model, the 'interpretation function' maps *from* the space of mental states *into* the space of real world states. In my model, the map goes in the opposite direction, and this is key.

2. The 'tower bridge' interpretation function takes individual concepts (eg. the number '2', or the idea of your friend Alvin) as input, and outputs their meanings; relationships between concepts (eg. '2+2=4', 'Alvin loves Bob') are then mapped to the corresponding relationships between their meanings through a sort of 'functorial' property of the representation function. In my model, the function takes *entire worldstates* as input (rather than pieces of them, such as the image of Alvin or Bob), and maps them to entire mental states (not just concepts, like the idea of Alvin or Bob)

3 Order and Disorder

> *According to convention there is a sweet and a bitter, a hot and a cold. According to convention, there is an order. In truth, there are atoms and a void.*
>
> —Democritus, 400 B.C.

'Order' and 'disorder' are subjective notions. I judge a system 'orderly' when I perceive a pattern or structure that *I* recognize —in other words, when the structure of the system fits neatly into one of *my* mental categories. Consequently, my perception of 'order' depends upon my own idiosyncratic vocabulary of mental categories. I perceive a system as 'disorderly' simply because its structure is foreign to *my* framework of mental categories.

But what of the well-known **Second Law of Thermodynamics**? The Second Law **(2LT)** is usually formulated:

> Isolated physical systems *must* proceed inexorably towards a state of maximum disorder.

Unfortunately, this formulation is misleading, and the meaning of **(2LT)** is often misconstrued. Since it is considered incontrovertible Physical Law, **(2LT)** is then used spuriously to 'deduce' all sorts of fallacies. The misconceptions about **(2LT)** are twofold:

- That **(2LT)** is a deterministic principle (hence the word 'inexorably'), which is true with absolute certainty (hence, the italicised *'must'*).

- That **(2LT)** is an *objective* statement about the *actual* state of the physical system.

(2LT) is *not* a deterministic or absolute statement. It is a *probabilistic* statement. Nor is it a statement about the *actual* state of a physical system —rather, it is a statement about the 'perceived' state. A more accurate (but less snappy) reformulation of **(2LT)** reads:

> With extremely high probability, an isolated physical system will proceed towards a state of maximum *perceived* disorder, where (with extremely high probability) it will remain indefinitely.

The 'extremely high probability' here is, for most macroscopic physical systems, so close to probability one that, we can, for all practical purposes, consider it to be absolute certainty. Nonetheless, there is always extremely small (but nonzero) probability that the system will deviate from 'maximum disorder'. The probability of this event is is so tiny that it can, for all practical purposes, be regarded as 'impossible'. It is comparable to the probability that one thousand monkeys banging on typewriters will accidentally produce *Hamlet*, or the probability of flipping a fair coin and coming up 'heads' one million times in a row. But it is not *absolutely* impossible, and this must be understood if one is to properly understand **(2LT)** .

The concept of 'maximum *perceived* disorder' is a bit more slippery. The purpose of this chapter is twofold:

- To explain the idea of 'perceived' order/disorder.

- To explain why **(2LT)** is a natural consequence of the nature of perception.

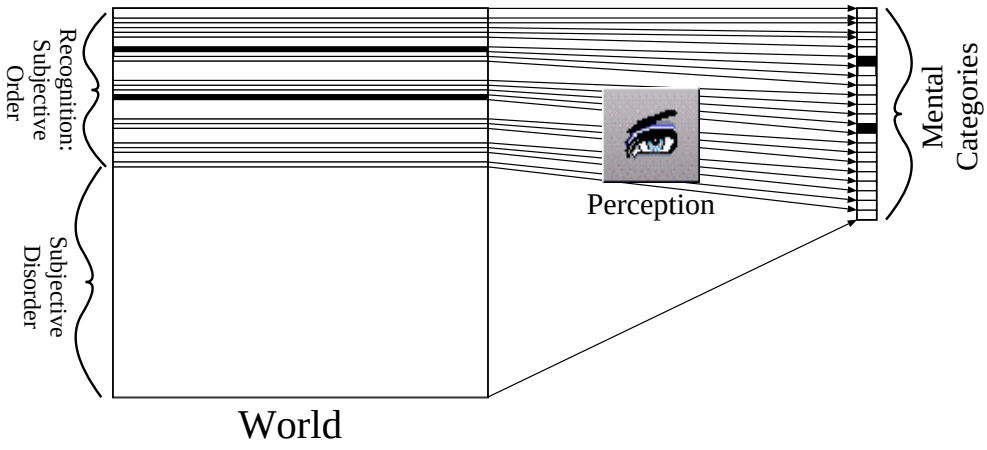


Figure 3.1: Perception divides up the space of worldstates into pieces corresponding to mental categories.

(i) Microstates vs. Macrostates; Noumena vs. Phenomena

In Chapter 2, I introduced a model of perception as a function $p : \mathcal{W} \longrightarrow \mathcal{M}$, where $\mathcal{W}$ is the statespace of the world, and $\mathcal{M}$ the mental statespace of the perceiver, who we will call Persephone. Each element $m \in \mathcal{M}$ corresponds to the a specific mental *percept*, and the preimage $\mathcal{C}_m = p^{-1}\{m\} \subset \mathcal{W}$ is the set of all world states 'recognized' by Persephone as that particular percept. Two worldstates w and w' in $\mathcal{C}_m$ are *indistinguishable* to Persephone; they generate the same percept in her mind, and thus, they are perceived by her as 'identical'. We can think of $\mathcal{C}_m$ as a *mental category*. Let's consider a couple of examples.

Example: *Water*

Consider a glass of water. The glass contains approximately 10^{23} water molecules, in constant motion. Even if your eyes were sharp enough to see the individual molecules, your mind could never simultaneously apprehend the positions and velocities of all of them simultaneously. You have no knowledge of the 'microscopic' properties of molecules, but only of 'macroscopic' properties, like the fact that the water is cold at the bottom of the glass but warmer at the top, or the fact that it sloshes around when you stirs it with a spoon.

This leads physicists to distinguish between the **microstate** of the glass of water (a precise specification of the position and velocity of each molecule, which requires about 6×10^{23} variables), and the **macrostate:** large-scale observable properties like a temperature gradient or aggregate motion (sloshing).

The microstate/macrostate dichotomy is somewhat analogous to Kantian dichotomy of *noumena* vs. *phenomena*. In terms of the model of perception we have developed, microstates correspond to *worldstates* (in $\mathcal{W}$), and macrostates correspond to *percepts* (in $\mathcal{M}$). Each

macrostate thus represents a very large collection of microstates. The act of *perception*, in this context, is called *measurement*.

In principle, we could formulate a physical model of the water in terms of its microstates, but this would be useless: we can't directly observe microstates, and even if we could, we can't compute formulae with 6×10^{23} variables. Instead, we must formulate a theory in terms of macrostates. Such a theory must necessarily be *probabilistic* in nature, since we are only working with 'approximate' (ie. macroscopic) information about the system. We hope to identify *statistical regularities* in the evolution of the macrostates. This is the basic paradigm of statistical mechanics, of which classical thermodynamics is one branch. The aforementioned Second Law is one of these 'statistical regularities'. The miracle of statistical mechanics is this: *As the number of particles in a system becomes very large, certain statistical regularities become near certainties.*

Example: *'Random' Numbers*

Consider the three following 30-digit numbers:

$$\left.\begin{array}{l} 01234\ 56789\ 01234\ 56789\ 01234\ 56789; \\ 00005\ 00040\ 00300\ 02000\ 10000\ 00000; \\ 010101\ \ 010101\ 010101\ 010101\ \ 010101. \end{array}\right\} \tag{3.1}$$

In each case, you recognize a pattern, which allows you to represent the entire number in your mind at once. Hence, your 'percept' of the number is an exact representation of it. Now consider the 30-digit random number:

$$97238\ 33463\ 64832\ 39798\ 53562\ 95141 \tag{3.2}$$

You think you see it, but you don't, really. Your mental percept cannot immediately and simultaneously represent the totality of this number. For example, you probably couldn't copy it down without repeatedly checking against the original. Eventually, if you studied (3.2), you might memorize it. You might begin to spot patterns —or perhaps, *invent* patterns —and use them as mnemonics.

However, even if you memorized (3.2), it wouldn't help you 'recognize' another random 30 digit number (except one closely related). There are 10^{30} such numbers —no matter how many you memorized, you couldn't ever become personally acquainted with more than a tiny minority. The three examples (3.1) —and all other numbers with recognizable patterns —are part of this tiny minority. The vast majority of the 10^{30} numbers have no recognizable pattern. They thus appear 'random'.

However, the key phrase is '*recognizable* pattern'. For example, consider the well-known digits sequence

$$\pi \ = \ 3.14159\ 26535\ 89793\ 23846\ 36433\ 83279\ldots$$

Now look at (3.2) again. It is just the first 30 digits of π, written backwards, excluding the leading '3'. Suddenly, it is no longer 'random' at all, but instead, highly ordered.

These examples show that the world is much vaster and more complex than a human brain. Hence, $\mathcal{W}$ is a much larger space than $\mathcal{M}$, and the map $p : \mathcal{W} \longrightarrow \mathcal{M}$ is necessarily many-to-one. In other words, for most percepts $m \in \mathcal{M}$, the corresponding mental category $\mathcal{C}_m$ must contain many elements.

(ii) Information content in mental categories

If $\mathcal{C}_m$ is a small subset of $\mathcal{W}$, then the percept m provides highly 'specific' information about the world; if Persephone perceives m, then she has an almost complete specification of the worldstate w which generated that percept. On the other hand, if $\mathcal{C}_m$ is a *large* subset of $\mathcal{W}$, then the percept m provides only vague and nonspecific information about w. To understand this, consider the following assertions[1] about the location of a buried treasure:

(m_1) 'It is somewhere on planet Earth.'

(m_2) 'It is buried under the Royal Ontario Museum, in Toronto, Canada.'

(m_3) 'It is buried exactly 4.3 metres below a spot which is 17.1 metres west and 23.7 metres south of the intersection of Bloor street and Avenue road.'

In this example, w is the location of the buried treasure (so that $\mathcal{W}$ is the space of all possible locations —say, the space of all points on earth). Percept (m_1) is extremely vague; $\mathcal{C}(m_1) = \mathcal{W}$ is huge, so this tells us nothing. Percept (m_2) is more specific, and $\mathcal{C}(m_2)$ is a smaller subset of $\mathcal{W}$; the set of all $w \in \mathcal{W}$ where the treasure is underneath the Museum. Percept (m_3) is the most specific, and $\mathcal{C}(m_2)$ is a very small subset of $\mathcal{W}$.

As this example illustrates, *the information content of a percept m is inversely proportional to the size of $\mathcal{C}(m)$*. Hence, if Persephone wants specific information about her world, she wants, whenever possible, to perceive percepts whose corresponding mental categories are as small as possible as subsets of $\mathcal{W}$.

(iii) The importance of being biased

Now Persephone has a problem. She can only divide up $\mathcal{W}$ into a limited number of mental categories (corresponding to the size of $\mathcal{M}$), but she wants most of these mental categories to be *small* (ie. high in information content). At the same time, they must together cover all of $\mathcal{W}$, which is large.

One solution is to choose an 'equitable' perception scheme where all mental categories are of equal size (Figure 3.2A). The problem is that, in this case, *all* the mental categories

[1]For the purposes of this example, I will gloss over the distinction between (linguistic) assertions and (mental) percepts.

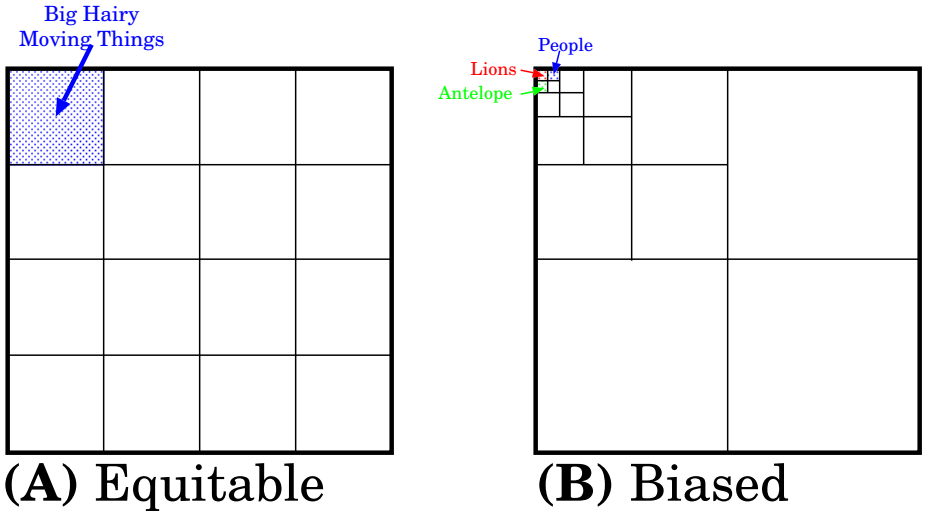


Figure 3.2: **(A)** An equitable perception scheme with 16 mental categories. **(B)** A biased perception scheme, also with 16 mental categories.

are large, and *all* percepts are uselessly vague. Had our ancestors possessed such vague perceptual apparatus, they would have been eaten by lions (which our ancestors would have perceived as 'big hairy moving things').

Instead, Persephone must chose a mental classification scheme which is precisely discriminates distinctions which affect Persephone's well-being, and glosses over distinctions which do not. I'll refer to such a perceptual scheme as **biased** (Figure 3.2B). The key point is this: to extract *any* useful information from its environment, a person of limited cognitive resources *must* have a highly biased perception scheme.

For example, Persephone the Paleolithic hunter-gatherer needs a very specific perception of the species of plants and animals in her world, but requires only a vague perception of minerals or geological formations. In her mind, most rocks would just engender the percept, 'rock'. On the other hand, Persephone the post-industrial geologist has a extremely detailed perception of various minerals, but is largely oblivious to vegetation. Persephone aspires to perfection in all things, and undoubtably would prefer to have detailed perception of flora, fauna, and geology. But her mental resources are limited, so she must prioritize.

If the hunter-gatherer and the geologist were to meet, they would perceive 'order' in very different situations. Where the hunter-gatherer sees a 'random jumble of boulders', the geologist sees a detailed record of 500 million years of geological history. Where the geologist sees only a 'forest', the hunter-gatherer sees ten edible plant species, thirteen poisonous ones, one tree which is good for making houses, another for canoes, another for weapons, etc.

Regardless of the perceptual scheme Persephone possesses, most of her repertoire of percepts is exhausted in a relatively detailed classification of some small part of $\mathcal{W}$. The

remaining few percepts are left to cover the vast majority of $\mathcal{W}$, and will represent vague categories like 'rocks', 'forest', or just, 'unrecognizable'. In her mental 'map' of $\mathcal{W}$, Persephone has only enough 'ink' to chart a small portion with any degree of detail, and she must leave the vast majority of the map virtually blank, with only the words, *terra incognita*.

(iv) Measuring the size of Mental Categories

To define a 'biased' perceptual scheme, I employed the 'size' of the mental category $\mathcal{C}(m)$ as a subset of $\mathcal{W}$. What exactly does this mean? There are several ways that the concept of 'size' can be made precise, depending on the mathematical structure of $\mathcal{W}$.

Cardinality: Suppose $\mathcal{W}$ and $\mathcal{M}$ are large finite sets. We suppose that $\mathcal{W}$ is much larger. For example, perhaps $\mathcal{M}$ contains 1 000 000 distinct states, while $\mathcal{W}$ contains 1 000 000 000 distinct states. Then clearly, the average percept $m \in \mathcal{M}$ must represent a category $\mathcal{C}(m)$ containing about 1000 distinct worldstates.

Dimension: If $\mathcal{W}$ is a vector space[2] or a manifold[3], then the size of $\mathcal{W}$ can be measured by its *dimension:* the number of distinct coordinates needed to exactly specify a point in $\mathcal{W}$. For example, if $\mathcal{W}$ is 100-dimensional, and $\mathcal{M}$ is 20-dimensional, and $p : \mathcal{W} \longrightarrow \mathcal{M}$ is a 'reasonable' function[4], then for percept $m \in \mathcal{M}$, the category $\mathcal{C}(m)$ will be an 80-dimensional submanifold in $\mathcal{W}$. In other words, *perception* provides Persephone with exact information about 20 coordinates, while leaving the other 80 entirely unspecified.

Diameter: If $\mathcal{W}$ is a *metric space*[5], then the size of a subset $\mathcal{C}(m) \subset \mathcal{W}$ is its *diameter* –the maximum distance between two points in $\mathcal{W}$.

For example, suppose $\mathcal{W}$ was a map of the world, and $\mathcal{M}$ was the set of countries. For any country $m \in \mathcal{M}$, the set $\mathcal{C}(m)$ is just the territory occupied by the country m on the map. The assertion, 'You are in Luxembourg' describes your location much more precisely than the assertion, 'You are in Canada', because the diameter of Luxembourg is much smaller. Two people in Luxembourg could be separated by a distance of at most 50 km. Two people in Canada could potentially be 3000 km apart.

Probability: If $\mathcal{W}$ is a *probability space*, then the size of a subset $\mathcal{C}(m) \subset \mathcal{W}$ is just its probability as a random event. High probability events are unsurprising, and thus, provide relatively little information. *Low* probability events *are* surprising, and provide more information.

[2] A mathematical generalization of a line, plane or 3-dimensional volume.

[3] A mathematical generalization of a curve or surface.

[4] That is, *linear* or *differentiable*.

[5] That is, a mathematical structure where one can measure the 'distance' between any two points in $\mathcal{W}$.

For example, suppose I flip a coin one hundred time, and tell you how many times it came up 'heads'. Thus, $\mathcal{W}$ is a set containing $2^{100} \approx 10^{30}$ elements (all possible sequences of coin flips), while $\mathcal{M}$ contains 101 elements (the numbers from 0 to 100). It is highly probable that I will flip approximately 50 heads; maybe 51 or 48. But probably not 75. And almost certainly not 100 heads in a row. The (unsurprising) statement, 'I flipped 53 heads and 47 tails' contains very little information, because there are about 10^{30} possible sequences of a hundred coin flips which yield 53 heads and 47 tails. However, the (very surprising) statement, 'I flipped 100 heads in a row' contains a *lot* of information: there is only *one* possible sequence of coin flips leading to this state. The statement, 'I flipped 99 heads' contains slightly less information, but still a lot. There are one hundred possible sequence of coin flips leading to this state, which is more than one, but still much, much less than 10^{30}.

Information Content: Suppose percept m corresponds to mental category $\mathcal{C}(m)$, which has probability P. The previous reasoning leads us to define the *information content* of the percept m as:

$$I(m) \;=\; -\log_2(P)$$

Thus, low probability percepts are accorded high information content, while high probability percepts get low information content. In the previous example, the statement 'I flipped 100 heads' has information content 100, and 'I flipped 99 heads' has information content 93.3, while 'I flipped 53 heads' has information content less than 1.

Amongst these notions of size, the probabilistic formulation is the most natural and versatile. We will use the accompanying notion of information content to characterize subjective order and disorder. Persephone perceives *subjective order* if she perceives a percept of *high* information content. For example, a sequence of 100 'head' coin flips in a row would be perceived by most people as a suspiciously 'ordered' sequence of events. Persephone perceives *subjective disorder* when she perceives a percept of *low* information content. Thus, flipping 53 heads seems quite 'disorderly', because a such a sequence would appear 'patternless'.

This perception is *subjective* because it depends entirely upon the specific patterns Persephone is looking for —ie. the specific information she chooses to extract from the coin flips. Until now, her perception of the coin flips has been limited to merely counting heads. But suppose instead she treated the sequence of heads and tails as a *binary number*: a sequence of 100 bits. Perhaps then she notices that the formerly 'random' sequence of 53 heads and 47 tails is actually a binary representation of the first 30 digits of π. By a change in perspective, a worldstate formerly perceived as 'disorder' suddenly becomes highly orderly.

(v) Ergodicity and the Second Law

If order and disorder are mere artifacts of perception, then is **(2LT)** also an artifact? From the right 'perspective', is the universe is actually proceeding towards a state of *greater*

order? Unfortunately, no.

As we've seen, anyone with limited cognitive resources, to survive, must adopt a *biased* perception scheme which provides highly detailed information about certain regions of $\mathcal{W}$, while leaving large parts of $\mathcal{W}$ as *terra incognita*. However, physical systems evolves along their own path, heedless of how we chose to perceive them. A system $\mathcal{S}$ is said to be *ergodic* if, over time, it tends to spend roughly equal amounts of time in all parts of its statespace $\mathcal{W}$. Hence, if Persephone were to inspect $\mathcal{S}$ at a random moment, she would be equally likely to see it in one part of $\mathcal{W}$ as in any other part. But wait: Persephone doesn't 'see' the actual state of $\mathcal{S}$; she only sees her *perception* of it. And a very large part of $\mathcal{W}$ is relegated to a relatively small number of Persephone's percepts; a *terra incognita* which she only vaguely perceives as 'random' or 'disordered'.

In other words, if $\mathcal{S}$ is ergodic, then, most of the time, when Persephone looks at $\mathcal{S}$, she will perceive it to be in a state of 'subjective disorder'. Furthermore, this is true *independent* of the exact nature of her perceptual representation. Different people will perceive subjective disorder in different situations (for example, the head-counter vs. the binary number watcher), but everyone will perceive disorder 'most' of the time.

This yields a more precise statement of the Second Law of Thermodynamics:

> Let $\mathcal{O}$ be a (biased) observer, and let $\mathcal{S}$ be an ergodic system. An observation of $\mathcal{S}$ by $\mathcal{O}$ will, with high probability, yield a percept of low (subjective) information content (ie. of high subjective disorder) to $\mathcal{O}$.
>
> The less information a percept p contains (ie. the more 'disorderly' it is), the more likely that $\mathcal{O}$ will observe p.

Even if $\mathcal{S}$ *begins* in a state of high (subjective) order, it will soon (by virtue of ergodicity) leave this state and enter a more (subjectively) disorderly one; hence we *perceive* that $\mathcal{S}$ 'proceeds towards a state of maximum disorder'.

Some key observations about this formulation of **(2LT)** :

(2LT) is a statement about the *perception* of $\mathcal{S}$, not the reality.

(2LT) is only true when perception is *biased*, and $\mathcal{S}$ is an *ergodic* system.

(2LT) is a statement about the higher relative *probability* of 'disordered' percepts.

The greater the bias of our perception, the greater the relative probability of subjective disorder. The stronger the ergodicity of the system[6], the faster the perceived 'decay of order' will be. **(2LT)** is appears an 'inexorable principle' of thermodynamic systems precisely *because* thermodynamic systems are rapidly ergodic[7], and because we perceive them in a highly biased way.

[6] That is, the speed with which the system traverses its statespace.

[7] This is called **Boltzmann's Ergodic Hypothesis**.

(vi) Thermodynamic Entropy

This formulation of **(2LT)** is not the quantitative one found in most physics texts. The quantitative version reads:

> In any closed thermodynamic system, the entropy never decreases, and usually increases over time.

The word 'entropy' rich in connotation and misconception. Let's be precise about its physical meaning:

- Entropy is a property of *macrostates* (ie. 'percepts'), *not microstates*.
- Entropy is not an absolute quantity (like mass), but a *relative* one (like potential energy). We can only speak of potential energy *gap* between two states; likewise, we can only speak of the entropy *difference* between two (macro)states of a system.

The entropy difference between two macrostates is usually defined through a certain integral. However, this definition is equivalent to the following one:

> Let m_0 and m_1 be two macrostates of a system $\mathcal{S}$, containing N particles. The entropy difference $h(m_1)-h(m_0)$ is proportional to the negative difference of their information contents. Formally, $h(m_1) - h(m_0) \;=\; -c \cdot N \cdot \Big(I(m_1) - I(m_0)\Big)$.

Here, c is some constant, which depends upon the choice of physical units, and which we can assume is equal to 1. Hence, our observation that ergodic systems *minimize* information content is equivalent to the assertion that they *maximise* entropy.

In the appendix at the end of the chapter, I discuss some simple examples that show this 'informational' definition is equivalent to the 'textbook' definition of thermodynamic entropy.

(vii) Work vs. Heat

Stuart Kauffman[20] complains that physics makes no clear distinction between 'work' and 'heat'. Both are forms of energy. The difference is that *work* is 'useful' or 'structured' energy (eg. the 'work' of lifting a brick against gravity, or of compressing a cylinder of gas), while *heat* is 'useless' or 'waste' energy (ie. the heat released when the brick is dropped and hits the floor, or when the cylinder explodes and the pressure is released).

Thermodynamics was originally developed to characterize the efficiency of machines; that is, to quantify how much of the energy they consumed was converted into *work*, and how much was wasted as *heat*. In the 'macroscopic' regime of thermodynamics, there is a clear distinction between the work and heat. In the microscopic regime, however, this

distinction breaks down, and it is precisely this regime which interests a theoretical biologist like Kauffman.

Cellular metabolism uses energy from exothermic reactions (eg. the oxidation of sugars) to drive endothermic reactions (eg. protein synthesis), and these processes explicitly involve *microstates*, not macrostates, so macroscopic concepts like 'entropy' are not applicable. Indeed, the metabolic chemistry is driven by thermal energy, so that 'heat' actually *becomes* 'work' (hence the importance of body temperature in homeothermic animals).

In the long term, we can watch an organism eat and metabolize, and conclude that its survival and growth constitute 'work', while its radiated heat and excreta are 'heat'. But when we look closely, this distinction, like so many others, becomes fuzzy.

It seems that 'work' is energy that is *ordered*, while 'heat' is energy that is *disordered*. But the terms 'order' and 'disorder' are subjective, and really just describe the *information content* of a percept. Hence, we can reformulate the distinction between work and heat in subjective, informational terms:

- **Work** is energy whose form and structure is *known*. In other words, it is energy accompanied by *information*.
- **Heat** is energy whose form and structure is *unknown*. It is energy *without* information.

Obviously these are extremes; any quantity of energy lies somewhere between 'pure work' (about which we know everything), and 'pure heat' (about which we know nothing).

Thus, lifting a brick is 'work' because we have precise information about the form of the energy (gravitational potential), from which we can draw useful consequences. The thermal energy of a fallen brick is 'heat' because we no longer have any precise information about its form (only that it somehow involves the thermal vibrations of 10^{25} brick molecules).

(viii) Disorder vs. Complexity

The perception of disorder is closely related to perceived *complexity*. We perceive a system as 'ordered' if it is amenable to a 'simple' description, and 'disordered' if it is not. For example, the sequence

01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01 01

is 'ordered', because we can simply describe it as 'Thirty repetitions of 01'. The sequence

16437 32478 95956 25040 43543 20394 32039 34950 34550 23030 23303 23430

is 'disordered', because it seems to admit no description shorter than merely enumerating the entire sequence verbatim.

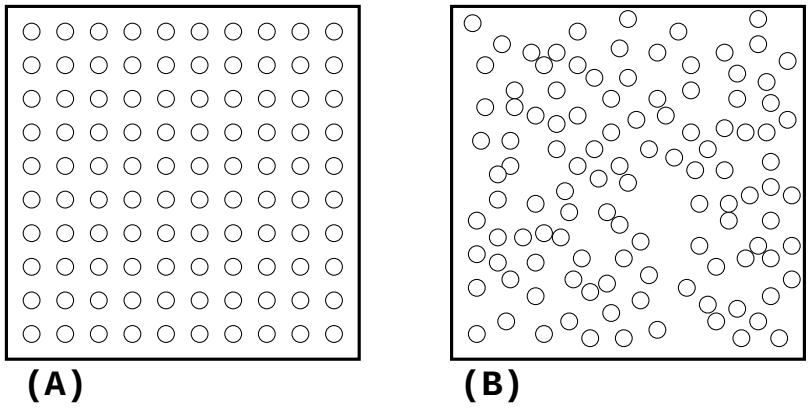


Figure 3.3: Order seems related to complexity.

Another example: Figure 3.3(A) seems 'ordered' and 'simple' because it admits a simple description: 'A 10×10 array of circles.' Figure 3.3(B) seems 'disordered' and 'complex' because the only (apparent) description is the picture itself.

Here, we are implicitly identifying the *complexity* of an object with the *length of our description* of that object. Hence, we are really observing:

> The perceived disorder of a state is proportional to the size of a complete description of that state.

This makes sense. Things appear 'ordered' if they fit well into our mental categories. Since our language is adapted to our mental categories, things which fit well into mental categories will be easily (ie. concisely) expressible in language. Things which defy easy mental categorization will also defy easy verbal description.

Note that, just as 'disorder' is *observer*-dependent, 'complexity' is *language* dependent; things which are very difficult to describe in one language may have a very simple description in others. The goal of science is in large part the search for the 'right language' to describe natural phenomena; a language in which an accurate description will be as simple as possible. This is the intuition behind *Occam's razor*: If your theory necessitates verbose explanations or descriptions of phenomena, then you are probably using the wrong language, and perhaps the wrong theory (see §(vii)).

Kolmogorov-Chaitin complexity: This characterization of 'complexity' in terms of 'minimum description length' is called **Kolmogorov-Chaitin complexity** [5, 6, 42]. Given a language $\mathcal{L}$ and a system $\mathcal{S}$, the KC-complexity of $\mathcal{S}$, relative to $\mathcal{L}$, is the minimum length of a complete description of $\mathcal{S}$ in $\mathcal{L}$. We write this as $\mathbf{KC}_{\mathcal{L}}(\mathcal{S})$

KC-complexity thus seems entirely language-specific, and thus, rather meaningless. However, the KC-complexity in any two languages $\mathcal{L}_1$ and $\mathcal{L}_2$ are *asymptotically equivalent* in the

following sense: there is some constant c so that, for any object $\mathcal{S}$, the difference between $\mathbf{KC}_1(\mathcal{S})$ and $\mathbf{KC}_2(\mathcal{S})$ will always be less than c. Thus, in theory, if $\mathcal{S}$ is an 'extremely complex' object (so that $\mathbf{KC}_1(\mathcal{S})$ is much larger than c), then $\mathbf{KC}_1(\mathcal{S})$ and $\mathbf{KC}_2(\mathcal{S})$ will be roughly equal in size.

This seems to contradict my earlier assertion that complexity (and with it, perceived disorder) are observer-dependent. However, you should keep in mind that the constant c is extremely large; to be precise, it is the complexity of a *complete translation scheme* from $\mathcal{L}_1$ to $\mathcal{L}_2$ and vice versa. Hence, the 'asymptotic equivalence' $\mathbf{KC}_1(\mathcal{S})$ and $\mathbf{KC}_2(\mathcal{S})$ should be treated like statement, 'One thousand monkeys pounding on typewriters will, after sufficient time, produce *Hamlet*,' or like the Poincaré Recurrence Theorem, which states that, 'after sufficient time', an ergodic system (say, a mixture of ink an water) will return to its original state (ie. spontaneously unmix). In all three cases, the statement, although true, involves such unimaginably vast timescales or complexities as to be meaningless for practical purposes.

Appendix: Examples of Thermodynamic Entropy

To see that our 'informational' definition of thermodynamic entropy is equivalent to the 'textbook' definition, we will look at some simple examples. Recall that $I(m_0) = -\log(P_0)$, where P_0 is the probability of macrostate m_0. Thus,

$$h(m_1) - h(m_0) \quad = \quad N \cdot \Big(I(m_0) - I(m_1) \Big) \quad = \quad N \cdot \Big(\log(P_1) - \log(P_0) \Big) \quad = \quad N \cdot \log\left(\frac{P_1}{P_0} \right)$$

Let $\mathcal{S}$ be a system of N identical gas molecules. In this case, the *microstate* of the system specifies the 3-dimensional positions and velocities of N particles; hence, it is a vector containing $6N$ coordinates. Thus, $\mathcal{W} = \mathbb{R}^{6N}$.

Let $\mathbf{X} \subset \mathbb{R}^3$ be some region of three dimensional space, and let m be the macrostate corresponding to the assertion, 'The gas is in region $\mathbf{X}$'. This tells us nothing about the position of any *individual* particle —only that all particles are in $\mathbf{X}$. Thus, if $\mathbf{p} \in \mathbb{R}^{3N}$ is the $3N$-dimensional vector specifying the positions of all particles, then we know: $\mathbf{p} \in \mathbf{X}^N$. If $\mathbf{X}$ has volume V, then the volume of $\mathbf{X}^N$ in $\mathbb{R}^{3N}$ is just V^N.

If a particle s (of mass 1) has velocity $\mathbf{v}(s) = (v_1(s), v_2(s), v_3(s))$, then its kinetic energy is $e(s) = v_1^2(s) + v_2^2(s) + v_3^2(s)$. The *total* kinetic energy of system $\mathcal{S}$ is then:

$$E \;=\; \sum_{s \in \mathcal{S}} e(s) \;=\; \sum_{s \in \mathcal{S}} v_1^2(s) + v_2^2(s) + v_3^2(s)$$

Represent the velocities of all particles in $\mathcal{S}$ with a $3N$-dimensional 'collective velocity vector' $\mathbf{V} = [V_1, V_2, \ldots, V_{3N}]$, containing $v_1(s), v_2(s), v_3(s)$ for all $s \in \mathcal{S}$. Then

$$E \;=\; V_1^2 + V_2^2 + \ldots V_{3N}^2$$

The temperature T is the *average* kinetic energy: $T = E/N$. Hence $E(\mathcal{S}) = N \cdot T$.

If m is the macrostate corresponding to the assertion, 'The gas has temperature T', then we know nothing of the velocity individual particles —only that their *total kinetic energy* is $N \cdot T$. We don't know how the energy is distributed amongst the particles —all we know is that $\mathbf{V} \in \mathbb{S}(N \cdot T)$, where $\mathbb{S}$ is the sphere of radius $\sqrt{N \cdot T}$ in $\mathbb{R}^{3N}$. Observe that the *surface area* of this sphere in $\mathbb{R}^{3N}$ is $C_N \cdot (N \cdot T)^{3N/2}$, where C_N is a constant[8].

Now consider the following scenarios:

Adiabatic Expansion: Suppose the gas originally occupies a region $\mathbf{X}_0$ of volume V_0, and we allow it to expand to occupy a region $\mathbf{X}_1$ of volume V_1, without changing temperature. Then its thermodynamic entropy will increase by an amount proportional to $N \cdot \log\left(\frac{V_1}{V_0}\right)$

To see this, suppose m_0 is the macrostate '$\mathcal{S}$ is contained in $\mathbf{X}_0$ at temperature T', then the corresponding family of microstates is $\mathcal{C}(m_0) = \mathbf{X}_0^N \times \mathbb{S}(N \cdot T)$, a region of volume $V_0^N \cdot C_N \cdot (N \cdot T)^{3N/2}$. Assuming the system is ergodic, the *probability* of this region is directly proportional to this volume, we conclude that $P_0 = V_0^N \cdot C_N \cdot (N \cdot T)^{3N/2}$

Likewise, if m_1 is the macrostate '$\mathcal{S}$ is contained in $\mathbf{X}_1$ at temperature T', then $\mathcal{C}(m_1) = \mathbf{X}_0^N \times \mathbb{S}(N \cdot T)$ is a region of probability $P_1 = V_1^N \cdot C_N \cdot (N \cdot T)^{3N/2}$. Thus, the ratio of the two probabilities is just:

$$\frac{P_1}{P_2} = \frac{V_1^N \cdot C_N \cdot (N \cdot T)^{3N/2}}{V_0^N \cdot C_N \cdot (N \cdot T)^{3N/2}} = \left(\frac{V_1}{V_2}\right)^N$$

(because all other terms cancel). Thus, according to our definition:

$$h(m_1) - h(m_0) = \log\left(\frac{P_1}{P_0}\right) = N \log\left(\frac{V_1}{V_2}\right)$$

Heating: Suppose the gas is contained in a region $\mathbf{X}_0$, and originally has temperature T_0. If we heat the gas to temperature T_1 without allowing it to leave $\mathbf{X}_0$, then its thermodynamic entropy will increase by an amount proportional to $\frac{3}{2}N \cdot \log\left(\frac{T_1}{T_0}\right)$.

To see this, suppose m_0 is the macrostate '$\mathcal{S}$ is contained in $\mathbf{X}$ at temperature T_0', and m_1 is the macrostate '$\mathcal{S}$ is contained in $\mathbf{X}$ at temperature T_1'. Then:

$$\begin{aligned} \mathcal{C}(m_0) &= \mathbf{X}^N \times \mathbb{S}(N \cdot T_0) \quad \text{has probability} \quad P_0 = V^N \cdot C_N \cdot (N \cdot T_0)^{3N/2}, \\ \text{and } \mathcal{C}(m_0) &= \mathbf{X}^N \times \mathbb{S}(N \cdot T_1) \quad \text{has probability} \quad P_1 = V^N \cdot C_N \cdot (N \cdot T_1)^{3N/2}, \end{aligned}$$

So that

$$\frac{P_1}{P_2} = \frac{V^N \cdot C_N \cdot (N \cdot T_1)^{3N/2}}{V^N \cdot C_N \cdot (N \cdot T_0)^{3N/2}} = \left(\frac{T_1}{T_0}\right)^{3N/2}$$

[8]To be precise, $C_N = \frac{2\pi^{3N/2}}{\Gamma(3N/2)}$, where Γ is the Gamma function. For example, $C_8 = \frac{2\pi^{12}}{\Gamma(12)} = \frac{2}{11!}\pi^{12}$.

and thus,

$$h(m_1) - h(m_0) \;=\; \log\left(\frac{P_1}{P_0}\right) \;=\; \frac{3N}{2}\log\left(\frac{T_1}{T_2}\right)$$

Similar reasoning can be applied to standard thermodynamic scenarios, like:

Combination: If $\mathcal{S}_1$ and $\mathcal{S}_2$ are two disjoint systems, and $\mathcal{S}$ is their aggregate, then the entropy of $\mathcal{S}$ is the sum of the entropies of $\mathcal{S}_1$ and $\mathcal{S}_2$.

Thermal Equilibration: If $\mathcal{S}_1$ and $\mathcal{S}_2$ are at different initial temperatures and are placed in thermal contact, they will exchange energy until they reach the same temperature, which is the macrostate of maximal entropy.

Pressure Equilibration: If $\mathcal{S}_1$ and $\mathcal{S}_2$ are at different initial pressures and are allowed to interact through a piston, they will move the piston until they reach the same pressure, which is the macrostate of maximal entropy.

4 Language and Discourse

What is a *language*? In Chapter 2, I vaguely described language as the arrangement of sequences of 'signifiers' (verbal utterances, written symbols, pictures, gestures, facial expressions, movements of gaming tokens, etc.) to convey meaning. To be more precise, let $\mathcal{A}$ be the set of all signifiers. For example:

- In a *written* language, $\mathcal{A}$ consists of all letters, punctuation marks, and other written characters.

- In a *spoken* language, $\mathcal{A}$ is the set of all phonemes. $\mathcal{A}$ may also include verbal inflections, facial expressions, or gestures, to the extent that these convey semantic content.

A *speech act* is then some finite sequence $a_0, a_1, \ldots, a_n$ of these signifiers. The set of all such finite sequences is denoted A^*. Speech acts are not just *words* or *sentences* because we rarely communicates in single sentences. Instead, we usually communicate by long series of sentences: paragraphs, monologues, essays, novels, etc. Hence, the sequence $a_0, a_1, \ldots, a_n$ may be quite long.

(i) Constraints and Formal Languages

Of course, not all sequences are allowed; some are ruled out as 'ungrammatical' or 'nonsensical'. For example 'ewjd asdf. wae, asd' is not a valid speech act in written English. Thus, we might start by saying:

> *A* **language** *is some subset* $\mathcal{L} \subset \mathcal{A}^*$.

The there is an extensive mathematical theory of 'formal languages' of this kind[18]. The 'rules' of the language constrain which sequences in $\mathcal{A}^*$ are admissible to $\mathcal{L}$.

For example, the written English language imposes constraints at several levels:

Spelling constrains the juxtaposition of individual letters: 'piece' is admissible, but 'pees' is not.

Grammar constrains the arrangement of words into sentences: 'The moon was a ghostly galleon' is admissible, while 'crepuscular exegete lugubrious' is not.

Semantics requires these sentences to be meaningful. Hence, 'Green ideas dream furiously' is inadmissible[1].

Coherence requires successive sentences to be semantically related. There should be a clear progression of ideas. Pronouns in later sentences should have clear references in earlier sentences.

Note that, while **Spelling** and **Grammar** can be *formally* specified, **Semantics** cannot. Nonetheless, there is no doubt that semantic restrictions exist; everyone will immediately agree that 'Green ideas dream furiously' is nonsensical. The **Coherence** constraint is even vaguer, but is somehow related to the semantics of the language.

Thus, we see that the *meaning* of language is intimately related to the *constraints* that determine which speech acts are admissible.

The Language of Mathematics Let $\mathcal{A}$ be the alphabet of mathematics, containing Greek and Roman letters, quantifiers like "$\exists$" and "$\forall$", brackets, logical relation symbols, etc. Let $\mathcal{L}_{\text{math}} \subset \mathcal{A}^*$ be the *Language of Mathematics.* Sequences in $\mathcal{L}_{\text{math}}$ must be collections of sentences satisfying the constraint:

WFF: Each sentence must be a well-formed formula (ie. 'grammatically correct' in a mathematical sense). thus, '$\}2\{\{\}5x\forall\exists\}$', is inadmissible, but a formula like '$2+2=5$' is admissible (even if it is false!)

[1] This example is due to Chomsky.

According to the 'Platonist' school of mathematical ontology, mathematical objects exist independent of our imaginations. As such, there are certain statements about them which are true, and others which are false, regardless of whether we recognize them as such. For example, it was true that Olympus Mons was the highest mountain on Mars, even before we discovered this fact by launching spacecraft to that planet. In the same way, it was true that there are an infinity of prime numbers, even before there were hominids who understood numbers.

Let $\mathcal{L}_{\text{true}} \subset \mathcal{L}_{\text{math}}$ be the *Language of true mathematics*. Sequences in $\mathcal{L}_{\text{true}}$ must be collections of sentences satisfying two constraints:

WFF: Each sentence must be a well-formed formula

Truth: Each sentence must be true. Thus, '$2 + 2 = 5$' is now inadmissible.

$\mathcal{L}_{\text{true}}$ isn't very practical, since we often don't *know* what is true. Indeed we often *can't* know, since we can't make 'direct mental contact' with complex mathematical objects. It seems possible to appreciate the truth of '2+2=4' by simply imagining 4 objects; that is, by making 'direct mental contact' with the objects '2' and '4'. However, it is seems impossible to appreciate the truth of abstract mathematical propositions in this manner. For example, I can't appreciate the truth of the assertion, 'There is no largest prime number' simply by 'imagining all the prime numbers' in my mind. Instead, I can only recognize the truth of this proposition by constructing a *proof* of it.

Indeed, according to the 'Formalist' school of mathematical ontology, mathematical objects do not exist independent of our imagination of them. Our 'subjective experience' of these objects is just the experience of manipulating formal symbols according to formal rules. Thus, for example, when you contemplate the number 2 you are *not* making 'direct mental contact' with some abstract Platonic ideal of 'twoness'; you are simply manipulating certain symbols according to certain conventions. What is 'true' about mathematics is simply what is true about these symbols.

So now consider $\mathcal{L}_{\text{pr}}$, the *Language of provable mathematics*. Sequences in $\mathcal{L}_{\text{pr}}$ must be collections of sentences in $\mathcal{L}_{\text{math}}$ satisfying two constraints:

WFF: Each sentence must be a well-formed formula

Logic: Each sentences is either an axiom, or follows from *previous* sentences according to explicit formal deductive rules.

According to the Formalist school, the *meaning* of formulae in $\mathcal{L}_{\text{math}}$ is determined not by their 'truth' or 'falsehood', but by their 'provability' —ie. by their membership or non-membership in $\mathcal{L}_{\text{pr}}$. It is hard to argue with this view, because as mathematicians, we can rarely directly apprehend 'truth'; all we can perceive is *provability*.

(ii) Probabilistic Constraints

So far we're defining a language as a subset $\mathcal{L} \subset \mathcal{A}^*$, consisting of all 'admissible speech acts'. However, the distinction between 'admissible' and 'inadmissible' is not black and white. For example, not all English writing satisfies **Grammar**. Inadvertent grammatical mistakes are commonplace, and many authors deliberately flout grammatical conventions. The constraint of **Spelling** is violated either by mistake or by the introduction of neologisms (eg. Derrida's 'differance'). Finally, the **Semantics** constraints is clearly ambiguous and subjective: what is a 'sensible' sentence, anyway? 'The moon was a ghostly galleon' is not 'sensible' if interpreted literally, but 'makes sense' if we understand it metaphorically.

So, we should regard language constraints as *absolute*, but only as *probabilistic*. To call a certain speech act 'inadmissible' only means that it is extremely *improbable* in ordinary speech, except perhaps within certain specific contexts. We thus get a *probability distribution* μ on $\mathcal{A}^*$.

Each person speaks a unique language, reflecting her personal instantiation of **Grammar** and **Spelling** constraints. For example, some people chronically misspell particular words (eg. 'peice' vs. 'piece'), or are prone to a particular grammatical error ('John and me went for lunch'). Also, different people have different conceptions of what is semantically sensible. Thus, each person's language dictates a *different* probability measure on $\mathcal{A}^*$.

By assigning a probability to every speech act a person may perform, we implicitly encode her *beliefs and values*. She is highly unlikely to say things she does not believe. The statistical regularities of a speaker's language embody not only the *syntax* of the language, but the speaker's entire world view.

Suppose you were observing a speaker of an alien language, and making a record of statistical regularities in her speech. You notice that certain words never appear in certain relations to each other, whereas other words often do. Without understanding the language, however, it is impossible to determine which statistical regularities are due to formal, syntactic considerations, and which are due to the speaker's belief system. This raises the question: if the two are not empirically distinguishable, is there really a clear distinction between them?

In $\mathcal{L}_{\text{pr}}$ (the Language of Provable Math), the admissibility criteria were entirely *formal*. There was no distinction between exclusion based upon nebulous 'semantic' properties, and exclusion based upon explicit 'syntactic' properties —they were one and the same. Indeed, this is the defining property of formal systems.

(iii) Discourse and Ideology

Each speaker speaks her own language, with its own unique stochastic properties. However, speakers belonging to the same social group (tribe, nation, social class, political party, profession, academic specialty, etc.) can be expected to speak 'similar' languages. The statistical properties of their languages will be similar, which means that their languages embody

roughly the same syntactic regularities, and also roughly the same values and beliefs. In other words, common statistical properties can reflect the prevailing *ideology* or *culture* of a social group.

Semioticians refer to the common language of a social group as a **discourse**. The idea is not that all members of the group display exactly the same statistical properties in their speech; clearly, this is false, since each person speaks a unique language. Instead the discourse of a social group is the probability measure on $\mathcal{A}^*$ obtained by *averaging* over all members of the group. If the group is relatively homogeneous, then this average process will be a good approximation of the behaviour of any member of the group

A recurring theme in semiotics is the embodiment of *ideology* within *discourse*. The statistical properties of the discourse reflect not only the *conscious* beliefs of the speakers, but also their *unconscious assumptions*. From this premise, some people draw a radical and disturbing conclusion: since the discourse manifests as formal restrictions upon what speech acts are 'admissible', the discourse places limits upon *what can be said*, and thus, inadvertently *reinforces* the ideology it embodies. If you must articulate your ideas within the constraints of the discourse, then those constraints can make it difficult —maybe impossible —to articulate ideas which contravene the ideology embodied by that discourse. For example:

Political discourse is often manipulated to reinforce ideology. Orwell imagined a totalitarian state which systematically manipulated the English language until only 'doublespeak' was possible. However, Chomsky [28] has argued that a totalitarian apparatus is unnecessary, and that even in democratic societies, the political discourse is controlled by and for the ruling elite.

Military propaganda provides the most transparent example of discourse manipulation. Military operations are always 'defensive'. Every country has a Department of Defense —no one has a Department of Aggression. Insurgents are 'freedom fighters' if they work for you, but 'terrorists' if they work for the opposition.

Scientific discourse embodies Kuhnian *paradigms*: implicit judgements about what methodologies are 'scientific', which questions are 'interesting', and which assertions are 'sensible'. These paradigms reinforce acceptance of the orthodox theories. Unorthodox theories are rejected because they are literally *unspeakable* within the discourse.

Technological discourse is especially vulnerable to manipulation, because no preexisting language exists to describe new technology. Language must be *invented* —usually by the very people who want to sell products. For example, while a '`80386-SX`' chip sounds *more* powerful than a '`80386`', it is actually less so. An '`80486 DX4`' sounds 'four times' faster than an '`80486`'; actually, it is one fourth the speed. The acronym '`RISC`' sounds flashy, but it would be less impressive if people knew it stood for 'Reduced Instruction Set Chip'.

Gender roles manifest in subtle ways in our discourse. Feminine titles are often obtained by 'feminizing' a default masculine form (eg. 'host' vs. 'hostess', 'prince' vs. 'princess'); this suggests the male is the 'rule' and the female the 'exception'. Worse, 'feminine' words often have diminutive or patronizing connotations (eg. 'suffragette'.) Some words have such strong gender connotations that we only specify gender when these connotations are contradicted (eg. 'male nurse', 'lady doctor'; no one feels it necessary to say 'female nurse' or 'male doctor').

Philosophy itself, according to Jacques Derrida, is just another literary genre, whose texts should be 'deconstructed' to uncover the subtext of ideological assumptions beneath superficially rational arguments.

But are speakers really 'constrained' by their discourse? This is the question of *agency*: is the discourse a voluntary product of the speakers, or are the speakers merely vehicles through which the discourse is realized? Do I *speak* the discourse, or does the discourse instead *get spoken* by me?

A similar question could be asked of mathematics. Is the Language of the Provable a voluntary product of mathematicians, or are we instead merely the instruments by which the Language is realized? Do I *prove* theorems, or do they *get proved* by me? This is reminiscent of the question: are mathematical truths 'discovered' or 'invented'? One thing seems clear: although I am 'free' to try to prove any theorems I like, I *am* constrained to write proofs which are deemed 'rigorous' according to certain formal standards. If I do not satisfy these standards, then my 'proof' will be rejected by the mathematical community as inadmissible.

In a similar way, *every* social group enforces an ideology by rejecting as 'inadmissible' those speech acts which do not satisfy the constraints of a certain discourse. These constraints are usually unwritten and largely unconscious (indeed, by definition, if the admissibility constraints could be explicitly codified, then the discourse would be a form of mathematics). However, in every discourse, the constraints reflect prevailing assumptions about what is 'true' or 'rational' or 'coherent', what is an 'empirical fact'; what is 'virtuous'; etc, and thus, can enforce ideology.

(iv) Discourse and Thought

To the extent that I form my thoughts *within* the discourse, my thoughts are constrained by it. This is the *Sapir-Whorf hypothesis* [36, 37, 44]. The 'strong' version of this hypothesis says that I simply cannot *think* outside of the discourse. This is clearly nonsense: it assumes that I *think* in English (or in some discourse), which I clearly do not. For example, when contemplating mathematics, I think nonlinguistically, in terms of pictures, spatial intuitions or logical abstractions.

However, it is naïve to think that I develop a complete and perfect idea in 'purely mental' form and then 'translate' it into verbal discourse. In reality, my thoughts become fully formed

only when I attempt to *verbalize* them, and it is *then* that I often realize what I 'really mean'. What *seems* like a rigorous proof or a precise definition often disintegrates when I attempt to articulate it. By the time I've obtained a satisfactory written expression, the proof is often much different than the 'mental' version I began with.

Likewise, in philosophy, I often figure out what I 'really think' in the process of trying to explain it to someone else. Again, the discourse (and its limitations) are explicitly involved in the formulation of the idea.

Furthermore, intellectual activity does not take place in a vacuum, but rather, in a community. Most of what I 'know', I learned from someone else. Most of what I think is shaped by the ideas of others. But when we learn or teach ideas within the medium of a discourse, it shapes those ideas. Most obviously, the expressive limitations of the discourse (see Chapter 2) fundamentally limit the ideas we *can* learn or teach within it.

In a debate, our judgements of the logic of an argument are unconsciously shaped by discourse. We tend to accept an argument that 'sounds good'. Indeed, making an argument 'sound good' has a name: it's called *rhetoric*. Even fallacy can be made to appear logical by manipulation of language; this is called *sophistry*.

Conversely, we tend to automatically dismiss assertions that 'sound incoherent'. For example, academic 'outsiders' are often ignored or dismissed because they don't know the 'right jargon'. Even knowing the jargon isn't enough: there are subtle and intangible stylistic conventions which function as shibboleths for each academic culture; an 'insider' can spot a paper written by an 'outsider' literally within a paragraph, by (unconsciously) noticing the violation of these conventions. Overworked or lazy academics tend to dismiss outsiders as 'cranks', without carefully considering their ideas.

(v) Abstraction Levels in Computation

Computers must ultimately be usable by human beings, and thus, must be able to exchange information with humans. Since the binary machine code of a computer is incomprehensible to humans, and since human languages are likewise incomprehensible to machines, some mechanism must exist to translate machine-readable information into human-comprehensible form, and vice versa. This translation mechanism is called an **interface**.

It is not merely *users* who must interact with machines, but also *programmers*. Programmers also need to speak to the machine through intermediaries, which translate from human-level languages to machine-level code. These intermediaries are variously called **programming languages**, **environments**, **platforms**, or **operating systems**.

Machines must also exchange information with one another. Different machines speak different internal "languages", and the information must pass through physical channels with physical limitations. Thus, it is necessary to use some encoding scheme to take information from one machine, pass it through a channel, and deliver it in usable form to another machine. This is encoding scheme is called a **protocol**.

In short: since the fundamental purpose of a computer is to receive, process, and transmit information, a fundamental part of computer design is the design of *signifier systems* for the reception and transmission of this information.

In general, there are many "layers" of translation between any human user and a computer. Each of these layers translates from one "language" to another. The "lowest" level language is the binary language of the machine: bits moving through memory registers, processed mindlessly by the CPU. At the "highest" level is the language spoken by the user (often, this is not really a "language" at all, but rather, a visual interface). In between is a hierarchy: languages are increasingly abstract, human comprehensible, and "purpose-oriented" as one rises to the top, and increasingly concrete, mechanical, and "structure-oriented" as one descends to the bottom.

These layers are called **abstraction levels**. There are many reasons why so many abstraction levels are interposed between human and machine. A few of them are:

- Different languages are suited to different purposes. The language which is ideal for a novice user to speak is not ideal for someone who wants detailed and precise control of the machine.
- The use of abstraction levels allows us to *specify* how a certain program should work at a "high" level of abstraction, and then *implement* this specification at a "lower" level of abstraction. This has two advantages:
  - She who constructs the specification at the high level need not have any knowledge of how the lower level works. This allows her to focus on her area of expertise.
  - The lower-level implementation can be modified without requiring any modification of the higher level specification. This means that a program can be repaired or optimised without any visible change in functionality from the user's perspective.

Machine-to-machine communication is mediated through protocols which are also normally layered in a hierarchy of abstraction levels, for similar reasons.

In short: an abstraction level is a *language* —a representation system. It exhibits constraints like any other discourse, and is subject to the same issues of expressive completeness, incommensurability, and reification which affect any representation system. If a language makes certain kinds of information difficult or impossible to communicate, then it can can distort its speakers' communicative intent. If a person's experience of the world, or of other people, is mediated through her computer, then the limitations of that medium have a direct impact on her experience. To the extent that the computer mediates her reality, the computer *is* her reality. The limitations inherent in the computer's communication systems become the limitations of her reality.

A well-designed hierarchy of abstraction levels imposes no real limitations on the information which can be transmitted through its layers of translation. However, many computer systems are badly designed, and severely restrict the ways the humans can interface with the machine —and, through the machine, with one another. This is often simply due to incompetent engineering, but it sometimes embodies ideology. Some examples:

Interoperability: When two programs can exchange information and work cooperatively —in other words, when the speak the same language(s) —they are called *interoperable*. When the maker of the leading software package **W** wants to eliminate competitor **U** in the same market niche, it will often deliberately design **W** to be *non*interoperable with **U** software. This marginalizes the minority who use **U**, and coerces users to switch to **W**, just so that they can communicate with the **W**-using majority.

Indeed, suppose **W** is made by company **M**, which also makes several other programs –say, **O** and **E**. Then **M** will enhance interoperability between **W**, **O**, and **E**, while simultaneously sabotaging interoperability with rival programs. Thus, the market dominance of one product —say, **W** —can be used to coerce users to chose **E** over rival product **N**, even if, all things being equal, they prefer **N** to **E**.

Programs communicate with one another in many ways and at different levels, and the web of transactions within a computer is quite complex. Issues of interoperability are often subtle and multifaceted, and many users do not realize the extent to which the functionality of their software —and thus, their ability to use it —is dictated by these issues.

Front Ends: Fischer-Price vs. Fighter Jet: The *front end* of software is the interface most users deal with. The job of a front end is twofold:

Format the information about the current operating state of the software so that the user can quickly and easily extract essential information.

Facilitate the easy and rapid execution of complex tasks.

To achieve the first goal, the front end *suppresses* the majority of information, and only presents the 'relevant' details. To achieve the second, the front end expedites 'important' tasks with obvious buttons or easily accessible menus.

This judgement of 'relevancy' or 'importance' clearly depends upon the user and on her goals. Thus, a good front end can be *configured* by the user to display whatever information she desires, and suppress the rest. It can be modified to expedite exactly those tasks which the user most often performs.

A *bad* front end insists on hiding certain facts, and forces others into prominence, regardless of the user's priorities. It makes certain tasks easy, at the expense of making others arcane or impossible. Instead of the machine adapting to the user, the user must adapt to the machine. The user learns to avoid attempting certain tasks. This avoidance becomes

habit, and the habit becomes unconscious. The user has internalized the limitations of the machine.

In other words, a bad front end imposes *value judgements* about what the user's relationship to the technology should be. Two extremes:

- The 'Fischer-Price' front end treats the user like an infant, hiding almost all information, and providing a facile, 'idiot proof' control structure. The interface is both patronizing and disempowering. It engenders a false sense of technical mastery in the incompetent, while actually hindering the performance of complex tasks. Metaphorically, the interface is like a child's toy mixmaster, with only one button: if there's only one button, you can never press the wrong one.

- The 'Fighter Jet' front end overwhelms the user with a barrage of irrelevant technical information, and presents a cryptic and baroque control structure. By intimidating anyone unwilling or unable to invest hours studying instruction manuals, the interface feeds technophobia and 'learned helplessness', while concentrating power in a technocratic elite.

Ethnocentric Encoding: Sometimes, data encoding standards embody ethnocentric worldviews. Two examples:

Mailing Addresses: Software often has a datastructure to represent a mailing address. Some of these datastructures can only accept addresses from certain countries (usually the United States). For example, the datastructure may only accept a 5 digit numerical 'Zip code' (as found in the U.S.), rather than a 6 digit alphanumeric postal code (as in the U.K. and Canada), or it may require a 2-letter 'state code'. Indeed, often, it isn't even possible to specify a country. The design thereby embodies an implicitly 'Americocentric' worldview.

Language support: Most computer software was developed by and for English-speaking people, and until recently, could not properly represent the accents and non-English characters (eg. 'Ł', 'æ', 'ø') found in many European languages. Even when European languages are properly supported, languages employing non-Latin alphabets often aren't.

5 Mind and Meaning

What is meaning? Loosely speaking, this question splits into three parts:

1. How do languages convey meaning? How do words mean? This is the problem of *semantics*.

2. How do *thoughts* contain meaning? This is the problem of *intentionality*.

3. How do signs in general (eg. gestures, behaviours, cultural artifacts) convey meaning? This is the problem of *semiotics*.

Semiotics investigates meaning in an explicitly social or cultural context, by examining how signs convey meaning in a particular community or cultural milieu. I explore semiotics a bit in Chapter 8. In this chapter, I will consider semantics and intentionality.

Naïvely, words (and thoughts) contain meaning by 'representing' things. In chapter 2 I presented a simple model of 'representation': a language $\mathcal{L}$ *represents* ideas by means of a function $s : \mathcal{M} \longrightarrow \mathcal{L}$, where $\mathcal{M}$ is the space of mental states. I proposed this 'Function Model' to examine issues like 'ineffability'. The Function Model was presented without justification, and indeed is flawed, in several ways:

- The Function Model does not explain *how* a sentence represents a particular idea. Take the sentence, 'Gödel showed that we can never know if the Zermelo-Fraenkel Axioms are consistent.' What, exactly, makes this a sentence 'about' someone named Gödel, and not about Heraclitus?

- The Function Model can be sensibly applied to *perception* (via a map $p : \mathcal{W} \longrightarrow \mathcal{M}$) and to *speech* ($s : \mathcal{M} \longrightarrow \mathcal{L}$), but it doesn't address other, equally 'meaningful' mental representations. For example:

  Speculation concerning hypothetical worlds. For example, 'If Turing's group had not cracked the `ENIGMA` cipher, the Nazis would have won the war.' How can we make a 'meaningful' statement about an event that never happened, or a universe that doesn't exist?

  Mathematics, which concerns real but abstract entities. For example, the sentence, 'There are an infinity of prime numbers' seems to be 'about' things called 'prime numbers'. But these do not live in the physical world. Perhaps this sentence is about an *idea* in my mind. But then what is the *idea* about?

  Fantasy about unreal but concrete entities. For example, what is a Sherlock Holmes story really 'about'?

For the Function Model to account for these forms of representation, we must postulate ontologically dubious objects such as the 'space of all possible mathematical universes' or the 'space of all possible Sherlock Holmes stories'.

- The Function Model posits a direct, unambiguous correspondence between individual speech acts and subsets of mental state space. In other words, when someone performs the same speech act, she is always in the same 'family of mental states'. But this is contradicted by instances of dishonesty or confusion, and by metaphorical, ceremonial, or narrative uses of speech.

  For example, if I say, 'I see a storm approaching', this could mean several things:

  – I honestly believe I see an oncoming storm.
  – I am hallucinating.
  – I am lying.
  – I am quoting someone.
  – I am acting in a drama, or part of a ritual, and this is one of my 'lines'.
  – I am speaking metaphorically of an imminent (nonmeteorological) disaster.

Rather than a direct correspondence, it seems there is only a *probabilistic correlation* between my mental states and my speech acts.

(i) Mentation as Mechanism

The operation of the mind can be loosely described as follows:

> *Each moment, I experience sensations.*
> **(1)** *These sensations combine with my current mental state, causing a new mental state.*
> **(2)** *My new mental state then possibly causes me to behave in some way.*

Formally, we can represent process **(1)** by a transformation $(i, m_0) \mapsto m_1$, where i is the sensory *input*, m_0 is my *current* mental state, m_1 is my *subsequent* state. We can then represent process **(2)** by a transformation $m_1 \mapsto o$, where and o is my behavioural *output*.

If $\mathcal{I}$ is the space of all sensory 'inputs', $\mathcal{M}$ is the space of mental states, and $\mathcal{O}$ the space of behavioural 'outputs', we have a pair of functions:

(1) $\phi : \mathcal{I} \times \mathcal{M} \longrightarrow \mathcal{M}$, so that $m_1 = \phi(i, m_0)$.

(2) $\psi : \mathcal{M} \longrightarrow \mathcal{O}$, so that $o = \psi(m_1)$.

However, we can't assume that a particular mental state will *always* react to a particular sensory input in the same way. We must allow some randomness. Rather than a 'fixed' value in $\mathcal{M}$, we will let $\phi(i, m_0)$ be a *random value*, determined by some probability distribution on $\mathcal{M}$, which we will denote $\Phi(i, m_0)$. The special case when $\phi(i, m_0)$ is a fixed value is just the case where $\Phi(i, m_0)$ is a 'point mass': a distribution concentrating all its probability at a single point m_1 in $\mathcal{M}$.

Thus, Φ is a function from $\mathcal{I} \times \mathcal{M}$ into $\mathbb{P}(\mathcal{M})$, where $\mathbb{P}(\mathcal{M})$ is the space of all *probability distributions* on $\mathcal{M}$. We say that ϕ is a **stochastic function** from $\mathcal{I} \times \mathcal{M}$ into $\mathcal{M}$.

Likewise, instead of a function $\psi : \mathcal{M} \longrightarrow \mathcal{O}$, we must represent the 'behavioural output' with a stochastic function $\Phi : \mathcal{I} \times \mathcal{M} \longrightarrow \mathbb{P}(\mathcal{O})$, where $\mathbb{P}(\mathcal{O})$ space of all *probability distributions* on $\mathcal{O}$.

Note that I am not asserting that a human mind 'is' a stochastic function —this is obviously ridiculous. I am only saying that the mind *can be modeled* by a stochastic function, for our present purposes. Note also that this model does not commit us to any position on issues like materialism vs. dualism, free will vs. determinism, or computationalism vs. noncomputationalism. The model can accommodate each of these 'isms' as follows:

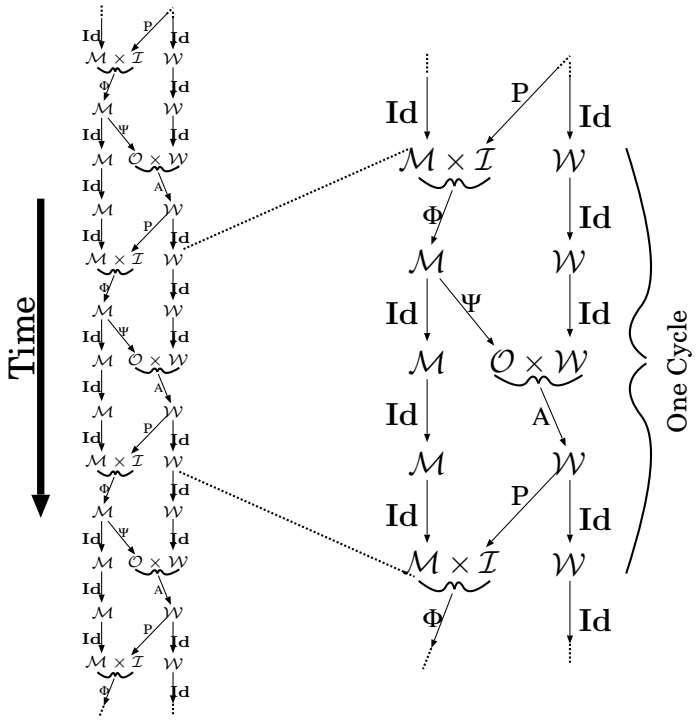
- *Materialism* is equivalent to asserting that $\mathcal{M}$ is the statespace of some physical system (ie. the brain). *Dualism* is the assertion that $\mathcal{M}$ is the statespace of some nonphysical system (ie. the 'soul').

- *Determinism* is the assertion that the function ϕ is deterministic —ie. that $\Phi(i, m_0)$ is a point mass in $\mathcal{M} \times \mathcal{O}$ for every $i \in \mathcal{I}$ and $m_0 \in \mathcal{M}$.

- *Computationalism* is the assertion that $\mathcal{I}$, $\mathcal{M}$, and $\mathcal{O}$ are all *countable* sets (ie. we can identify them with the set $\{0, 1, 2, 3, \ldots\}$) and that $\phi : \mathcal{I} \times \mathcal{M} \longrightarrow \mathcal{M} \times \mathcal{O}$ is deterministic and *computable* —ie. there exists a Turing machine which computes ϕ.

Our sensations are caused by the state of the world, and our actions change the world's state. We represent this with a pair of (stochastic) functions $P : \mathcal{W} \longrightarrow \mathcal{I}$ and $A : \mathcal{O} \times \mathcal{W} \longrightarrow \mathcal{W}$. *Perception* is described by P; if the world is in state $w \in \mathcal{W}$, then I will experience (ie. *perceive*) the (random) sensation $P(w) \in \mathcal{I}$. *Action* is described by A; if the world is in state w_0 and my behaviour is $o \in \mathcal{O}$, then this will change the (random) worldstate to $w_1 = A(w, o)$.

Thus, the cycle of interaction between myself and the world is described by the following diagram:



(ii) Semantics and Correlation

The semantics of language is determined by the correlation between utterances, thoughts, and events in the world. If Aletheia is speaking to Boromeo, we must distinguish between three kinds of semantics in Aletheia's utterances.

Intended Semantics correlates Aletheia's utterances with her own mental states.

Perceived Semantics correlates Aletheia's utterances with Boromeo's mental states.

Effective Semantics correlates Aletheia's utterances with the state of the world.

To be precise: let $\mathcal{L}_A$ be the set of possible utterances Aletheia could make, and let $\ell \in \mathcal{L}_A$ be a particular utterance; for example, the words, *'Il y a un chien.'* Let $m_A \in \mathcal{M}_A$ be Aletheia's (unknown) mental state, let $m_B \in \mathcal{M}_B$ be Boromeo's (unknown) mental state, and let $w \in \mathcal{W}$ be the (unknown) state of the world.

- Let $\mathcal{S}_A \subset \mathcal{M}_A$ be a certain subset of Aletheia's mental states —for example, the set of all mental states where Aletheia believes she is seeing a dog. Then $\mathcal{S}_A$ is the the *intended semantics* of ℓ if the conditional probability that $m_A \in \mathcal{S}_A$ is very high, given that Aletheia has just uttered ℓ.

- Let $\mathcal{S}_B \subset \mathcal{M}_B$ be a subset of Boromeo's mental states —for example, the set of all mental states where Boromeo believes Aletheia thinks she is seeing a dog. Then $\mathcal{S}_B$ is the the *perceived semantics* of ℓ if the conditional probability that $m_B \in \mathcal{S}_B$ is very high, given that Boromeo has just heard Aletheia utter ℓ.

- Let $\mathcal{S}_W \subset \mathcal{W}$ be a subset of worldstates —for example, the set of worldstates where a dog is present. Then $\mathcal{S}_W$ is the the *effective semantics* of ℓ if the conditional probability that $w \in \mathcal{S}_W$ is very high, given that Aletheia has just uttered ℓ.

An utterance may have some semantics without others. For example, If Boromeo doesn't understand French, then '*Il y a un chien*' has no *perceived* semantics for him. Also, utterances concerning Aletheia's mental states ('I feel sad'), mathematical abstractions ('5 is a prime number'), hypothetical scenarios, memories, or predictions need not correlate with the current world state, and thus, may have no *effective* semantics. Also, if Aletheia is delusional (for example, she frequently hallucinates dogs) then even an apparently concrete utterance need not have effective semantics.

(iii) Intentionality and Correlation

If the semantics of an abstract utterance ('5 is a prime number') is defined by correlating it with mental states, then we must next define the intentionality of mental states. Like semantics, intentionality is determined by correlations between mental states, perceptions, actions, and events in the world. We can distinguish five kinds of intentionality.

Sensory Intentionality correlates states of the world with sensations (ie. 'percepts'). The sensory intentionality of a percept is, loosely speaking, the set of worldstates which tend to coincide with that percept being experienced. It is determined by the stochastic function P.

Perceptual Intentionality correlates sensations with mental states. The perceptual intentionality of a concept is, loosely speaking, the set of percepts which tend to coincide with that concept coming to mind. It is determined by the stochastic function Φ.

Conceptual Intentionality correlates mental states ('concepts') with other mental states. The conceptual intentionality of a concept m is, loosely, the set of *preceding* concepts which can trigger concept m, and the set of *succeeding* concepts which might be triggered *by* concept m. Conceptual intentionality is determined by the stochastic function Φ.

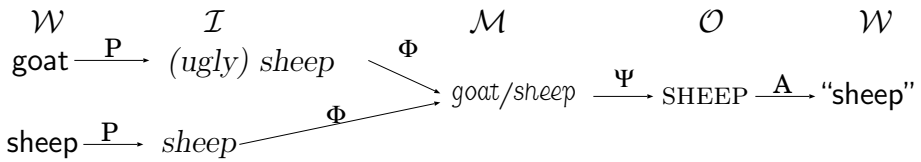
Inceptional Intentionality correlates mental states with an intent or will to act (we will use the word 'incept' to mean an intent to act, since unfortunately the word 'intention' already means something else in this discourse). The inceptional intentionality of a

concept is the set of incepts it tends to trigger. Inceptional intentionality is determined by stochastic function Ψ.

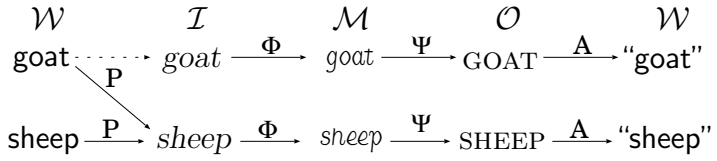
Behavioural Intentionality correlates incepts with behaviours, where a 'behaviour' is something which changes the state of the world in some way. It is determined by stochastic function A.

For example, consider a scenario where Aletheia, viewing some goats in the distance, remarks, 'Look at those ugly sheep.' We can interpret this remark in several ways:

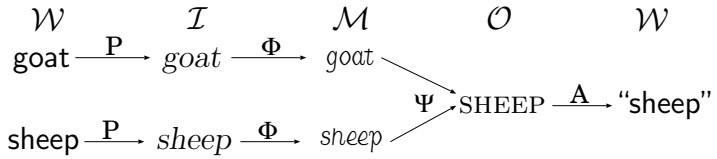
- Aletheia does not mentally distinguish goats from sheep. She can *see* a difference (and hence, perceives goats as *'ugly' sheep*). But there is only a single *goat/sheep* concept in her mind, whose (sensory) intentionality correlates it to percepts of both *normal sheep* and *ugly sheep*, and whose (inceptional) intentionality correlates it with the verbalization "sheep".



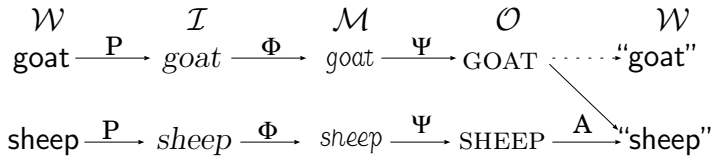
- Aletheia *does* mentally distinguish goats from sheep, but has poor eyesight. She has distinct concepts of *goat* and *sheep*, with distinct perceptual intentionalities (correlating respectively to percepts of *goat* and of *sheep*), and distinct inceptional intentionalities (the verbalizations "goat" and "sheep"). Hence, her perceptual and conceptual intentionalities are like ours, but her *sensory* intentionality correlates the *sheep* percept to worldstates manifesting either sheep or goats (because of poor eyesight).



- Aletheia has good eyesight and mentally distinguishes goats from sheep, but she speaks a unique dialect of English where the word SHEEP is an abstract term (like 'ungulate' or 'mammal') which applies equally well to goats or sheep. Hence, her sensory, perceptual, and conceptual intentionalities are like ours, but her *inceptional* intentionality correlates both the *goat* concept and the *sheep* concept with the verbalization SHEEP.



- Aletheia has good eyesight and distinguishes goats from sheep both mentally and linguistically. However, she has a rare speech impediment which causes her to often mix up certain words. In particular, she often says "sheep" when she *means* to say "goats". Hence, her sensory, perceptual, conceptual, and inceptional intentionalities are like ours, but her *behavioural* intentionality correlates the utterance "sheep" with both the desire (ie. incept) to say SHEEP and the incept to say GOAT.



This 'pentapartite' account of intentionality obviates some of the problems of standard 'causal' accounts of mental representation. In the standard account, we say a particular the mental state represents sheep if it is *caused* by the sight of sheep, or *causes* utterances involving sheep. This account is unsatisfactory because it cannot account for perceptual or verbal errors.

For example, suppose that, at a particular distance D, under particular light conditions L, Aletheia *always* mistakes goats for sheep. Then the causal account says that Aletheia's *sheep* concept has the intentionality, '*either* sheep, *or* goats-seen-at-distance-D-under-light-conditions-L'. This absurd conclusion is called the *disjunction problem*[7]. With pentapartite intentionality, we can assign disjunctive *perceptual* intentionality to Aletheia (perhaps due to poor eyesight), while still allowing her other four intentionalities to be the same as ours.

For example, Aletheia believes *sheep only eat grass*; this is part of her *conceptional* intentionality involving her *sheep* concept (it somehow involves correlations with to her *grass* concept). Aletheia also believes *goats eat everything* —this is part of the conceptional intentionality of her *goat* concept. Her *goat* concept is (perceptually) correlated with her *goat* percept, and the *sheep* concept is (perceptually) correlated with her *sheep* percept, which is (sensorily) correlated with both sheep and goats-seen-at-distance-D-under-light-conditions-L. However, this does *not* mean that Aletheia believes that *'goats-seen-at-distance-D-under-light-conditions-L only eat grass.'*

Of these five kinds of intentionality, *conceptual* intentionality plays a special role, because it is the only way we can assign meaning to abstract or imaginary concepts. Aletheia's

concepts of *prime number* or *dragon* obtain their intentionality from their relationships to other concepts in her mind, *not* from relationships with real world states. We might assign *perceptual* intentionalities to the *prime number* and *dragon* concepts (for example, the sight of the numerals '2','3','5','7','11', etc. or pictures of dragons), but Aletheia may still have these concepts even if she has never *seen* a numeral or a picture (for example, she is blind, or lives in a preliterate society).

Likewise, we might characterize *prime number* or *dragon* with *behavioural* intentionalities (for example, Aletheia's tendency to verbally agree with statements like, '13 is a prime number' or 'Dragons breath fire', and verbally disagree with statements like '9 is a prime number' or 'Dragons have pink feathers'). But Aletheia may still *think* about prime numbers and dragons even if she is entirely paralyzed by a neuromuscular disease. We might dodge the 'paralysis' quibble by looking at *inceptional* intentionalities. But an entirely inceptional characterization is unsatisfactory, since it fails to describe the cascade of mental activity which seems to be an essential part of the 'meaning' of the *prime number* or *dragon* concepts. For example, Aletheia may mentally prove a theorem about prime numbers, or daydream a story about dragons, without ever entertaining the desire to express this theorem or story to another person.

So far I have vaguely described intentionality as 'correlation'. To get more precise, recall diagram (**??**) on page **??**. In terms of this diagram, *percepts* are subsets of $\mathcal{I}$, *concepts* are subsets of $\mathcal{M}$, and *incepts* are subsets of $\mathcal{O}$. Recall from Chapter **??** that *macrostates* are subsets of $\mathcal{W}$. Then:

Sensory Intentionality correlates *macrostates* (subsets of $\mathcal{W}$) to *percepts* (subsets of $\mathcal{I}$), via P.

Perceptual Intentionality correlates *percepts* (subsets of $\mathcal{I}$) to *concepts* (subsets of $\mathcal{M}$) via Φ.

Conceptual Intentionality correlates *concepts* other *concepts* via Φ.

Inceptional Intentionality correlates *concepts* (subsets of $\mathcal{M}$) to *incepts* (subsets of $\mathcal{O}$) via Ψ.

Behavioural Intentionality correlates *incepts* (subsets of $\mathcal{O}$) to *macrostates* (subsets of $\mathcal{W}$) via A.

I can't be more specific than this. For example, we *cannot* say, 'The perceptual intentionality of mental state $m_1 \in \mathcal{M}$ *is* percept $i \in \mathcal{I}$'. Recall that percepts trigger mental states *in combination* with preexisting mental states. Thus, we *can't* say, 'Percept $i \in \mathcal{I}$ causes mental state $m_1 \in \mathcal{M}$'. Instead, we must say 'Percept i, together with mental state m_0, causes m_1' —in other words, $\phi(i, m_0) = (m_1, o)$ (where $o \in \mathcal{O}$ is some incept). But since ϕ is a stochastic

function, we can't even rightly say this; we can only say, 'Percept i, together with mental state m_0, has a *high probability* of triggering a mental state inside the subset $\mathcal{M}_1 \subset \mathcal{M}$' —in other words, the probability of the subset $\mathcal{M}_1 \times \mathcal{O} \subset \mathcal{M} \times \mathcal{O}$ is (relatively) high, with respect to the distribution $\Phi(i, m_0)$. Even *this* is a hedge, since 'relatively high' is a necessarily vague term.

In short, the 'intentionality' of a mental state m_1 is *not* a set of clear and distinct 'links' between m_1 and specific percepts, incepts, or other mental states. Instead, the intentionality of m_1 is the entire context of correlations —both strong and weak —which exist between m_1 and every percept, every incept, and every concept, both prior and subsequent

I have said that a 'concept' is a subset of $\mathcal{M}$. For example, the concept of the number 2 is just the set $\mathcal{C}(\mathsf{Two}) \subset \mathcal{M}$ of all mental states where I am thinking about the number 2. Specific propositions involving the number 2 are subsets of $\mathcal{C}(\mathsf{Two})$ —for example, the concept of the proposition 'Two is an even number' corresponds to the subset $\mathcal{C}(\text{'Two is an even number'}) \subset \mathcal{C}(\mathsf{Two})$ of all mental states where I am thinking about this proposition. Attitudes towards this proposition (ie. 'propositional attitudes') are then further subsets. For example, the propositional attitude, 'I believe that two is an even number' is a subset of $\mathcal{C}(\text{'Two is an even number'})$ —the set of all mental states where I believe 2 is even.

This notion of 'concept' seems too liberal, because *any* subset of $\mathcal{M}$ qualifies, including bizarre and arbitrary ones. For example, one 'concept' is the set $\mathcal{C}$ of all mental states where a particular, arbitrary family of ten million neurons in my brain are firing. To put it another way: if an alien were examining my mental dynamics, how would she know which subsets of $\mathcal{M}$ to identify as 'real' concepts, and which to ignore?

One solution is to call subset $\mathcal{C} \subset \mathcal{M}$ a 'concept' *only* if inceptional intentionality correlates $\mathcal{C}$ strongly to specific speech acts (ie. the utterance, 'Two is an even number'). This is too restrictive, however, because we all possess 'private' concepts we find difficult or impossible to articulate in language.

Another solution is to recognize that only certain subsets of $\mathcal{M}$ qualify as *predictively useful* concepts, in the sense that they strongly correlated with particular percepts, incepts, worldstates, or other mental states. The 'concept' involving the arbitrary ten million neurons is probably not useful: depending on the activity of the other ten billion neurons in my brain, this 'concept' could precede or succeed pretty much *any* mental state, and could coincide with *any* percept or incept. However, a concept like the number 2 is likely to be noticeably correlated with certain percepts (ie. the perception of pairs of objects), verbalizations ('You two make a nice couple') or even other concepts (integer, even, prime, etc.)

This criterion of 'predictively useful' is vague, because the concept of 'strongly correlated' is vague. Depending upon what sorts of correlations you want, different concepts may appear to be the 'predictively useful' ones.

6 What is Identity?

You will soon die. That is, the person you are today is not the person you will be tomorrow. The 'you' of today will be replaced by someone slightly different. Everything you learn or experience changes you, and every time you change, you lose who you were and become someone else.

On short timescales, these changes are negligible, and do not challenge our sense of identity. On longer timescales, though, the little changes accumulate. Consider the experience of meeting an old friend after many years apart, and suffering a dismaying unfamiliarity. They aren't the same person at all, it seems. Or perhaps *you* aren't. These changes in identity are especially pronounced when someone undergoes a traumatizing experience, like a bout of mental illness or drug addiction.

Languages also evolve over time, to the dismay of those defending their linguistic identity against 'foreign influence'. The *Institute de la Langue Français* decries the 'corruption' of the French language by anglicisms like *le hamburger* and *le cellphone*. But languages exchange words as promiscuously as bacteria trade plasmids. English began as a hybrid of Gaelic, Anglo-Saxon and Norman French, and virtually all 'literary' English vocabulary derives from French parlance. Furthermore, almost all our technical jargon and neologisms are Greek or Latin imports. Does English become 'less English' when we import words like *samizdat*, *zeitgeist*, *bazaar* or *hibachi*?

Cultures have even more fluid identities. Are cultures susceptible to 'engineering'? Can mass ideology be molded through propaganda? Can demand be 'manufactured' by advertising? Does 'American cultural imperialism' (MTV, Hollywood, McDonalds, etc.) threaten other cultures with extinction, or merely influence their evolution?

Our intuitive sense of a persistent identity (personal, linguistic, or cultural) seems real and incontrovertible. How can we reconcile this sense with the reality of continual change? Given two systems $\mathcal{S}$ and $\mathcal{S}'$ (say, two persons, two languages, etc.), we ask: how similar need they be in order to be deemed 'the same'? And how can $\mathcal{S}$ remain 'the same' while evolving in response to external influences?

To preserve its identity, it seems that $\mathcal{S}$ must change *autonomously* in response to external influence. For example, learning a new fact changes you but does not disrupt your identity, because the change is autonomously controlled. Brain damage, surgical intervention, or narcotics, however, *do* disrupt your identity because you cannot control the change they inflict.

A **dynamical control system** is a system which changes its internal state autonomously, in a response to external input. The system has a space of internal states $\mathcal{M}$, and a space of inputs, $\mathcal{I}$; if the system is in state $m \in \mathcal{M}$ and experiences input $i \in \mathcal{I}$, it will change to some new state m', which is entirely determined by i and m. Mathematically, this is

described by a function

$$\phi : \mathcal{I} \times \mathcal{M} \longrightarrow \mathcal{M} \qquad \text{so that } m' \;=\; \phi(i, m)$$

(we already introduced this function in Chapter 5 §(i) —see page 46).

For example, the system might be a car; in this case, $\mathcal{M}$ is the current position and velocity of the car, and $\mathcal{I}$ is the set 'inputs' (steering wheel position, pressure on brake pedal, etc.) which the driver controls. The car's position and velocity evolve in response to the input signals sent by the driver.

In our model, $\mathcal{M}$ is the space of all personalities and mental states of all possible human beings, and $\mathcal{I}$ is the space of all possible sensory inputs. At any moment in time, your entire 'state of mind' is a point $m_0 \in \mathcal{M}$. As you experience sensations $i_0, i_1, i_2, \ldots$ from your environment, your mind evolves through a succession of states $m_1, m_2, m_3, \ldots$. Diagrammatically:

$$\begin{array}{ccccccccc} & i_0 & & i_1 & & i_2 & & i_3 & \ldots \\ & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\ m_0 & \Longrightarrow & m_1 & \Longrightarrow & m_2 & \Longrightarrow & m_3 & \Longrightarrow & \ldots \end{array}$$

The important point is that, although the state m_1 is 'caused' by sensation i_0, it is also determined by your current mindstate m_0. A different person (in state m_0') experiencing the same sensation, would respond in a different way (changing to state m_1').

A state m_* is said to be **reachable** from m_0 if there is a sequence of inputs $i_0, \ldots, i_{n-1}$ so that:

$$\begin{array}{ccccccccccccc} & i_0 & & i_1 & & i_2 & & i_3 & \ldots & \ldots & \ldots & i_{n-1} & \\ & \downarrow & & \downarrow & & \downarrow & & \downarrow & & & & \downarrow & \\ m_0 & \Longrightarrow & m_1 & \Longrightarrow & m_2 & \Longrightarrow & m_3 & \Longrightarrow & \ldots & \Longrightarrow & m_{n-1} & \Longrightarrow & m_* \end{array}$$

Thus, m_* is reachable from m_0 if it represents a person who you could potentially *become*, under the right circumstances. We will indicate this by writing $m \rightsquigarrow m_*$. The set of all points reachable from m is called the **reachable set** of m.

Not all states are reachable from m_0. Indeed, it is safe to say that the reachable sets of any two people are disjoint, because no sequence of experiences could lead one person to actually *become* another. And this is the touchstone of identity. We can say that a mindstate (ie. person) $m_* \in \mathcal{M}$ is 'the same person' as you if $m \rightsquigarrow m_*$ (ie. m_* is a possible 'future self') or if $m_* \rightsquigarrow m$ (ie. m_* is a possible 'past self'). In either case, we write $m \leftrightsquigarrow m_*$. More generally, we could say that m and m_* share the same identity if there is any chain $m \leftrightsquigarrow m' \leftrightsquigarrow m'' \leftrightsquigarrow \ldots \leftrightsquigarrow m_*$.

Clearly, the longer such a chain must be (or the greater the distance in time separating successive links in the chain), the more tenuous the relationship of identity. Nonetheless, it seems a universal article of faith that no two distinct human beings could ever be connected by such a chain; all of us are unique, and no one could ever 'become' someone else. This is not something we can prove, but it is something we certainly like to believe.

Notice that invasive changes (surgery, trauma, etc.) will move the point m to a new point m' which is probably *not* reachable from m; hence, they disrupt the continuity of identity.

The same formalism is applicable to culture, but this entity is far more amorphous. Let's say that a *society* is a population of individuals, each having her own personality. Thus, a society is a set of points in $\mathcal{M}$. Thus, in a society of P individuals can be described by a point in $\mathcal{M}^P$.

Each of these people evolves over time, in response to external influences, and one of the most important of these is her interaction with other people in the society. Other factors also influence the society, such as natural events, or interaction with *other* societies. Thus, the society as a whole can be described by a dynamical control system:

$$\Phi : \mathcal{I} \times \mathcal{M}^P \longrightarrow \mathcal{M}^P$$

One disadvantage of this model is that we must fix a specific population, P (thus, we are excluding births and deaths). A more versatile model is to represent a society as a *probability distribution* over $\mathcal{M}$, which represents the distribution of some very large (but nonspecific) number of points in $\mathcal{M}$. If $\mathbb{P}(\mathcal{M})$ is the set of distributions, then society is described by a dynamical control system of the form

$$\Phi : \mathcal{I} \times \mathbb{P}(\mathcal{M}) \longrightarrow \mathbb{P}(\mathcal{M})$$

Of course, we can't even begin to suggest equations to describe such a complex system. We expect, however, that the probability distribution describing society will exhibit a some degree of *clustering* (like a Gaussian). The reason is that individuals in a society tend to imitate one another, so that their personalities will stay relatively similar over time.

Thus, a *culture* is an (evolving) probability distribution on $\mathcal{M}$, the 'space of personalities'. We can again use the concept of *reachability* to decide whether two cultures are 'the same', and to decide when cultural change is a process of autonomous evolution (analogous to a person learning something), and when it is a traumatic and discontinuous event (analogous to brain damage or surgery).

We could probably develop a similar model of *collective linguistic evolution*: a population of 'speakers', determining an evolving probability distribution on the 'space of all possible languages'.

7 Science

(i) What Science is Not

A lot of metascientific discourse arises from naïve and idealized misconceptions of science. In philosophy of science, these misconceptions yield absurd conclusions. When questions of 'scientific validity' have ideological consequences (eg. creationism vs. evolution, behaviorism vs. cognitivism, economics vs. ecology), these misconceptions become pernicious. Thus, it is important to first identify some things which science generally *isn't:*

SCIENCE IS NOT EXACT: Scientific models are never exact, for three reasons:

System complexity: Most real systems are far to complex for us to have an exact model. If you're studying a living organism, a weather system, or an economy, the best you can hope for is a macroscopic, qualitative model which accounts for the large-scale features of the data. Thus, medicine, meteorology, and macroeconomics are inherently 'inexact' sciences.

Computational Complexity: Even if an exact model exists, rigorous computation within this model is usually impossibly complex. You must always make approximations and idealizations to get any kind of answer. For example, we habitually throw away 'small terms' in an equation. Functions are often approximated via Taylor series[1], and the higher-order terms are then discarded.

Measurement error: Physical measurements are always imprecise. The best you can do is quantify the expected error.

Science yields not exactitudes, but approximations. In special cases (eg. celestial mechanics), these approximations are in fact incredibly accurate. But it spurious to extrapolate from celestial mechanics to the rest of science.

[1] A **Taylor series** is way of representing a function as an infinite polynomial. For example, the *sine* function has Taylor series:

$$\sin(x) \quad = \quad x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \cdots \quad = \quad \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}.$$

It is common to approximate this infinite series with a 'truncated' finite polynomial. For example:

$$\sin(x) \quad \approx \quad x - \frac{x^3}{6} + \frac{x^5}{120}.$$

SCIENCE IS NOT DETERMINISTIC: Classical mechanics is perfectly deterministic, but more recent scientific models are not. There are three reasons:

System complexity: Imagine a one metre cube, filled with rapidly colliding ping pong balls. There is a hatch at the top, through which one ball can escape at a time.

In principle, this system is deterministic (the collisions and trajectories of the balls obey classical mechanics). Thus, we could, in principle, predict the next ball to escape the hatch. *In practice*, the system is far to complex to explicitly model, and we instead treat the next escaping ball as a random event. Indeed, this is often how the random winning number is generated in televised lotteries.

Chaos: Even if you constructed an explicit, deterministic model of the ping pong box, this model would be *chaotic*, in the sense that slightly different initial conditions diverge exponentially and lead to totally different outcomes. Thus, a tiny measurement error (which is inevitable) will rapidly ruin the accuracy of your predictions. Hence, although the model is *in principle* deterministic, it must *in practice* be treated as stochastic; this is the motivation for the *ergodic theory* [30, 43] of dynamical systems.

Intrinsic Randomness: Quantum mechanics says that microphysical processes are *intrinsically* stochastic. It follows: if large scale phenomena are grounded in microphysical processes, then they are also intrinsically random. For example, biological evolution is driven by mutation. The mutation of single nucleotide is a microphysical process, subject to quantum indeterminacy.

SCIENCE IS NOT SYLLOGISTIC: In contrast to the 'logico-deductive' examples common in philosophy of science literature, syllogisms rarely appear in real scientific discourse. Only in the simplest models can you unambiguously 'deduce' conclusions from premises in a syllogistic manner. Instead, usually you take a 'toy model' of the system under consideration, perform some rough calculations, and obtain some numbers which —after suitable interpretation —yield tentative predictions about the original system.

Logico-deductive philosophers of science often offer *medical diagnosis* as a prototypical example of 'scientific syllogism'. For example:

$$\Big(\text{Wet cough \& Yellow sputum} \Big) \Longrightarrow \Big(\text{Strep throat} \Big) .$$

The problem is, this is *not* a syllogism; it is a purely probabilistic statement, which employs an observed statistical correlation between symptoms and syndromes. A syllogism can never be wrong: it can never lead from true premises to false conclusions. But medical diagnoses *can* be wrong, and often are.

SCIENCE IS NOT STRONGLY PREDICTIVE: Notwithstanding the Logical Positivist ideal, real science often does not —indeed, *cannot* —make bold predictions. Part of the reason is that science is often inexact and probabilistic, as we've already seen. For example, meteorologists still can't correctly predict weather more than a few days into the future. This is not because meteorology is 'unscientific' —it's because the atmosphere is a vastly complex chaotic dynamical system.

Some scientific theories do not even *attempt* prediction. For example, Darwin's theory of Natural Selection is *not* a predictive theory. It does not predict how organisms *will* evolve, except in the vague sense that they will either adapt to changing environmental conditions, or die out[2]. Instead, Darwin's theory is 'retrodictive' theory, which explains natural history as a story of adaptation to various selection pressures.

I'll discuss the relationship between prediction and explanation more in §(iv).

SCIENCE IS NOT ONTOLOGICAL: Scientific theories make no claims about the existence of theoretical entities. Theoretical entities come in three flavours:

Elementary units: Scientists often hypothesize elementary units (eg. electrons, voltages) whose properties and interaction are the basis of the theory. Do these really exist?

The ontological status of electrons is as unknowable as it is irrelevant. 'Electron' is the name of a mathematical construct, a component of a model. It is a 'conceptual handle' which we attach to certain variables in our equations, to facilitate our intuitions. The value of the 'electron' construct is measured by the explanations and predictions we can obtain with it. Physicists do not 'believe in' electrons; they believe in the *usefulness* of the 'electron' concept.

Theorist's fictions: Do physicists believe in gravity? The naïve response is, 'Of course they do.' But actually, they don't. Relativistically speaking, there is no 'force of gravity'; there is curved spacetime. Yet the 'force of gravity' is ubiquitous in the discourse of (nonrelativistic) physics. Again, the 'existence' of gravity is irrelevant; its *utility* as a concept is the issue.

Electrons may not exist 'in reality', but at least electrons exist in the standard model of physics. Gravity doesn't even exist in the model. Gravity is a 'fictitious force'; an artifact of a noninertial reference frame, similar to *centrifugal force* (which keeps the water in a swinging bucket), *Coriolis force* (which makes hurricanes spiral), and the 'G-force' felt in an accelerating airplane.

These are examples of *theorist's fictions:* entities which we 'know' don't exist, but which are useful as conceptual aids. Some other common theorist's fictions:

[2]One exception: biologists now confidently predict that pathogenic bacteria *will* eventually evolve resistance to any antibiotic we develop.

Phonons are quanta of mechanical energy which propagate through a crystal lattice. Mathematically, it is convenient to treat them as particles.

Point particles do not exist in quantum mechanics. A quantum 'particle' is, by definition, a spatially distributed object defined by a sort of 'complex probability field' (the wavefunction). Nevertheless, it is often intuitively convenient for physicists to *pretend* that electrons, photons, etc. are point particles.

Rational Maximizers are the agents of microeconomic models. No one believes that real people behave this way.

Selection pressure is the stress of a hostile environment, which 'pushes' a species to evolve in certain directions.

Selfish Genes are genes anthropomorphised with the 'desire' to replicate, and the 'cunning' to innovate and adapt. Of course, real genes are mindless, but the selfish gene is a powerful in the discourse of evolutionary biology.

Anthropomorphisms: The Selfish Gene is one example of a class of theorist's fictions where we attribute mindless objects with desires or beliefs. For example, we commonly say that a thermodynamic system 'wants' to attain the state of minimal free energy. A moth 'wants' to fly in a straight line; it keeps circling the lamp because it 'thinks' the light is the sun.

The value of a theorist's fiction lies in compressing a complex bundle of ideas into an easily managed abstraction.

Collective/emergent entities: Systems with many interacting components often exhibit *emergent phenomena* which persist in time and entrain the coherent collective activity of millions of units. It is useful to give names to these collective phenomena, and develop models of their behaviour. Consider the following examples. In each case, the reality of the entity seems incontrovertible from a distance, but becomes fuzzy up close.

Statistical mechanics: Temperature, pressure, and current are *collective* quantities, obtained by averaging over large populations of molecules:

- 'Temperature' is *average kinetic energy* of a population of molecules. Individual molecules do not have a 'temperature' in a thermodynamically meaningful way.
- 'Current' is *average velocity* of a population of molecules; 'pressure' is proportional to the *average deviation* from this average velocity. An individual water molecule has a *velocity*, but it does not have a 'current' or 'pressure' in a hydrodynamically meaningful way.

Hurricanes: Where, exactly, is the physical boundary of a hurricane? When does it stop being a hurricane, and revert to a mere tropical storm?

Species: A 'species' is a population of organisms with similar genomes. But what is 'similar'? For sexually reproducing species, we can apply the test of 'mating to yield fertile offspring', but this is inapplicable for asexual species. On evolutionary timescales, when exactly does an old species cede to its successor?

Ecosystems: An 'ecosystem' is a self-contained web of interacting organisms. But how do we delimit this web? Any two species on earth are connected by *some* chain of interactions, so the smallest 'self-contained' ecosystem is in fact the whole biosphere.

Organisms: Even organisms are fuzzy, up close. Where is the boundary between self and nonself? When does the food you eat stop being 'foreign matter' and become part of you? Are your gut bacteria part of you? What about the mitochondria living symbiotically in your cells? Is an endogenous retrovirus 'part' of its host bacterium? Is a plasmid? What if it confers resistance to an antibiotic?

Life: Are memes alive? It seems not, since they can only exist and replicate within a suitable *host:* the mind of a human. But by the same argument, a bacteriophage virus is not alive, since it can only replicate in the context of as host bacterium. Likewise, mitochondria are not alive, since they can only exist and replicate within a host eukaryote. But by the same argument, *humans* aren't alive, since we can only exist and replicate in the context of *our* host organism, the biosphere.

These questions probably do not have meaningful answers. We are looking too closely at the boundaries of inherently fuzzy concepts. I am certainly *not* saying that temperature or species or hurricanes 'don't exist'. They *do* exist, as collective phenomena —in other words, as abstractions. But they *don't* exist in the same precise, concrete, delimited fashion as, say, electrons (to the extent that electrons exist).

Collective entities are just a particular kind of theorist's fiction; a white lie which facilitates comprehension of a complex theory. Indeed, even elementary units —being mere 'conceptual handles' for certain variables —are a theorist's fiction. And that is exactly the point.

SCIENCE IS NOT NEATLY CAUSAL: The classic model of scientific explanation involves a 'causal process' leading inextricably from initial conditions to observed consequences. Real science never works this way, for three reasons.

Indeterminism: Causality is usually seen as a kind of logical necessity:

$$\Big(\text{ Cause } \Big) \Longrightarrow \Big(\text{ Effect } \Big)$$

But science is usually *probabilistic*; initial conditions do *not* logically necessitate final outcomes, but at most influence their probability. We might develop a 'probabilistic' version of causality, as advanced by Suppes[41], but we then lose the neat distinction between causation and correlation.

A tangled skien: Causal processes are imagined as having an unambiguous 'path of propagation'. In simple physics models, you can sometimes trace the propagation of energy and information unambiguously from a 'cause' to an 'effect'. In more complex models, however, everything impinges upon everything else, and the causal web is so tangled as to render meaningless assertions of the form '**A** caused **B**'.

For example, in chaotic systems, tiny perturbations can 'cause' radically divergent outcomes. Can we really blame the proverbial butterfly in Rio for 'causing' the hurricane in Bermuda? A million other perturbations could also have 'caused' it. And further million could have prevented it, or exacerbated it, or delayed it, etc. At a certain point, it is meaningless to ask what 'caused' the hurricane.

Achronality: 'Causality' implies a process unfolding in time. As we will see in §(ii), many scientific models are *achronal*: there is no notion of time, hence, no notion of temporal causality.

SCIENCE IS NOT REDUCTIONIST: Science exhibits a rough hierarchy of *abstraction levels*, with lower levels providing the conceptual foundations for higher levels[3] (Figure **??**A). In theory, the 'axioms' of any biochemistry theory should be deducible as 'theorems' of quantum physics. In practice, however, this vision is impractical to realize. Even the simplest chemical system —say, a single water molecule —is far too complex to be rigorously mathematically analyzed when expressed in quantum terms.

We will probably never reduce theoretical biochemistry to a mere 'logical corollary' of physics. But this is unnecessary for the coherence and success of the scientific project. Biochemists don't *need* physics; their theories are intended to predict and explain *biochemical* phenomena. It would be intellectually satisfying to precisely formulate biochemistry in terms of physics, but it is not necessary for biochemists to do their job. It *is* useful, however, to at least roughly situate biochemical phenomena in the framework of quantum physics, so that biochemists can lean on this framework when necessary. For example, it is useful to know that, *in principle*, a DNA molecule can be modelled as a quantum system, even if such a model is far to complex to analyze in practice.

[3]This may seem rather backwards; after all, isn't physics *more* abstract than biology? But I am using 'abstraction level' in the sense of computer science, where software at a *lower* abstraction level (eg. the operating system) provides the 'platform' to build software at higher abstraction levels (eg. the word processor). In this sense, quantum physics is the 'platform' for quantum chemistry.

This *separation of abstraction levels* is crucial, because it allows theoretical evolution and revolution to occur independently at various levels. It is not necessary to have a completely articulated quantum theory to begin formulating a biochemical theory; nor is it necessary to understand all the minutiae of molecular biology before one can study the biology of cells. Furthermore, a scientific revolution at one level (say, the rejection of the Standard Model in physics) need not shatter the intellectual paradigms at other levels.

(ii) Scientific Models

It is important to distinguish between theories and models. A **model** is a mathematical construct used to explain, predict, or describe a *specific system*. A **theory** is a framework for the construction of models. I will discuss scientific theories later. Since theories exist to create models, it is useful to first discuss the kinds of models which exist in science.

I will describe these models as mathematical objects. This is not to say that all scientific models *are* formulated mathematically, or even that they *should* be, but only that, in principle, they *could* be. Note that 'mathematical' does *not* mean 'quantitative'. Mathematical models do not necessarily involve equations and numbers; they may instead involve precisely defined *qualitative* concepts, linked by relationships of logical implication or probabilistic correlation. For example, diseases and symptoms are (usually) not numbers; however, a medical diagnostic model *could* be formulated in terms of the correlations between certain symptoms and certain diseases[4].

Scientific theories fall into four broad (and nonexclusive) categories: *dynamical systems*, *stochastic processes*, *equilibrium models*, and *achronal models*.

1. Dynamical Systems are *deterministic* models of a system evolving in time. The initial conditions of the system define a point in a statespace $\mathcal{W}$. Each point in $\mathcal{W}$ lies on a unique *trajectory*, which tells us exactly how the future will unfold (Figure 7.1A).

'Events' correspond to subsets of $\mathcal{W}$. 'Causality' manifests as follows: an event $\mathcal{A}$ at time 0 'causes' an event $\mathcal{B}$ at time t if every point in $\mathcal{A}$ has a trajectory that passes through $\mathcal{B}$ at time t (Figure 7.1B)

Dynamical systems are usually formulated in terms of ordinary differential equations –the prototypical example is classical mechanics —or in terms of partial differential equations, eg. the Heat Equation or Navier-Stokes Equations.

2. Stochastic Processes are *nondeterministic* models of a system evolving in time. Stochastic processes are defined by assigning a *probability* to every possible 'history' the

[4]Of course, this is not how doctors think about medical diagnosis. However, this is *exactly* how computerised 'expert systems' perform medical diagnosis, and some of these systems perform as well, or better, than human doctors.

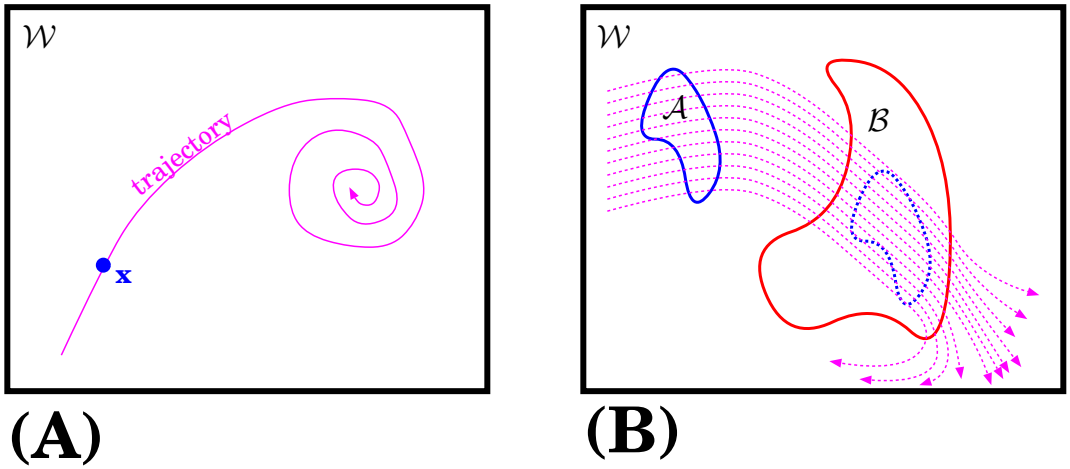


Figure 7.1: **(A)** In a dynamical system, each point $\mathbf{x}$ in the state space $\mathcal{W}$ has a unique *trajectory*, which describes its past and future. **(B)** Causality: Subset $\mathcal{A}$ flows into subset $\mathcal{B}$.

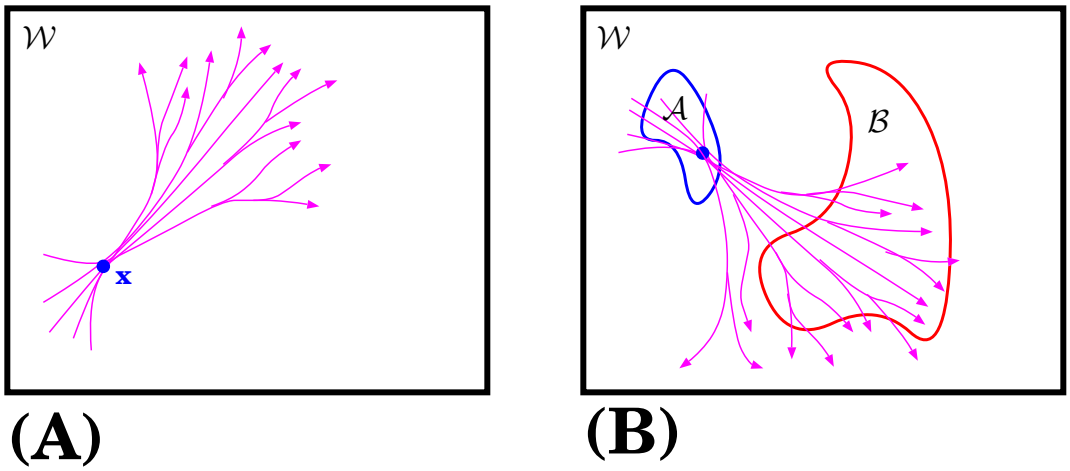


Figure 7.2: **(A)** In a stochastic process, each point $\mathbf{x}$ in the state space $\mathcal{W}$ has a probability distribution of possible *trajectory*, which describes its possible pasts and futures. **(B)** Causality: 'Most' of the probability from subset $\mathcal{A}$ flows into subset $\mathcal{B}$.

system's evolution over time –in other words, to every possible trajectory in $\mathcal{W}$ (Figure 7.2A).

There is only a vague notion of 'causality' in a stochastic process. An event **A** at time 0 'causes' an event **B** at time t if a trajectory which passes through **A** at time 0 is *highly probable* to pass through **B** at time t (Figure 7.2B).

For example:

Quantum mechanics is a deterministic theory (described by the Schrödinger equation) until the moment a measurement is taken, at which point the wavefunction 'collapses' in a random manner.

Natural selection is a stochastic process where a large population of genetically distinct replicators experience random mutations, which enhance or degrade their replication abilities. Over time the population as a whole evolves to favour genomes of greater 'fitness' (ie. replication ability), but *which* adaptations will be favoured is unpredictable, except on the most trivial 'fitness landscapes'.

A dynamical systems is special case of a stochastic process: one where all trajectories have probability 0 or 1.

3. EQUILIBRIUM MODELS describe a system which has attained a final rest state, and is *not* evolving in time. There is usually only a vague qualitative description of how the system arrived at equilibrium, how long it took, or what path it followed. For example:

Laplace's Equation describes the equilibrium concentration of a diffusing chemical, or an equilibrium temperature distribution.

Classical thermodynamics describes the final state of a closed thermodynamic system.

Microeconomics describes equilibrium allocation of resources and the prices of goods in a perfect market.

Kirchoff's Laws determine the currents and voltage gaps through the components of an electric circuit, and yield a static model of circuits made from simple components (eg. resistors, batteries) which converge to equilibrium 'almost instantaneously'. (Kirchoff's laws yield a dynamical model when we include slowly equilibrating units like capacitors).

Natural Selection, in Darwin's original formulation, can be seen as an 'equilibrium' model, because (unlike the 'stochastic' model) it provided only a vague qualitative description of *how* organisms evolve. The focus was on explaining their *current* form as the *equilibria* of a process of adaptation.

Many scientific fallacies arise from applying equilibrium models to nonequilibrium situations. For example:

- Classical thermodynamics does *not* apply to systems displaced from equilibrium, such as living organisms. Nevertheless, 'creation scientists' spuriously apply the Second Law of Thermodynamics to 'refute' Darwinism.

- The relevance of 'long term' market equilibria in *real* economies is questionable; this 'long term' may be too long a wait for the victims of market imperfections, and indeed may be so long that market conditions themselves change before the equilibrium is reached. As Keynes said, 'In the long term, we're all dead.'

4. ACHRONAL MODELS involve no explicit representation of time. An equilibrium model is obviously achronal. Other examples include:

Fermat's Principle: In this formulation of optics, a light ray travelling through an optical medium always 'chooses' the path which *minimizes total travel time*. From this premise, you can derive the usual laws of refraction, reflection, etc. Fermat's Principle seems 'chronal' (since it explicitly involves 'travel time'). However, it describes the light ray as an entity living 'outside of time', which first computes the travel time of all possible trajectories, and then picks the minimal one. Applying Fermat's Principle is more like solving a (timeless) optimization problem, rather than observing an unfolding evolution. (Of course, once we have the solution, we *interpret* it as a chronal trajectory.)

Lagrangian mechanics extends Fermat's principle to classical mechanics. In this formulation, a mechanical system 'chooses' the trajectory through statespace which minimizes an aggregate quantity called the *action*. Again, the formalism is inherently achronal, (although the solution describes a chronal process).

General relativity: A relativistic four-manifold possesses only a vague notion of 'time'. When we impose a local coordinate system on the manifold, we often identify a certain coordinate as 'time', but this is just a conceptual aid, and has no physical meaning. There is only a 'local' notion of time:

- At every point in spacetime, there is a *forward lightcone*, which confines the trajectory of any particle. This enforces a weak form of causality, by preventing particles from travelling 'backwards in time' and colliding with their past selves or otherwise 'changing history'.
- Each particle has its own 'subjective time': the flow of time you would experience if you 'were' that particle.

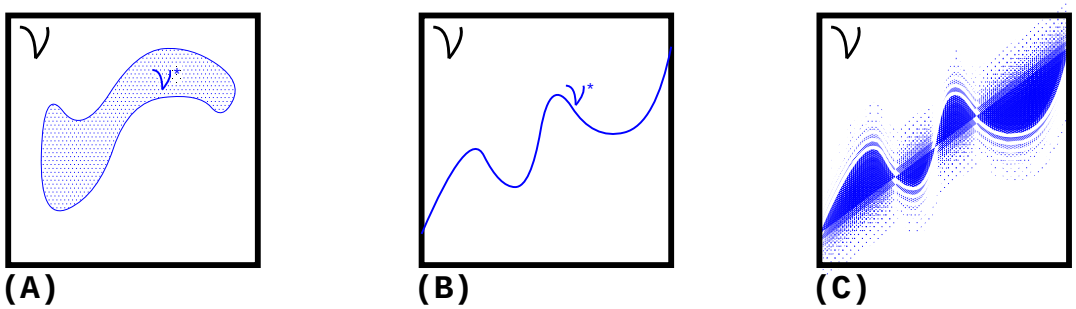


Figure 7.3: The space $\mathcal{V}$ of models **(A)** The constraints of a deterministic theory delimit a subset $\mathcal{V}^*$ of 'admissible' models. **(B)** If these constraints take the form of equations, then $\mathcal{V}^*$ will be a smooth variety within of $\mathcal{V}$. **(C)** The constraints of a probabilistic theory determine a probability measure on $\mathcal{V}$.

However, there is no global absolute measure of time.

In this model, matter and spacetime interact in an *achronal* manner. Contrary to popular belief, matter does not 'cause' spacetime to curve, nor does curved spacetime 'deflect' the trajectory of matter. Instead, the distribution of matter and energy throughout spacetime (described by the *stress-energy tensor*) must be 'compatible' with the curvature of spacetime (described by the *Einstein tensor*). The compatibility condition is the Einstein field equation. This equation has certain solutions (eg. the Robertson-Walker model), which we can *interpret* as an 'evolution in time'. But it is often more natural to view a solution as a four-dimensional, 'timeless' object.

(iii) Scientific Theories

A scientific theory provides a 'conceptual vocabulary' with which we construct models. This vocabulary usually consists of a collection of prototypical **objects** (eg. particles and fields; consumers and firms, etc.), each of which is defined by a (possibly infinite) bundle of (usually numerical) **attributes** (eg. mass, energy, vector fields, supply/demand curves). We define a model by postulating relationships between these prototypical objects —ie. by arranging them in some **configuration**. Let $\mathcal{V}$ be the space of all possible configurations; hence, $\mathcal{V}$ is the *space of models* of the theory. Each point in $\mathcal{V}$ corresponds to a specific, concrete model, a sort of 'virtual universe'.

The theory also imposes **constraints** on the interactions between objects. The constraints are usually equations or 'laws', which rule out most configurations (ie. elements in $\mathcal{V}$) as 'impossible'. Thus, the constraints define a subset $\mathcal{V}^* \subset \mathcal{V}$ of 'admissible models' (Figure 7.3A). For example:

- In a numerically quantified theory (eg. particles with quantitative attributes like position, momentum, energy, etc.), a model is specified by a (possibly infinite) list of numerical parameters. Thus, we imagine $\mathcal{V}$ as an infinite-dimensional space. The constraints form an (infinite) system of equations relating these parameters (eg. conservation laws), and $\mathcal{V}^*$ is the solution set to these equations; thus, $\mathcal{V}^*$ is a *smooth variety* (like a curve or a surface) inside $\mathcal{V}$ (Figure 7.3B).

  If we know the values of some attributes (eg. the momentum of object $\mathbf{x}$), we can use the constraint equations to solve for other attributes (eg. the momentum of $\mathbf{y}$). This is an example of *scientific inference* (see §(iv)).

- In a *time-dependent* theory (eg. a dynamical system or stochastic process) a point in $\mathcal{V}$ does not represent a single moment, but instead an entire *trajectory* —that is, a complete 'history of events'. Thus, $\mathcal{V}$ is the 'space of all possible histories'. The constraints impose relationships between *earlier* events and *later* ones, so that we can *predict* the future (or *retrodict* the past) through scientific inference.

- In a *probabilistic* theory (eg. a stochastic process), the constraints manifest as a *probability distribution* on $\mathcal{V}$ (Figure 7.3C). This distribution dictates, that certain configurations are much more likely than others. A deterministic model is the special case when all models are assigned either probability zero or one.

- Although most models are defined using 'objects' and 'attributes', this is not necessary. What's important is that the theory defines a space $\mathcal{V}$ of possible models; the language of objects and attributes simply provides one 'coordinate system' whereby we can identify points or regions in $\mathcal{V}$.

Finally, the theory provides **correspondence rules** to relate the models to empirical phenomena (for example, via 'operational definitions'). The correspondence rules tell us how to translate data from real measurements into the theory vocabulary, and how to translate inferences in the theory back into empirically testable predictions.

Example: (*Newtonian mechanics*)

The objects of this model are *particles* and *force fields*. A *particle* is described by one constant (its mass) and 6 variables (three position, three momentum). Thus a 'particle' is an element of $\mathbb{R} \times (\mathbb{R}^3 \times \mathbb{R}^3)$, and its trajectory is a function from $\mathbb{R}$ into $\mathbb{R}^3 \times \mathbb{R}^3 = \mathbb{R}^6$. The *force field* between two particles i and j is a 3-dimensional vector, which changes as a function of their respective positions and velocities, and also of time; hence, it is a function $f_{ij} : \mathbb{R} \times \mathbb{R}^6 \times \mathbb{R}^6 \longrightarrow \mathbb{R}^3$.

A model with n particles thus consists of n trajectories in $\mathbb{R}^6$ —or equivalently, *one* trajectory in $\mathbb{R}^{6n}$ —along with n^2 force fields $f_{ij} : \mathbb{R} \times \mathbb{R}^6 \times \mathbb{R}^6 \longrightarrow \mathbb{R}^3$.

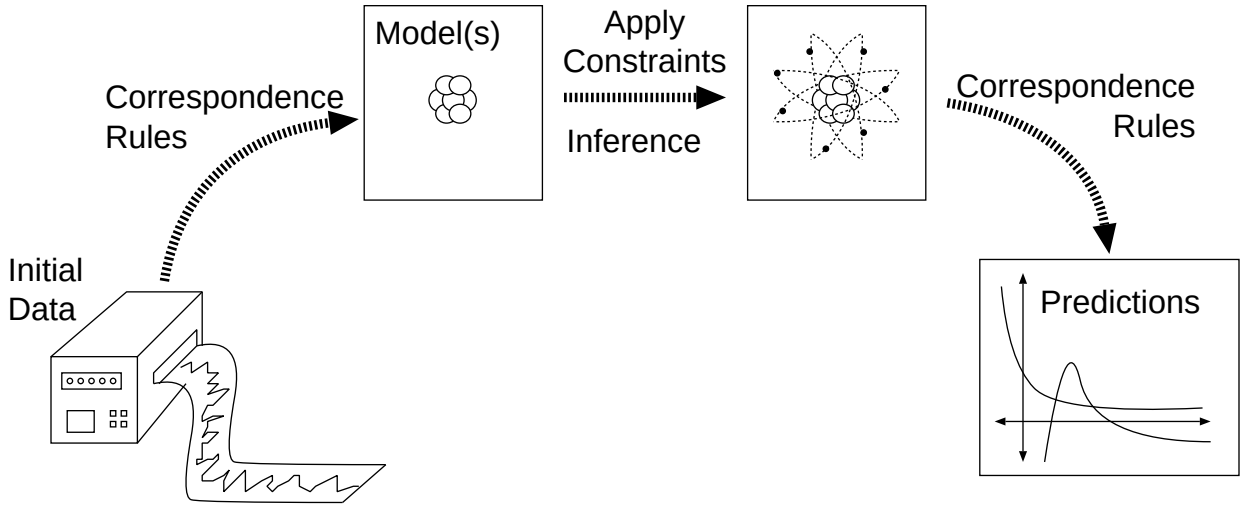


Figure 7.4: The process of scientific inference

Let $\mathcal{T}$ be the space of all system trajectories[5], and let $\mathcal{F}$ be the space of all interparticle force fields[6]. Thus, a model of n interacting particles consists of an element $\mathbf{t} \in \mathcal{T}$, and n^2 force fields $f_{ij} \in \mathcal{F}$; hence if $\mathcal{V}_n$ is the space of n-particle models, then $\mathcal{V}_n = \mathcal{T} \times \mathcal{F}^{n \times n}$. Hence, the space of *all* models (having any finite number of particles) is $\mathcal{V} = \bigcup_{n=1}^{\infty} \mathcal{V}_n$.

The constraints are Newton's Three Laws. The Third Law says that any particles i and j exert equal but opposite forces on one another: $f_{ij} = -f_{ji}$. The other two Laws dictate the particle trajectories; once the force fields and initial positions and velocities of the particles are specified, their trajectories are uniquely determined *for all time*. Hence, the vast majority of trajectories (elements of $\mathcal{T}$) are impossible; the set of 'admissible' trajectories is small enough that we can use the state of the particles at time zero to *infer* their state at later times.

The correspondence rules are pretty physically intuitive ('A force is a push or a pull', etc.). Toy models include: the simple pendulum and the two-body gravitational system.

(iv) Scientific Inference

The process of scientific inference involves four steps, represented in Figure 7.4

1. Begin with initial data from past observations or experiments.

2. Apply the *correspondence rules* to isolate the set of models in $\mathcal{V}$ which fit this data.

[5]Trajectories are smooth functions from $\mathbb{R}$ into $\mathbb{R}^{6n}$, so formally, $\mathcal{T} = \mathcal{C}^{\infty}\left(\mathbb{R};\ \mathbb{R}^{6n}\right)$.

[6]A force field is a smooth functions from $\mathbb{R} \times \mathbb{R}^6 \times \mathbb{R}^6 = \mathbb{R}^{13}$ into $\mathbb{R}^3$, so formally, $\mathcal{F} = \mathcal{C}^{\infty}\left(\mathbb{R}^{13};\ \mathbb{R}^3\right)$.

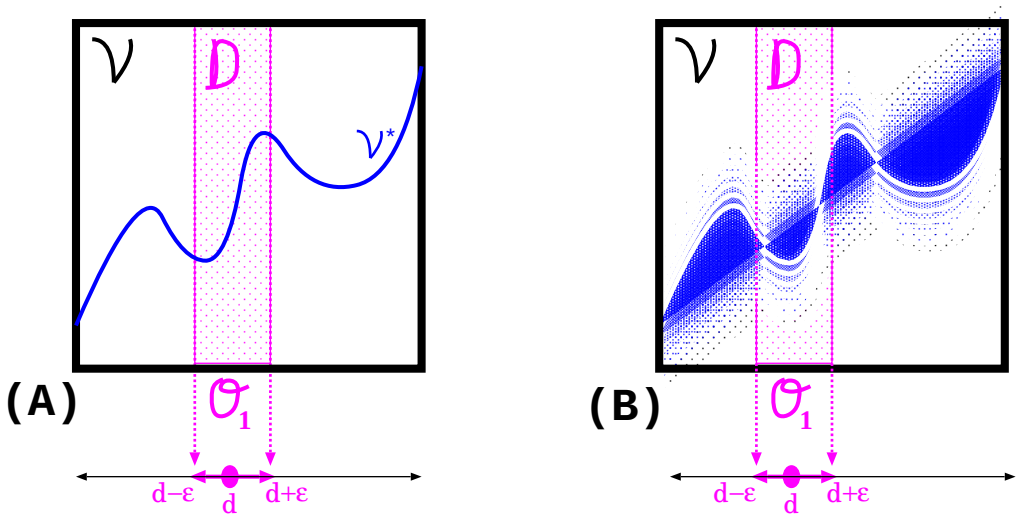


Figure 7.5: Given empirical data d (with experimental error ϵ), we consider the subset $\mathcal{D} = \{v \in \mathcal{V} \; ; \; d - \epsilon \; < \; \mathcal{O}(v) \; < \; d + \epsilon\}$ of all possible models consistent with this data.

3. Apply the *constraints* to determine/estimate the unknown parameters of these models.

4. Apply the *correspondence rules* to translate these estimates into empirical predictions.

To represent this mathematically, let's start with an example. Imagine an empirical measurement which produces a single real number as output (say, a temperature). For any possible model $v \in \mathcal{V}$, the interpretation scheme should *predict* a value for this measurement. In other words, the interpretation scheme defines a function $\mathcal{O} : \mathcal{V} \longrightarrow \mathbb{R}$. If $\mathcal{O}(v) = d$, this means, 'If the universe is in the state corresponding to the point $v \in \mathcal{V}$, and you perform this measurement, you will get value d.'

In this way, every possible measurement, test, or observation is identified with a function $\mathcal{O} : \mathcal{V} \longrightarrow \mathbb{R}$, which we will call an **observable**[7]. The *correspondence rules* therefore take the form of an (infinite) collection of observables $(\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \ldots)$ which correspond to our repertoire of empirical tests and measurements.

Our initial data is a a set of measurements, which fixes the value of some of these observables. For example, if we made three measurements, corresponding to observables $\mathcal{O}_1$, $\mathcal{O}_2$, and $\mathcal{O}_3$, and obtained measurement values d_1, d_2, and d_3, then our initial our initial data consists of the equations:

$$\mathcal{O}_1(v) = d_1; \qquad \mathcal{O}_2(v) = d_2; \qquad \text{and} \qquad \mathcal{O}_3(v) = d_3.$$

[7]For simplicity, we assume the observable yields numerical output —ie. that $\mathcal{O}(v)$ is a real number. But this is hardly necessary, and we could replace $\mathbb{R}$ with any other space.

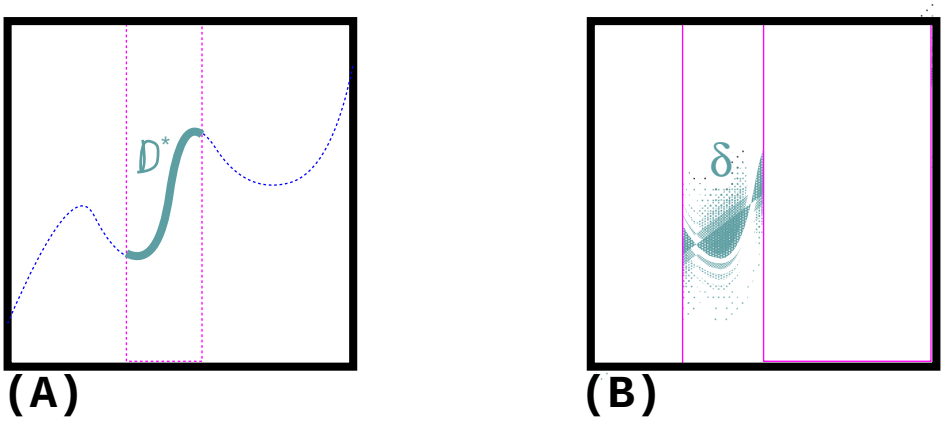


Figure 7.6: **(A)** Deterministic constraints restrict $\mathcal{D}$ to a subset $\mathcal{D}^*$. **(B)** Probabilistic constraints yield a probability measure δ on $\mathcal{D}$.

This data picks out a subset of $\mathcal{V}$ –namely, the set $\mathcal{D}$ of all models $v \in \mathcal{V}$ satisfying these equations. That is,

$$\mathcal{D} \;=\; \Big\{ v \in \mathcal{V} \;;\; \mathcal{O}_1(v) = d_1, \quad \mathcal{O}_2(v) = d_2, \quad \text{and} \quad \mathcal{O}_3(v) = d_3 \Big\}$$

In other words, $\mathcal{D}$ is the set of all possible models which *could* have yielded the measurements (d_1, d_2, d_3). Realistically, however, measurements never yield *exact* values, but always come with some error. In other words, our data will have the form:

$$\mathcal{O}_1(v) = d_1 \pm \epsilon_1; \qquad \mathcal{O}_2(v) = d_2 \pm \epsilon_2; \qquad \text{and} \qquad \mathcal{O}_3(v) = d_3 \pm \epsilon_3.$$

where $\epsilon_1, \epsilon_2, \epsilon_3$ are worst-case measurement errors. Hence, the correct collection of models is the set

$$\mathcal{D} \;=\; \Big\{ v \in \mathcal{V};\; (d_1 - \epsilon_1) < \mathcal{O}_1(v) < (d_1 + \epsilon_1), \qquad (d_2 - \epsilon_2) < \mathcal{O}_2(v) < (d_2 + \epsilon_2), \\ \text{and} \quad (d_3 - \epsilon_3) < \mathcal{O}_3(v) < (d_3 + \epsilon_3) \Big\} \qquad \text{(see Figure 7.5).}$$

Now we apply the constraints.

- In a *deterministic* theory, the constraints will *exclude* most elements of $\mathcal{D}$ as 'impossible', leaving us with a subset $\mathcal{D}^* \subset \mathcal{D}$ (see Figure 7.6A). This is the set of models of the initial data which are compatible with the theory.

- In a *probabilistic* theory, the constraints will determine a probability distribution δ on $\mathcal{D}$ (see Figure 7.6B). The distribution δ says that some models in $\mathcal{D}$ are 'more likely' than others.

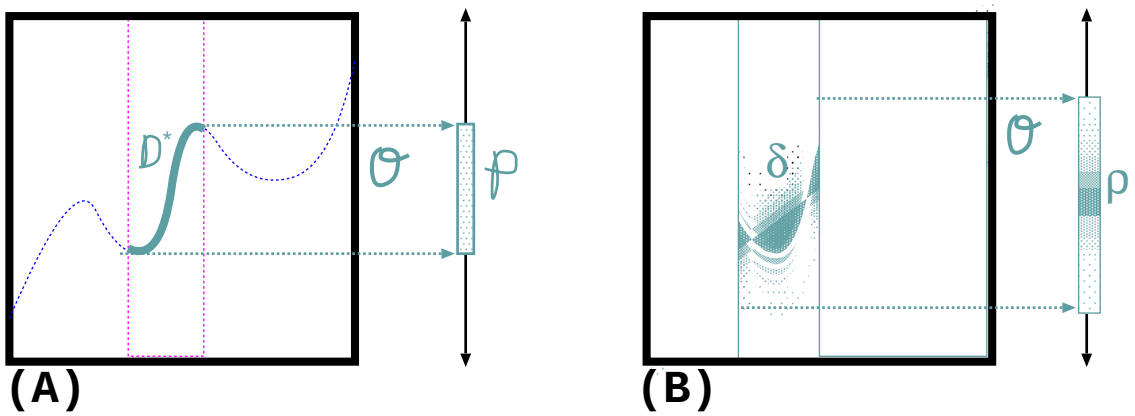


Figure 7.7: **(A)** Apply the observable $\mathcal{O}$ to $\mathcal{D}^*$ to get a range $\mathcal{P}$ of predicted measurement outcomes. **(B)** Apply the observable $\mathcal{O}$ to δ to predict a probability distribution ρ over the possible measurement outcomes.

Now, suppose we wanted to *predict* the value of some other observable, $\mathcal{O}$.

- In a *deterministic* theory, let $\mathcal{P} = \mathcal{O}(\mathcal{D}) \subset \mathbb{R}$; this is the range of *predicted measurement outcomes* (see Figure 7.7A). That is, $\mathcal{P}$ is the set of measurement outcomes for $\mathcal{O}$ which are compatible with the initial data. If $\mathcal{P}$ is a very small set, this means the data (d_1, d_2, d_3) yields a good prediction of $\mathcal{O}$. If $\mathcal{P}$ is large, then (d_1, d_2, d_3) does little to predict $\mathcal{O}$.

- In a *probabilistic* theory, let $\rho = \mathcal{O}(\delta)$. Then ρ is a probability distribution on $\mathbb{R}$, assigning a probability to every possible measurement outcomes (see Figure 7.7B). If ρ is highly concentrated in a small interval, this means that (d_1, d_2, d_3) yields a good prediction of $\mathcal{O}$.

(v) Description, Explanation, and Prediction

The three most common forms of scientific inference are:

Description: We take a collection of known facts and plug them into the theory. Manipulation of constraints allows us to infer these facts from one another, thus showing that they are mutually consistent. The chains of scientific inferences connecting the facts to each other can be translated into a story —a *description* of how these facts are related.

Explanation: We begin with a 'mysterious' or puzzling phenomenon (the *explanandum*), and a set of 'mundane' known facts (ie. initial data). By plugging the initial data into

the theory and manipulating constraints, we are able to infer the explanandum. Thus, we *explain* it by showing how it is entailed by the known facts.

Prediction: We plug a collection of known facts into the theory. We manipulate constraints, and make an inferences about as-yet unobserved phenomena or as-yet untried experiments. This yields experimentally testable *predictions*.

Notice that, at the formal level of the theory, *all three forms of inference are identical*:

- *Description* is inference between unsurprising and well-known facts.
- *Explanation* is inference from known 'mundane' facts to a known but 'mysterious' fact.
- *Prediction* is inference from known facts to yield a currently unconfirmed assertion.

To corroborate a theory, it is generally felt explaining a mystery is more impressive than merely redescribing known facts. And making a bold *prediction* (if empirically confirmed) is much more convincing than merely providing an *explanation* (even an explanation of a mystery). What justifies this ranking?

WHY EXPLANATION IS BETTER THAN DESCRIPTION: Explanation and description are formally identical exercises; the only difference is our subjective feeling that the explanandum is 'mysterious'. Why should a this make for a more impressive verification? The reason is that a fact is 'nonmysterious' if there are *already* plausible explanations for it (in terms of other theories). A fact is 'mysterious' if it so far has no good explanation. Thus, in describing known facts, a new theory does not distinguish itself from its rivals. In explaining a mystery, it demonstrates a clear superiority.

WHY PREDICTION IS BETTER THAN EXPLANATION: Once it is empirically confirmed, a *prediction* is indistinguishable an *explanation*. The theory agrees with the data in both cases. The only difference is in the order of events: in an 'explanation', the data arrived *before* the theory, in a 'prediction', the data arrived *after*. This is just an accident of history; why should it affect the extent to which the theory is 'validated' by the data?

If science were *exact*, then explanation and prediction really *would* be identical exercises, with a mere accident of history distinguishing them. But science is never exact. As we saw in §(i), approximations, 'guestimates', Taylor series truncations, 'scaling arguments', and other white lies are ubiquitous in scientific inference. These are tolerated out of practical necessity, and are vindicated when they correctly predict *unknown* phenomena. Explaining *known* phenomena using such inexact inference is less impressive; there is a suspicion that, with judicious approximations and truncations, you can 'rig' the calculations to 'explain' pretty much anything you want. Predicting the *unknown* provides an airtight alibi; you can't 'rig' the calculations to get the right answer, because you don't know what the right answer *is*, until after the prediction is tested.

(vi) Empirical theory validation

We trust scientific theories which have a long history of good agreement with experiment. What justifies this? When the empirical data agrees with a theory, in exactly what sense has the theory been 'validated'?

Suppose we are studying some system $\mathcal{S}$, which we will imagine is some stochastic process (special case: dynamical system). Our job as scientists is to develop a good *model* of this system. This model will be another stochastic process, which 'mimics' the first. A stochastic process is defined by assigning probabilities to trajectories through statespace, so our job is to *estimate the trajectory probabilities of* $\mathcal{S}$.

After observing the behaviour of $\mathcal{S}$ for a long time, we begin to notice certain correlations in its behaviour; a certain state at time 1 is strongly correlated with another state at time 2, certain trajectories are highly improbable, and so on. In other words, we accumulate a body of *empirical probability estimates*, based upon our *sample* of the behaviour of $\mathcal{S}$. If we are clever, we can subsume these estimates within some plausible model, $\mathcal{T}$.

We then proceed to 'test' $\mathcal{T}$, by taking further observations ('samples') of $\mathcal{S}$, and checking the empirically measured probability estimates against the theoretically predicted ones. They seem to agree. The question now becomes: how much trust should we place in our empirical probability estimates?

The probability theorists have developed 'sampling theorems' which answer this question. The simplest is the **Law of Large Numbers**, which basically says: if you flip a coin ten thousand times, and it comes up heads 5001 times, then it is extremely likely to be a fair coin. But if it comes up heads 8000 times, then it is extremely likely that it is an unfair coin, with a 30% bias in favour of heads.

A far-reaching generalization is the **Ergodic Theorem**, which roughly says:

> *Let $\mathcal{S}$ be an ergodic*[8] *stochastic process, and let $\mathbf{s} = (s_1, s_2, s_3, \ldots, s_n)$ be part of a trajectory of $\mathcal{S}$. Let $\mathbf{E}$ be some event, and empirically estimate the probability of $\mathbf{E}$ by c/n, where c is the number of times the trajectory $\mathbf{s}$ passes through $\mathbf{E}$.*
>
> *If n is large enough, then (with extremely high probability) c/n will be close to the true probability of $\mathbf{E}$.*

Here n is some large number, say, one million. *'Extremely high probability'* means 'as close to probability one as you want', and '*close to the true probability*' means 'with an error as close to zero as you want', where, in both cases, you may have to make n very large to get what you 'want'.

So, if you wait long enough, it is *extremely unlikely* that your empirically measured probability for $\mathbf{E}$ will be in error. Keep in mind that $\mathbf{E}$ can be any 'event', including, for example, events of the form, 'The system was in state $\mathbf{A}$ at time 0 and in state $\mathbf{B}$ at time t'.

[8] This is a technical assumption.

Hence, we can use this to estimate *correlations in time*, and indeed, any statistical property we want. Hence, we can estimate the 'true' trajectory probabilities of $\mathcal{S}$ to within any desired accuracy.

The Ergodic Theorem says that a sufficiently large body of empirical verification for your model virtually assures its correctness. There are two drawbacks:

- The Ergodic Theorem does not specify just *how* large n must be. A 'sufficiently large' body of verification may take a million years to acquire.

- The Ergodic Theorem says nothing about generalizing from our model of this one system to get models of other similar systems. In particular, it does not tell us how to extrapolate form a specific *model* to a general *theory*.

(vii) Occam's Razor; or, What is a good theory?

Suppose two scientific theories are equally compatible with the available empirical data. What makes one theory better than the other?

Robustness: Measurements are never precise, and calculations inevitably require approximations. In a good theory, small errors shouldn't rapidly magnify into large ones.

Computational Simplicity: In simpler models, solutions are easier to compute, and less likely to be wrong.

Amenability to Analogy: Reasoning by analogy is ubiquitous in science, and a good theory facilitates this, by allowing us to take an unfamiliar system and understand it by arguing that, in the important respects, it is 'isomorphic' with a familiar one.

These three considerations entail the following desiderata:

Minimal number of relevant variables: This clearly facilitates *Computational Simplicity*, because less variables means less computation. It also increases *Robustness*, because there are less opportunities for measurement error to corrupt the results.

Most importantly, it makes the theory *Amenable to Analogy.* The fewer the relevant variables, the more likely it is that two very different systems will be isomorphic. For example, Newtonian gravitation is powerful because, to model the behaviour of an object in a gravitational field, the *only* relevant variable is its mass. Thus, a stone and a potato of identical mass will exhibit identical behaviour, so we can reason by analogy from one to the other.

Decorrelation with distance: It is useful to know an event occurring far away or long ago will not influence the phenomenon you are studying. This radically reduces the number of relevant variables, because it means we can ignore most of the universe. Thus, for example, gravitational and electric fields obey *inverse square laws*, which means the influence of far away masses or charges is negligible.

Invariance and covariance: A theory is *invariant* if changing certain variables doesn't affect the application of the theory. This helps minimize the number of relevant variables. For example:

- An essential property of all scientific theories is *translation invariance*, which just means, 'The same scientific laws hold everywhere in space and time.'
- Classical mechanics also features *Galilean invariance*, which says, 'The laws of classical mechanics are the same, regardless of your velocity.'
- Special relativity in addition assumes *Lorenz invariance*, which says, 'The laws of electrodynamics (and in particular, the speed of light) are the same regardless of your velocity.'

A theory is *covariant* if changing a certain variable changes the application of the theory in a simple and predictable way. This makes the theory *Amenable to Analogy*. A good example is *scaling*. It is possible to apply the same equations to a rock falling to earth and an asteroid hurtling through space, despite the fact that they differ wildly in mass, velocity, and ambient gravitational field. The reason is because the equations of classical mechanics *scale* in very simple ways with changing mass, velocity, etc.

Differentiability and Continuity A good theory lends itself to approximation. This is clearly necessary for *Robustness*, and often crucial for *Computational Simplicity*, since it lets us get away with minor mathematical sleight-of-hand. It is also enhances *Amenability to Analogy*, because we can construct a useful analogy between two systems even when they aren't *exactly* the same. For example, a stone falling to Earth is *not* the same as an asteroid travelling through space, because the stone is subject to air resistance. However, this is a small error, which we can neglect for most purposes.

To approximate, we need functions which are *continuous* and hopefully *differentiable*. Continuity means that a small input error produces a small output error. Differentiability —a stronger condition —allows us to *extrapolate* observed trends.

For example, suppose a rocket is at position x, and is travelling at velocity v. It is firing its engines, so it has acceleration a. However, it is cutting back the thrust to these engines at rate r. In this case, we can extrapolate the position of the rocket at some time t in the near future as

$$x \ + \ v \cdot t \ + \ \frac{a}{2}t^2 \ - \ \frac{r}{6}t^3 \ + \ \epsilon$$

where ϵ is an error which we can guarantee is small, as long as t is small. This *Taylor polynomial*[9] extrapolation implicitly exploits differentiability.

In summary, a scientist building a theory has strong incentive to minimize the number of relevant variables and long-distance correlations, while maximizing the invariance, covariance, continuity, and differentiability of the theory. If we identify these properties with 'simplicity', then we can say that she has good reason to seek the simplest theory possible. This is perhaps a good formulation of Occam's Razor.

8 The Games People Play

In this chapter, I'll develop a model of social, political, and economic interactions as a 'game' between two or more players, and use this framework to investigate the nature of power, freedom, and political stability. The most important 'moves' in a social game are often 'communicative' —ie. attempts to manipulate other people's beliefs. Thus, the semiotics of communication is deeply implicated in game strategy, as is demonstrated by the role of advertising in the 'game' of economics, and the role of diplomatic posturing in the 'game' of international relations.

(i) Social Games

Let $\mathcal{W}$ be the set of all possible worldstates. Thus, the space of all *possible futures* is the space of all infinite sequences the form $\mathbf{w} = (w_1, w_2, w_3, \ldots)$, where $w_1, w_2 \ldots$ are all elements of $\mathcal{W}$. The sequence $\mathbf{x}$ describes a future where the world is in state w_1 tomorrow, state w_2 the day after, and so on. The space of all such $\mathbf{w}$ is denoted $\mathcal{W}^{\mathbb{N}}$.

UTILITY: Let's consider a particular player, called Xander. Xander has a **utility function**, which assigns to every possible future in $\mathcal{W}^{\mathbb{N}}$ a numerical value, indicating its relative level of "happiness" for him. In other words, we have a function:

$$U : \mathcal{W}^{\mathbb{N}} \longrightarrow \mathbb{R}$$

Thus, if $\mathbf{w}, \mathbf{v} \in \mathcal{W}^{\mathbb{N}}$ and $U(\mathbf{w}) > U(\mathbf{v})$, this means that Xander *prefers* the future $\mathbf{w}$ to the future $\mathbf{v}$ (since it would make him happier). Thus, U encodes all of Xander's desires and values, including the relative importance of short-term vs. long-term happiness for him

[9] See footnote 1 on page 57 of §(i).

(what economists call his *discount rate*). Different people have different utility functions, since they have different values and desires.

We will assume that Xander always attempts to maximise U. This does *not* mean that Xander is 'selfish'. For example, if Xander loves Ysolde, then her future happiness will be one of the variables determining the value of U. If Ysolde is happy in the future $\mathbf{w}$, but sad in the future $\mathbf{v}$, then $U(\mathbf{w})$ will be much larger than $U(\mathbf{v})$. Xander will try hard to make Ysolde happy.

We will also assume that utility is **cardinal**. This means that, if $U(\mathbf{w}) = 2 \cdot U(\mathbf{v})$, then $\mathbf{w}$ has 'twice' the utility of $\mathbf{v}$. This may seem meaningless: after all, what exactly does 'twice as happy' mean? To understand the meaning of cardinal utility, imagine a gambling scenario. You can choose one of two lottery tickets, **(A)** and **(B)**:

(A) offers you an all-expenses-paid, 14-day trip to Argentina, with a 1% chance of winning.

(B) offers an all-expenses-paid, 14-day trip to Brazil.

Which ticket do you want? If the ticket **(B)** offers a 99% chance of winning, you would probably take **(B)**. But if **(B)** offered a 0.0001% chance of winning, you'd take ticket **(A)**. Somewhere in the middle is a probability where the two tickets are equally desirable to you. Let's suppose that, when ticket **(B)** is at 3%, you'd choose it, but if it was 2.9%, you'd go for **(A)**. This means that a 1% chance of the Argentina trip is *worth about the same to you* as a 3% chance of the Brazil trip. Hence, the Argentina trip has *three times* as much value to you.

I am assuming that your decisions are 'rational'. For example, if a 3% Brazil ticket equals a 1% Argentina ticket, then a 6% Argentina ticket should equal a 2% Argentina ticket. Likewise, if a 1% Brazil ticket equals a 5% chance of a trip to Chile, then, 'rationally', a 1% *Argentina* ticket should equal a 15% Chile ticket. If your wagers are rational in this fashion, it follows that you are valuing the trips with a cardinal utility function (albeit unconsciously).

The relevance is this: in real social games, outcomes are never certain; *all* decisions are wagers. A 'perfectly rational' Xander will makes his decisions by trading off the utilities of the various outcomes against their relative probabilities. Of course, none of us are perfectly rational in this sense, but I'll use this approximation here in order to build a mathematical model. In the 'ticket' example, the probabilities were given to you; in real life, you have to estimate them. You estimate probabilities by means of a vast (and mostly unconscious) body of knowledge and intuition about which events are likely and which are not and how they are correlated —in short, by means of your 'worldview'.

WORLDVIEWS: Xander also has a **worldview**, ξ, which reflects his beliefs about what sorts of events are likely or unlikely, what sorts of correlations one can expect between events, how other people are likely to behave, and so on. In the language of Chapter 7§(ii), ξ is a sort of personal scientific *model* of the universe. We will treat this model as a stochastic process

(page 63). In other words, ξ assigns a probability to every *possible history* —every trajectory of the universe through its statespace $\mathcal{W}$. Such a trajectory has a *future* (a sequence $(w_1, w_2, w_3, \ldots)$ in $\mathcal{W}^{\mathbb{N}}$), a *present* (a single element $w_0 \in \mathcal{W}$), and also a *past* (a sequence $(\ldots, w_{-3}, w_{-2}, w_{-1})$; we will assume the past is infinite, for simplicity). Together, the past, present, and future form a bi-infinite sequence $(\ldots, w_{-3}, w_{-2}, w_{-1}, w_0, w_1, w_2, w_3, \ldots)$. The space of such sequences is called $\mathcal{W}^{\mathbb{Z}}$, and ξ is a probability distribution on $\mathcal{W}^{\mathbb{Z}}$, which assigns a probability to every possible sequence of events. Thus, ξ implicitly encodes any logical relationships or statistical regularities in which Xander believes (often unconsciously).

Xander *might* be deluded. The *real* logical relationships and statistical regularities of the universe (ie. the 'Laws of Nature') are given by *another* probability measure, ρ. Xander clearly believes and hopes that ρ is close to ξ; his 'delusions' are just the disparities between ρ and ξ.

Obviously, real people do not make decisions by consciously assigning probabilities to various scenarios ("There is a 0.734 probability that buying this car will make me 20% happier"). Even when we *do* estimate likelihoods, we don't employ some conscious, logically consistent theory, but instead resort to intuition (or prejudice). Thus, I'm not saying that ξ is a consciously articulated, 'scientific' model in Xander's mind. Instead, I'm saying we can model Xander's decisions *as if* they were the result of some worldview ξ, which arises from largely unconscious, often 'irrational' mental processes. Thus, ξ implicitly encodes the full human complexity of Xander's psychology.

INFORMATION: Xander has some information about the present (his sensations) and some information about the past (his memories). This information determines a subset $\mathcal{I} \subset \mathcal{W}^{\mathbb{Z}}$; the set of all sequences $(\ldots, w_{-2}, w_{-1}, w_0, w_1, w_2, \ldots)$ consistent with Xander's memories and sensations. Note: $\mathcal{I}$ is not a 'set of facts' —rather, it is the 'set of possible worlds' that are *consistent* with Xander's facts. Thus, the *smaller* $\mathcal{I}$ is (ie. the more restricted the set of possible worlds), the *more* Xander (thinks he) knows about the state of the word.

From $\mathcal{I}$ and ξ, Xander can tentatively *predict* the future. He estimates the probability of future events by applying the distribution ξ, *conditioned* on $\mathcal{I}$. For example, if $\mathcal{E} \subset \mathcal{W}^{\mathbb{Z}}$ is some event, then Xander's estimates the likelihood of $\mathcal{E}$ as the conditional probability $\xi[\mathcal{E} \;\langle\!\langle\; \mathcal{I}]$. If $\xi[\mathcal{E} \;\langle\!\langle\; \mathcal{I}] = 0$, then Xander 'deduces' from the information $\mathcal{I}$, that $\mathcal{E}$ is impossible. If $\xi[\mathcal{E} \;\langle\!\langle\; \mathcal{I}] = 1$, then he 'deduces' from $\mathcal{I}$ that $\mathcal{E}$ is a certainty.

An important prediction for Xander is his **expected utility level**:

$$\mathbb{E}_\xi[U \;\langle\!\langle\; \mathcal{I}]$$

This is the *average utility* Xander can expect to obtain in all possible futures consistent with $\mathcal{I}$, with probabilities given by ξ. For example, suppose there were only two possible futures

in $\mathcal{W}^{\mathbb{N}}$ that were consistent with $\mathcal{I}$ —call them $\mathbf{v}$ and $\mathbf{w}$. Suppose that:

$$\begin{array}{rclcrcl} \xi[\mathbf{v} \ \langle\!\langle \ \mathcal{I}] & = & 0.7; & & U(\mathbf{v}) & = & 5 \\ \xi[\mathbf{w} \ \langle\!\langle \ \mathcal{I}] & = & 0.3; & \text{and} & U(\mathbf{w}) & = & 10. \end{array}$$

Then Xander's expected utility is:

$$\begin{array}{rcl} \mathbb{E}_\xi\,[U \ \langle\!\langle \ \mathcal{I}] & = & \xi[\mathbf{v} \ \langle\!\langle \ \mathcal{I}] \cdot U(\mathbf{v}) \ + \ \xi[\mathbf{w} \ \langle\!\langle \ \mathcal{I}] \cdot U(\mathbf{w}) \quad = \quad (0.7) \cdot 5 \ + \ (0.3) \cdot 10 \\ & = & 3.5 \ + \ 3.0 \quad = \quad 6.5. \end{array}$$

More generally, suppose there were N possible futures, say $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$, having conditional probabilities

$$\xi[\mathbf{w}_1 \ \langle\!\langle \ \mathcal{I}] \ = \ p_1; \qquad \xi[\mathbf{w}_2 \ \langle\!\langle \ \mathcal{I}] \ = \ p_2; \qquad \ldots \qquad \xi[\mathbf{w}_N \ \langle\!\langle \ \mathcal{I}] \ = \ p_2.$$

and utilities

$$U(\mathbf{w}_1) \ = \ u_1; \qquad U(\mathbf{w}_2) \ = \ u_2; \qquad \ldots \qquad U(\mathbf{w}_N) \ = \ u_N$$

Then Xander's expected utility is:

$$\mathbb{E}_\xi\,[U \ \langle\!\langle \ \mathcal{I}] \quad = \quad p_1 \cdot u_1 \ + \ p_2 \cdot u_2 \ + \ \ldots \ + \ p_N \cdot u_N.$$

(A similar formula holds for an infinity of possible futures, but involves some technicalities which we will forego).

Xander may also *speculate* on possible scenarios; for example, he may hypothesise some state of current/future affairs, represented by a subset $\mathcal{H} \in \mathcal{W}^{\mathbb{Z}}$; and then consider the expected utility

$$\mathbb{E}_\mu\,[U \ \langle\!\langle \ \mathcal{I} \cap \mathcal{H}]$$

This is Xander's expected utility consistent with the information $\mathcal{I}$ and the hypothesis $\mathcal{H}$. If $\mathbb{E}_\mu\,[U \ \langle\!\langle \ \mathcal{I} \cap \mathcal{H}] \ > \ \mathbb{E}_\mu\,[U \ \langle\!\langle \ \mathcal{I}]$, this means that the hypothetical state of affairs represented by $\mathcal{H}$ is desirable to Xander. Conversely, if $\mathbb{E}_\mu\,[U \ \langle\!\langle \ \mathcal{I} \cap \mathcal{H}] \ < \ \mathbb{E}_\mu\,[U \ \langle\!\langle \ \mathcal{I}]$, this means $\mathcal{H}$ is undesirable to him.

ACTION: Xander also has a repertoire of **actions**. He can perform one action during each moment of the game, and it will modify the current game state. To be precise, there is a *partition* of $\mathcal{W}$ into a collection of subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_N$ such that:

(i) These sets are *disjoint*: for any i and j which are not equal, $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$.

(ii) These sets together *cover* $\mathcal{W}$; that is, $\mathcal{W} = \bigcup_{n=1}^{N} \mathcal{X}_n$.

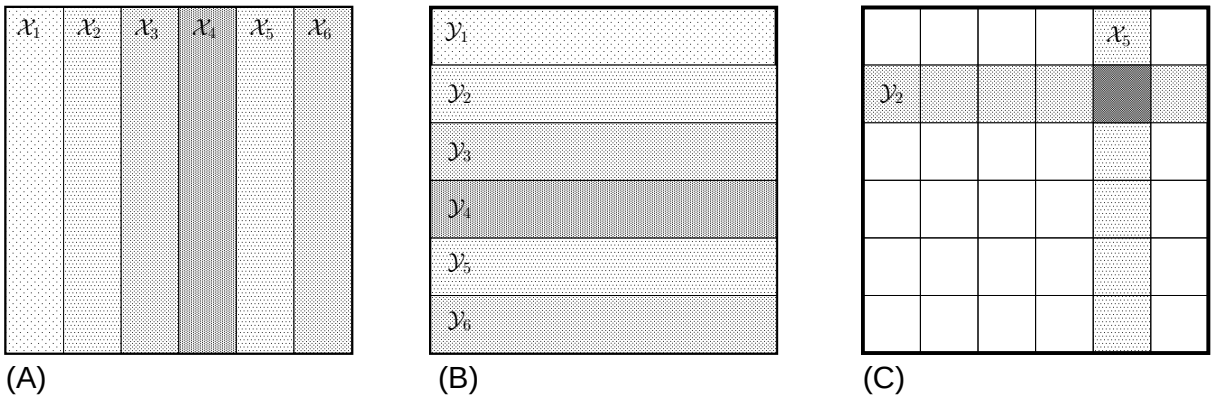


Figure 8.1: **(A)** Xander's actions partition $\mathcal{W}$ into $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_6$. **(B)** Ysolde's actions partition $\mathcal{W}$ into $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_6$. **(C)** If Xander chooses $\mathcal{X}_5$, and Ysolde chooses $\mathcal{Y}_2$, then $w_0 \in \mathcal{X}_5 \cap \mathcal{Y}_2$.

In other words, the family $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_N$ represents a collection of alternatives which are **(i)** *mutually exclusive* and **(ii)** *exhaustive* (see Figure 8.1A).

Xander's action during any round is to choose one of the sets $\mathcal{X}_1, \ldots, \mathcal{X}_N$; this forces the current world-state to be a member of that subset.

The actions of each player partition the space in a different way. For example, suppose that Ysolde's actions determine a partition $\mathcal{Y}_1, \ldots, \mathcal{Y}_M$ (Figure 8.1B). If Xander chooses action $\mathcal{X}_5$, and Ysolde chooses $\mathcal{Y}_2$, then we know that $w_0 \in \mathcal{X}_5 \cap \mathcal{Y}_2$ (see Figure 8.1C). If, in addition, Zarathustra has partition $\mathcal{Z}_1, \ldots, \mathcal{Z}_L$, and Zarathustra chooses $\mathcal{Z}_7$, then $w_0 \in \mathcal{X}_5 \cap \mathcal{Y}_2 \cap \mathcal{Z}_7$. The actions of other players will further constrain w_0.

Some remarks:

- I have assumed that Xander has a finite repertoire of actions (ie. that the partition $\mathcal{X}_1, \ldots, \mathcal{X}_N$ is finite) for simplicity of illustration. In principle, Xander's partition could be infinite or even uncountable[1]. Even if finite, we can assume the repertoire is extremely large.

- The intersection $\mathcal{X}_5 \cap \mathcal{Y}_2$ is always nonempty. If it were empty, this would mean that Xander action $\mathcal{X}_5$ somehow 'preempts' Ysolde's ability to choose $\mathcal{Y}_2$, which makes no sense, since they act simultaneously. Similarly $\mathcal{X}_5 \cap \mathcal{Y}_2 \cap \mathcal{Z}_7$ is nonempty, etc.

For example, in a model of economic behaviour, Xander's repertoire might include the following actions:

- Production (eg. making a shoe).

[1]In this case, rather than a partition, we would use a *sigma-algebra* —see Appendix D

- Research (eg. looking for a good strudel).
- Communication (eg. advertising the shoe to Ysolde)
- Exchange (eg. exchanging shoe for money, exchanging money for strudel)
- Consumption (eg. eating the strudel)

We assume that Xander is **rational**, meaning that he will always choose the sequence of future actions (ie. the **strategy**) which yield the *greatest expected utility* according to his worldview ξ, utility function U, and current information $\mathcal{I}$.

COMMUNICATION: **Communication** is any action whose purpose is to *change the information state of other players*. For example, if Ysolde chooses action $\mathcal{Y}_1$, then Xander's information-state goes from $\mathcal{I}$ to $\mathcal{I} \cap \mathcal{Y}_1$. He has just learned something new (namely, that Ysolde chose $\mathcal{Y}_1$).

The *significance* of action $\mathcal{Y}_1$ to Xander is how it affects his conditional probability estimates. For example, if $\mathcal{E}$ is some event, and $\xi\left[\mathcal{E} \;\langle\!\langle\; \mathcal{Y}_1 \cap \mathcal{I}\right]$ is much smaller than $\xi\left[\mathcal{E} \;\langle\!\langle\; \mathcal{I}\right]$, then, through action $\mathcal{Y}_1$, Ysolde has 'informed' Xander that $\mathcal{E}$ is much less likely than he previously thought it was.

Most of Ysolde's actions carry little 'information', and will not strongly affect Xander's conditional probability estimates. Those that do can be divided into several (vague and overlapping) classes:

- *Explicit communication:* Ysolde speaks directly to Xander, and the meaning she *intends*, the *literal* meaning of her words, and the meaning Xander *perceives* all coincide. This is the simplest form of communication, and the most often studied in the philosophy of language.
- *Deception:* Ysolde speaks falsely, but *intends* that Xander should believe her (ie. that his perceived meaning should be the literal meaning of her words.)
- *Miscommunication*: Xander's perceived meaning differs from Ysolde's intended meaning.
- *Coded communication:* Ysolde speaks to Xander, and her intended meaning coincides with his perceived meaning, but differs from the literal meaning.

  For example, the spies in cloak-and-dagger stories often employ various code phrases ("The dog barks at midnight." = "We act tonight.", etc.)
- *Metacommunication:* Ysolde speaks to Xander, and her words carry an *explicit* meaning, but also imply an additional 'coded' meaning.

For example, when a reference letter excessively praises a job candidate's punctuality and sartorial sensibilities, but conspicuously fails to discuss his intellectual abilities or work habits, it *metacommunicates* a poor opinion him, while explicitly seeming to praise him.

- *Nonverbal communication:* Ysolde performs an action involving no overt verbal communication, but fraught with communicative significance nonetheless.

  In social settings, the crucial example is *eye contact*. By meeting someone's eyes, you implicitly communicate that you are aware of them, and aware that they are aware of you, and aware that they are aware that you are aware of them, etc. Eye contact opens a 'channel' over which to transmit other nonverbal signals. It tells the recipient, 'I know you are watching, and thus, my gestures are *not* random —they are are intended to be seen and interpreted.'

  For example, at a cocktail party, Xander catches Ysolde's glance. If either looks away immediately, this says the eye contact was accidental, with no communicative intent. But each holds eye contact, and a channel is established. With an inquiring arch of his eyebrows, Xander (nonverbally) communicates a romantic interest. Ysolde's expression (nonverbally) acknowledges her awareness of his interest. Without breaking eye contact, she raises her glass and drinks, clearly displaying the wedding ring on her hand, and (nonverbally) communicating her disinterest in him.

- *Inadvertent Revelation:* Ysolde's actions unintentionally reveal information to Xander (for example, she blushes or stammers when certain topics arise in conversation).

- *Dissimulation:* This is like *Deception*, but rather than speaking a sentence whose literal meaning is false, Ysolde instead acts in a way which she *intends* Xander to perceive as an *Inadvertent Revelation*. A lot of emotionally manipulative behaviour in dysfunctional human relationships is of this kind.

Social, political, and diplomatic communication is complicated because the participants (legitimately) assume that almost all explicit communication is deceptive, and that the *real* communication is coded or nonverbal. For example, consider the following description of the 1911 'Moroccan crisis':

> *When a rebellion against the sultan... resulted in a French expedition to occupy Fez,... the German government protested. ...A German warship was sent to the Moroccan port of Agadir, ostensibly to protect German nationals there (there were none). The real aim was to show that Germany meant business and to frighten France into agreeing to compensation for her.*
>
> *...A government in Paris which had recently been disposed to be conciliatory now found it very difficult... to make concessions to Germany....; it could not allow itself to appear to be weak in defending French interests.*

> *....It was felt by London that a gesture was needed to show British concern. A speech was made by a minister, which, whatever it was intended to mean, was taken as a warning that if France found herself at war with Germany, Great Britain would support her.* [33, pp.201-202]

Here, the Germans interpret the French occupation of Fez as *inadvertently revealing* French intentions to encroach upon Morocco. The German protestations *explicitly communicate* that they will not tolerate such encroachment. To make sure that their protest is not misinterpreted as a bluff (ie. *deception*), the Germans send a warship is to *nonverbally communicate* that they are serious. Ironically, the German actions actually increase French intransigence; the French feel they cannot 'climb down', lest this be interpreted as *inadvertently revealing* vulnerability. The British minister's speech is then interpreted by Germany as *metacommunicating* British support for the French.

In social, political, and diplomatic 'games', we must treat the players' actions as *signs*. The interpretation of these signs is *semiotics*. Thus, semiotics is deeply implicated in political and social issues. I'll return to this in §(iii).

(ii) Power and Reciprocity

Power is the central concept of political philosophy. But what exactly *is* Power? Where does it come from? How is it acquired, and how is it maintained? 'Power comes from the barrel of a gun,' said Mao. But this can't be its only source. Is the *economic* power of the wealthy really just a disguised threat of violence? If Power always comes from the threat of violence, then why don't all governments devolve into military dictatorships?

I argue that power is a consequence of *reciprocity* in player interactions. For example, Ysolde interacts with Xander mainly through *bargaining*, which means influencing his behaviour by manipulating his incentives. Xander's *incentives* are determined by his expected future utility. In this analysis, Ysolde's *power* is her (perceived) ability to affect Xander's future utility.

INCENTIVE: Xander has **incentive** to choose action $\mathcal{X}_1$ when he decides that **expected utility**, if he chooses $\mathcal{X}_1$, is greater than his expected utility for any other action. Formally:

$$\mathbb{E}_\xi\left[U \ \langle\!\langle \ \mathcal{X}_1 \cap \mathcal{I}\right] \quad \geq \quad \mathbb{E}_\xi\left[U \ \langle\!\langle \ \mathcal{X}_k \cap \mathcal{I}\right], \qquad \text{for all } k \neq 1 \tag{8.1}$$

We assume that Xander is a 'rational maximiser', and chooses the action which maximises his expected utility. Note that this is a very weak notion of 'rationality': the inequality in (8.1) depends on Xander's worldview ξ and on his information $\mathcal{I}$. If he has a highly demented worldview, or is wildly misinformed, then his choice may be neither rational nor maximal, according to our estimation.

The value $U_{\max} = \mathbb{E}_\xi [U \ \langle\!\langle \ \mathcal{X}_1 \cap \mathcal{I}]$ is Xander's *maximal expected utility under any course of action*; we will call this his **prospect**. Note that Xander's prospect is determined by his current information $\mathcal{I}$. In other words:

$$U_{\max}[\mathcal{I}] \quad = \quad \max_k \ \mathbb{E}_\xi [U \ \langle\!\langle \ \mathcal{X}_k \cap \mathcal{I}]$$

Power: Ysolde may be able to affect Xander's prospect. If Ysolde takes action $\mathcal{Y}_+$, perhaps she can improve Xander's prospect:

$$U_{\max}[\mathcal{I} \cap \mathcal{Y}_+] \quad > \quad U_{\max}[\mathcal{I}]$$

This is **benevolent power**. Conversely, if she takes action $\mathcal{Y}_-$, perhaps she can diminish Xander's prospects:

$$U_{\max}[\mathcal{I} \cap \mathcal{Y}_-] \quad < \quad U_{\max}[\mathcal{I}]$$

This is **malevolent power**.

Bargaining and Influence: By offering benevolent power, and/or threatening with malevolent power, Ysolde can **influence** Xander, by creating incentive for him to act in certain ways. For example, suppose she wants him to choose action $\mathcal{X}_*$. If $\mathbb{E}_\xi [U \ \langle\!\langle \ \mathcal{X}_* \cap \mathcal{I}] = U_{\max}[\mathcal{I}]$, then $\mathcal{X}_*$ is already an optimal choice for Xander, so she need not intervene. However, suppose that

$$\mathbb{E}_\xi [U \ \langle\!\langle \ \mathcal{X}_* \cap \mathcal{Y}_+ \cap \mathcal{I}] \quad > \quad U_{\max}[\mathcal{I}] \tag{8.2}$$

Hence, Ysolde can *offer* action $\mathcal{Y}_+$ *in exchange* for Xander choosing $\mathcal{X}_*$, and it is clearly preferable for him to comply. In economic terminology, Ysolde's benevolent power is her *buying power*. The exercise of this power is either a *purchase* or a *bribe* (depending on context).

Conversely, suppose that

$$\mathbb{E}_\xi [U \ \langle\!\langle \ \mathcal{X}_* \cap \mathcal{I}] \quad > \quad \mathbb{E}_\xi [U \ \langle\!\langle \ \mathcal{Y}_- \cap \mathcal{I}] \tag{8.3}$$

Then Ysolde can *threaten* action $\mathcal{Y}_-$ *unless* Xander chooses $\mathcal{X}_*$; again, it is clearly preferable for him to comply. Malevolent power usually manifests as military strength; its exercise is usually called *extortion*.

(iii) The Semiotics of Action

Note that the inequalities (8.2) and (8.3) both involve Xander's worldview ξ and current information $\mathcal{I}$. Hence, Ysolde's influence over Xander is determined by his worldview and information —ie. his *beliefs*. For her bargaining to succeed, it is not important whether Ysolde can (or will) *actually* improve/diminish Xander's utility; what is important is that

he *believes* she can (and will). Ysolde's key strategy is to manipulate Xander's information state $\mathcal{I}$ (through 'communication') to create the *image* of power, and to maintain a credible reputation. To do this, she must have some model of Xander's mental processes —ie. some kind of psychology.

PUBLIC IMAGE It is more important to create a credible *illusion* of power than it is to actually *have* it. Let's look at some examples of how this is done.

Diplomacy and Disinformation: States dissimulate through military 'feints' and deceive with disinformation. Almost all diplomatic speech is coded or deceptive, and every policy decision has implicit (nonverbal) communicative intent. Information is crucial; hence the importance of espionage and counterespionage.

The 'Morocco crisis' example (page 83) already showed the role of semiotics in diplomatic communication. Another episode from the same era illustrates the crucial importance of credibility in diplomacy.

> *...there at last appeared to have been a slight improvement in Anglo-German relations as the two powers worked together in the London Conference....* [and] *negotiated secretly over a proposed railway from Berlin to Baghdad and the possible fate of* [the Portuguese Empire]. *....Taking the responsiveness of the British on these matters to mean they lacked confidence,* [Germany] *hopefully speculated that Great Britain might not, after all, be serious about backing France, should Germany attack her.* [33, p.205]

Thus, British support for France (though genuine) lacked *credibility* from a German perspective. This (mistaken) assessment later made the Germans more willing to pursue the disastrous *Schlieffen Plan*[2], thereby igniting the First World War. Perhaps, had the British threat been more *credible*, Germany would have been less willing to support Austria-Hungary in the Balkan brinkmanship which lead to hostilities with Russia, and the war might have been forestalled.

Advertising: In a free markets, firms gain economic influence by manipulating consumers' beliefs about the benevolent power of their products or services. Everyone agrees that deliberately deceptive advertising should be illegal. However, even advertising which is not *literally* false can still mislead or manipulate.

The 'neoconservative' view is that, if advertising makes no false assertions, then it is at worst 'informationally neutral', and cannot be pernicious. However, this view reflects

[2]In the Schlieffen Plan, the Germans (having declared war on Russia) invaded Belgium in order to preemptively attack France (Russia's ally), so as to 'prevent' a two-front war. Britain declared war on Germany (ironically, in defense of Belgian, not French, territorial integrity) and thus, the war began.

When the advertisement shows	It wants to associate the product with
Happy children playing, mother relaxing	Good parenting
Yee-haw adventuring in the rugged outback	Freedom, individualism, self-reliance
Ostentatious luxury	Wealth, status, 'success'
Nubile, scantily clad beautiful people	Sex
Tranquil sunset on a beach	Contentment, personal fulfillment

Table 8.1: Advertising translation table

a facile, overly literal notion of communication. As we've seen, much communication (and in some contexts, *most* communication) is implicit, not explicit. Deception is a crude and primitive kind of dishonesty. *Dissimulation* is much more effective and has the advantage of plausible deniability (no one can *prove* you intended to dissimulate).

Advertisers rarely try to manipulate you through literally false assertions. Instead, effective advertising manipulates your information-state in subtle and unconscious ways, to change your assessment of conditional probabilities, and thus, your incentives and decisions. We often say that ads try to *associate* a product with some (often totally unrelated) emotion or goal (Table 8.1). What this means is that the advertiser is trying to boost your (unconscious) estimate of the *correlation* between (buying) the product and (obtaining) the goal.

Indeed, Klein claims that, in media-saturated, post-industrial societies, marketing is less about selling physical *products*, and more about selling intangible *brands*:

> *What was changing* [in the 1990s] *was the idea of what... was being sold. The old paradigm had it that all marketing was selling a product. In the new model, however, the product always takes a back seat to the real product, the brand....*
>
> ...[A] *new consensus was born: the products that will flourish in the future will be the ones presented not as "commodities" but as concepts: the brand as experience, as lifestyle.*
>
> *Ever since, a select group of corporations has been attempting to free itself from the corporeal world of commodities, manufacturing, and products.... Anyone can manufacture a product... Such menial tasks, therefore, can and should be farmed out to contractors and subcontractors ....(ideally in the Third World, where labour is dirt cheap, laws are lax and tax breaks come by the bushel). Headquarters, meanwhile, is free to focus on the real business at hand —creating a corporate mythology powerful enough to infuse meaning into these raw objects just by signing its name.* [22, pp.21-22]

This analysis is fascinating, because a brand is an entirely *cultural* construct. A brand

is nothing more than a widely shared belief that a certain name or logo is strongly correlated with certain positive values, emotions, or experience. The job of advertising is to disseminate and cultivate this belief —to create a 'corporate mythology' which becomes an enduring feature of the collective cultural landscape.

The power of the brand has nothing to do with the concrete qualities of a particular product. Indeed, advertisers prefer *not* to cultivate psychological associations between the brand and concrete qualities, but instead, to cultivate associations with abstract intangibles like 'love', 'freedom', or 'rebellion'. The more abstract and intangible these positive associations become, the less vulnerable they are to rational scrutiny and deconstruction.

For example, it is suboptimal to cultivate a belief that 'Coke tastes good' or 'Chryslers are reliable'. These are concrete, specific statements which can be empirically tested and disproved (either you like the taste of Coke or you don't; either your Chrysler breaks down or it doesn't). It is much more effective to cultivate an association that 'Coke is passion' or 'Chrysler is freedom'. These associations (I won't call them 'beliefs' since they make no literal sense) connect the brand with intangible qualities, and are unfalsifiable (because they are, in fact, nonsensical).

Nevertheless, they profoundly affect people's buying decisions. In 1989, global corporate spending on advertising exceeded $240 billion ([32, pp.171-172], cited in [23, p.155]); in the same year, United States companies alone spent $125 billion [22, p.11]. By 1997, the United States spent $185 billion/annum on advertising [22, p.11]. If real corporations even vaguely resemble the profit-maximizers of microeconomic textbook fiction, this indicates that advertising is overwhelmingly important to sales. In other words, the manipulation of people's beliefs and expectations (primarily at an irrational level) has enormous influence on their purchasing habits.

In the mathematical framework of classical economics, advertising has no effect on economic behaviour. The 'rational maximiser' of the classical framework is *a priori* unsusceptible to psychological manipulation, because she has no 'psychology' to manipulate. However, the 'Social Game' framework of this chapter provides a model of a 'rational maximiser' who *is* susceptible to advertising[3]. Advertising changes her *information state* $\mathcal{I}$, thereby altering her estimation of her *expected utility* under various strategies (according to her worldview ξ), and potentially changing her optimal strategy (so that, for example, it includes buying a Coke). For advertising to be effective, it must take advantage of preexisting irrationality within ξ. But ξ is a model of human psychology, which contains irrationality aplenty.

[3]The flipside is that classical economic models are simple enough to be explicitly formulated and even solved, whereas the Social Game model is far to complex for practical purposes.

When the propaganda shows	It wants to associate the party with
Conspicuously ethnic people	Racial tolerance and diversity
Grandparents playing with babies	Tradition, conservatism, 'family values'
Pastoral landscapes	Commitment to the environment
Lab-coated workers in a clean, high-tech workplace	Economic prosperity, 'good jobs at good wages'
Big guys working in heavy industry	'Labour' values: respect for the hardworking man
Soft-focus shots of babies and happy, pregnant women	Pro-life policies

Table 8.2: Political propaganda translation table

Propaganda: Politicians gain political influence by manipulating voters' beliefs about the benefits of a party or policy. As with corporate advertising, overt *deception* is rare, but *dissimulation* is ubiquitous. Political speech is often coded (eg. talk of 'immigration reform' is often a coded appeal to racist sentiments) or nonverbal (eg. projecting an 'image' of competence, integrity, and 'leadership'). Some examples of implicit communication appear in Table 8.2.

Poseurs: We all cultivate a personal image. Our clothing, possessions, and physical posture are all forms of nonverbal communication: actions which are intended to manipulate other's beliefs about us —ie. to manipulate their information state. Even *speech* often communicates implicitly as well as explicitly: if Felipe conspicuously displays his erudition about Derrida and Foucault at a party, we conclude he is educated (but we also suspect that he wants to *appear* educated).

It is a good strategy to nonverbally communicate certain aspects of one's 'identity'. For example, clothing communicates a lot about our class (labour, professional, academic, business, etc.), values (materialism vs. spiritualism, conservatism vs. liberalism, traditionalism vs. futurism, etc.) and personality (conformity vs. individualism, playfulness vs. sobriety, etc.). Our wardrobe thus attracts other people with similar values: potential collaborators, friends, or lovers. Other aspects of our identity we conceal: we don't to reveal anxiety or insecurity, or declare our ideology in a hostile context.

Hence, *image-management* is a basic (perhaps *the* basic) social behaviour. However, we reserve a special contempt for people who take it too far, whose speech and dress is *nothing but* image-management, because we resent their (too obvious) attempt to manipulate our beliefs

Reputation and 'Face' In many a situations, both players must simultaneously commit to a strategy, and neither can act in response to the other. Suppose we have a **Prisoner's Dilemma** situation, as illustrated by the payoff matrices in Table 8.3. Here, each player

Xander's Expected Utility		
	Ysolde Cooperates	Ysolde Defects
Xander Cooperates	+9	−1
Xander Defects	+10	0

Ysolde's Expected Utility		
	Ysolde Cooperates	Ysolde Defects
Xander Cooperates	+9	+10
Xander Defects	−1	0

Table 8.3: Expected utilities for Xander and Ysolde in the Prisoner's Dilemma

receives a small advantage for `Defecting`, regardless of the other player's behaviour. Thus, it is in the interest of each player to `Defect`, and hope that the other does not. Hence, 'rationally', both players `Defect`, even though *both* could do better if they `Cooperate`.

The mutual `Cooperate` solution can only be achieved if both players have *full information* about the other's actions. However, whenever there is ambiguity about the other player's actions, the result will be mutual `Defection`. This explains a lot of real life situations, such as the 'Tragedy of the Commons'[4]. For similar reasons, free markets need a judiciary to enforce contracts, since the parties may otherwise breach the contract in Prisoner's Dilemma type situations.

The Prisoner's Dilemma is somewhat obviated if we include the importance of *reputations.* A player's reputation determines his **credibility** in future bargaining sessions, and thus, determines his **influence** over other players. To be concrete, suppose Ysolde is extremely powerful, but also known to be completely untrustworthy. Remember that Ysolde's influence over Xander is her ability to create a difference in his **expected utility** $\mathbb{E}_\xi[U]$. But if Ysolde has no credibility with Xander then, according to ξ, there is no strong probabilistic correlation between his actions and her reactions. It seems equally likely to Xander that Ysolde will `Cooperate` or `Defect`, *regardless* of what he does. Thus, he has no real *incentive* to **cooperate** .

Thus, if Ysolde betrays a lot of people, and these betrayals become widely known, then her bargaining ability will be greatly diminished. Since all influence is exercised through bargaining, she will lose much of her influence.

Some real-life examples:

- In small communities, the 'Tragedy of the Commons' is less likely, because everyone is aware of your activities, and you face a strong risk of social sanction if your abuse the collective resource.

- International treaties have no judiciary to enforce them, but signatory nations have

[4]This is an economic scenario where people degrade a collectively owned resource because, even though it is in their *collective long-term* interest to preserve the resource, it is in their *individual short-term* interest to despoil it.

a vested interest in not capriciously abrogating treaty commitments: abrogation will cost them credibility in future negotiations.

- In a free market, firms have reputations, and the desire to maintain this reputation motivates them to behave responsibly, and provide consistent quality in their products.

- Axelrod [3] and Danielson [9] have run computer simulations where thousands of virtual agents compete in 'Iterated Prisoner's Dilemma' games. Agents repeatedly play one another, and the outcome of previous encounters can influence future strategy. For example, if agent **X** `defected` against agent **Y** in rounds 87 and 88, then this information may cause **Y** to `defect` against **X** in round 89.

  Unsurprisingly, totally unscrupulous agents (who always `Defect`) did well at the beginning, but badly in the long run, because, metaphorically speaking, they acquired a 'bad reputation', which hurt them in future encounters. On the other hand, more 'ethical' agents did better in the long run, where they could benefit by establishing 'good reputations' with one another.

In diplomatic or political situations, players are often forced to make clearly suboptimal decisions, because of the long-term consequences to their reputation. For example:

- In games of military 'brinkmanship', both sides find it difficult to 'climb down', even when it becomes clear that the costs of the imminent confrontation far exceed any possible gain. The reason is that a withdrawal will be seen as a sign of weakness, decreasing the credibility of future threats.

- Governments refuse to negotiate with terrorists, even when the terrorists' demands (say, amnesty for six political prisoners) are trivial compared to their threatened retaliation (executing one hundred hostages). The reason is that any negotiation will send the message that terrorism is an effective tool to extract concessions, thereby increasing the likelihood of future terror acts.

(iv) Psychology

To manipulate Xander, Ysolde must first have a *model* of his current information-state. Conversely, when Xander trusts Ysolde because of her 'reputation', he is implicitly using a *model* of Ysolde to forecast her future behaviour. When Xander forms a long-term strategy ('First I'll get an education, then I'll get a job, then I'll get a house', etc.), he is using a *self-model* to forecast his *own* behaviour. Thus, a complete account of social games must describe Xander's psychological *models* of himself and other players.

Xander has several ways to model Ysolde, none of which is completely satisfactory.

Behaviorism: Xander can model Ysolde as nothing but an ensemble of (probabilistic) responses to various inputs. In other words, his model consists of a *stochastic function* Φ_Y; given any input $\mathbf{i}$, he predicts Ysolde's response as a random action with probability distribution $\Phi_Y(\mathbf{i})$.

The problem is that this model doesn't allow Xander to explicitly represent *Ysolde's* model of *him*, which he needs in order to reason about issues of 'reputation' and 'face'. For Xander to reason, 'I should not `defect` against Ysolde today, or else she may `defect` against me tomorrow.', his model of Ysolde must be sophisticated enough to represent her knowledge of him and his past behaviour.

Of course, we could explicitly build this information into the function Φ_Y, for example:

$$\Phi_Y(\text{I defect today}) \quad = \quad 99\% \text{ chance that Ysolde defects tomorrow.}$$

But this rather artificial solution becomes unwieldy when we model the enormously complex reputational reasoning of real social or diplomatic situations.

Infinite Regress: The Social Game model describes Ysolde with three pieces of data: her utility function U_Y, her worldview Υ, and her information-state $\mathcal{I}_Y$. Thus, one naïve approach is to endow Xander with a 'simulacrum' of Ysolde: a triple $(U'_Y, \Upsilon', \mathcal{I}'_Y)$ where U'_Y is his *estimate* of U_Y, etc.

The problem is that this begets an infinite regress of self-reference. Xander's model of Ysolde must include a model of *her* model of *him*. But then *her* model should include a model of *his* model of her model of him, and so on. While amusing, this is not strategically useful.

Perfect Empathy: To avoid infinite regress, we can assume that $U'_Y = U_Y$, $\Upsilon' = \Upsilon$, and $\mathcal{I}'_Y = \mathcal{I}_Y$. Thus, Xander has *perfect knowledge* of Ysolde (and vice versa). This makes the model simple, but is highly unrealistic. In reality, people's ignorance of each other's beliefs and motives is a major factor in their decisions.

'Like me, only different': In this model, Xander imagines Ysolde as a copy of himself, with small deviations. In other words, $U'_Y = U_X + D_Y$, $\Upsilon' = \xi + \delta_Y$, and $\mathcal{I}'_Y = \mathcal{I}_X + \mathcal{D}_Y$. Here, D_Y, δ_Y, and $\mathcal{D}_Y$ are small 'deviations' which represent (for Xander) how Ysolde differs from him. As he learns more about her, he modifies D_Y, δ_Y, and $\mathcal{D}_Y$ appropriately.

This seems plausible. In real life, a common mistake is to assume others are too much like ourselves. Communication breakdowns often occur because we assume that someone else enters the conversation with the same values or background knowledge as us.

SENTIENCE: Hominids evolved as social animals, living in tightly knit communities. Our ancestors have been playing social games for a very long time, and we almost certainly have evolved specialised neurological 'wetware' for reasoning about social situations, modelling other players, etc. As we've seen, a crucial part of your model of another player is your model of their model of *you*. Thus, this hypothetical wetware must contain *self-modelling* capabilities. Perhaps this is the neurological origin of our self-awareness.

(v) Social and Political Stability

Cultures are self-perpetuating: people raised within a certain culture assimilate and recapitulate its social norms and behavioural codes. Rarely do they challenge convention, even when it seems obvious that they might personally benefit from doing so. Why? What enforces social conformity? Indeed why are governments stable? Why don't democracies become dictatorships? And why aren't dictators overthrown by their own henchmen?

To answer these questions, we must model society as a social game, and then model the game as a *dynamical system*[5].

In every round of a social game, each player has an optimal choice, determined by his worldview, utility, and current information. If we exactly knew ξ, U, and $\mathcal{I}$, we could thus predict his behaviour. The past actions of other players have partly determined his information, and his actions now will influence other players in turn. Nonetheless, the evolution of the game is (in principle) completely predictable. In other words, the game is a **dynamical system**.

This dynamical model of game evolution provides a good model of social stability. We represent society as a social game in which all citizens participate. Their social, economic, and political choices are all simply 'moves' in this game. If the social game is a dynamical system, then a dynamical attractor corresponds to a *stable society* —that is, a stable social, political, and economic order.

The simplest dynamical attractors are fixed points, limit cycles, and quasiperiodic systems. A *fixed point* is a scenario where all players converge upon a single optimal strategy, which they repeat forever. A *limit cycle* is a situation where each player reiterates the same stereotypical sequence of actions forever, in a sort of 'ritual'. A *quasiperiodic system* is similar, except that it allows players some small degree of variation in their ritual.

Clearly, none of these realistic model of society. A real society corresponds to a chaotic attractor, where small changes can trigger large long-term consequences, and history never repeats itself. Nevertheless, the attractor as a whole will be *resilient* to small perturbations. This means that the society will not collapse or radically transform because of random minor events.

[5]See page 63.

(vi) Freedom

Despite its ubiquity in both philosophy and political rhetoric, 'freedom' is a poorly understood idea. What is freedom? A naïve answer is, 'Xander is *free* if he can choose any action in his repertoire.' But in this sense, Xander is *always* free. Unless someone sticks electrodes in his brain to puppeteer his movements, he is 'free' to take any action he is physically capable of performing. (Of course, being rational, Xander will always act so as to maximise his expected utility.)

But 'freedom' is really much more complicated than this. Consider two situations:

(A) *I offer to sell you a used computer for $300.*

(B) *I point a gun at you and demand $300.*

In both situations, you are equally 'free' to choose whether or not to give me money. However, we feel you are 'more free' in **(A)**. The difference lies in the *consequences* of your choice; in terms of the 'Game' model, the difference is in your *expected utility*.

EXTORTION: Perhaps when we say 'freedom', we really mean *freedom from extortion*. Ysolde **extorts** Xander when she issues an **ultimatum** which offers him a 'choice' between an bad outcome (complying with her odious demands) and *worse* outcome (suffering her retaliation). Formally, if Xander's 'compliance' is $\mathcal{X}_*$ and Ysolde's 'retaliation' is $\mathcal{Y}_-$, we have:

$$U_{\max}[\mathcal{I}] \quad > \quad \mathbb{E}_\xi\left[U \;\langle\!\langle\; \mathcal{X}_* \cap \mathcal{I}\right] \quad > \quad \mathbb{E}_\xi\left[U \;\langle\!\langle\; \mathcal{Y}_- \cap \mathcal{I}\right] \tag{8.4}$$

This is 'extortion' because Xander's expected utility under either choice is worse than $U_{\max}[\mathcal{I}]$, which is what his prospect *would* have been if Ysolde just left him alone.

Let $\mathcal{Y}_u$ be Ysolde's communication of the ultimatum. Thus, assuming Xander has no better options than the two shown in (8.4), we have the inequality:

$$U_{\max}[\mathcal{I}] \quad > \quad U_{\max}[\mathcal{I} \cap \mathcal{Y}_u] \tag{8.5}$$

In other words, his prospects after she issues the ultimatum are worse than they were before. It is the inequality in (8.5) which characterizes extortion.

EXPLOITATION There is more to freedom than freedom from exploitation. Consider the following situations:

(C) *You are hanging from a cliff. For a mere $300, I offer to lift you to safety.*

(D) *We are standing on a cliff. I threaten to push you over the edge, unless you give me $300.*

According to our previous analysis, **(D)** is extortion, but **(C)** is not. In **(C)**, your prospect is *already* bad, and I am actually offering to *improve* it... for a price. Nonetheless, most people would find my behaviour in **(C)** almost as morally repugnant as in **(D)**. I am taking advantage of your desperation to extract concessions. This is not extortion, but *exploitation*.

Somehow, exploitation has to do with the difference between *wants* and *needs*. Consider the following scenario:

(E) *I own a beautiful painting. You want it. For a mere $300, I offer to sell it to you.*

What exactly makes **(C)** different from **(E)**? Naïvely, the answer is that you 'need' to be lifted to safety, but you don't 'need' the painting. Here, 'need' refers to issues of peril and safety. In **(C)**, your life is in immediate danger. But what about long-term danger? Here's a situation common in places with no public health-care:

(F) *You are dying from a terminal illness. For a mere $300 000, I offer you the life-saving treatment you need.*

Is this exploitation? You face certain death, albeit somewhat delayed. But what about the mere *risk* of death?

(G) *There is a 0.1% chance that you will die in a workplace accident. For a mere $300 000, I can improve safety conditions in your workplace, and eliminate this risk.*

Extortion has an exact definition, but clearly, exploitation is more slippery. Just when is your situation 'desperate' enough that my bargaining is exploitative? This issue divides political 'Left' from 'Right'. Loosely speaking, 'Leftist' economists regard exploitation as a real phenomenon, and are concerned with its prevention. 'Rightist' economists care mainly about freedom *from extortion*, and often deny the reality of 'exploitation' altogether, because it lacks a clear definition.

The 'Right' thus arrogates intellectual superiority, since 'Leftists' cannot even precisely define their key theoretical concept. For the Left to have intellectual credibility, it must define 'exploitation' as precisely as we've defined extortion. Some possible definitions:

Marxist 'surplus value' Marx defines the 'exploitation' of labour as *surplus value*: the difference between the wages that Workers receive to produce commodity **C**, and the price at which the Capitalist can sells commodity **C**. In other words, the measure of 'exploitation' is precisely the *profit* of the capitalist[6].

This definition is severely flawed, because 'surplus value' plays several important and legitimate economic roles:

- It covers the cost of *overhead* (ie. purchasing and maintaining physical capital).

[6]....hence capitalism is 'expoitative' by definition, Q.E.D.

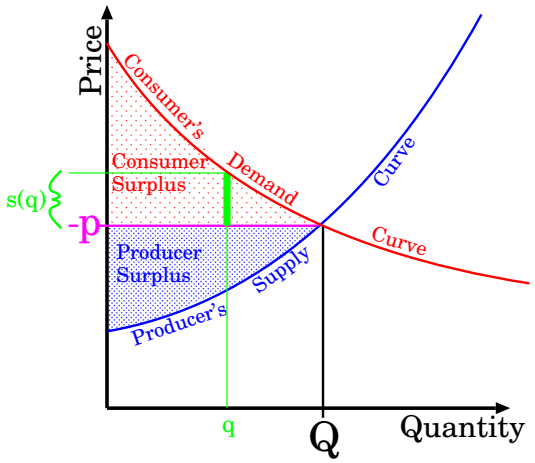


Figure 8.2: Consumer surplus vs. Producer surplus

- It provides wealth for further *investment* (making economic growth possible).
- It acts as the 'wage' which the Capitalist pays *herself* for the nontrivial labour involved in setting up and maintaining an enterprise.
- It provides a reward (ie. incentive) for investing personal wealth into production capital (thereby creating jobs and goods), instead of spending it on personal indulgence.
- It compensates the Capitalist for risking personal wealth on uncertain business ventures (instead of hoarding it).

Consumer Surplus vs. Producer Surplus Suppose a Xander (a consumer) and Ysolde (a producer) willingly enter a transaction where Xander willingly purchases a quantity Q of some commodity (say, slices of cake) of p dollars per unit. Both Xander and Ysolde willingly participate in the transaction, so we assume it is mutually beneficial. However, we can meaningfully ask the question, 'Who benefitted more'? One way to measure this is to compare the *consumer surplus* to the *producer surplus*.

In Figure 8.2, Xander's *consumer surplus* is the area of the region *above* the horizontal line p, and *below* the 'Consumer's Demand Curve'. The idea is this: distance $s(q)$ between these two curves at the point q on the 'Quantity' line is the difference between the price Xander would be *willing* to pay for his qth slice of cake, and the *actual* price he paid. Thus $s(q)$ measures Xander's opinion of how 'good a deal' he's getting —his *surplus* —for the qth slice. The aggregation of all these $s(q)$ (for q from 0 up to Q) is Xander's total surplus.

Likewise, Ysolde's *producer surplus* is the area of the region *below* the horizontal line p, and *above* the 'Producer's Supply Curve'. The reasoning is similar.

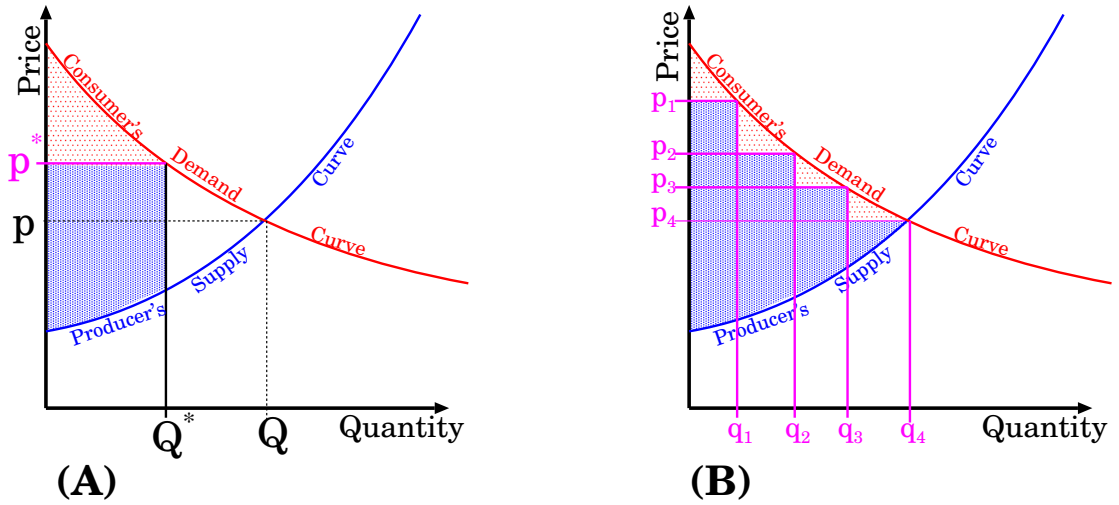


Figure 8.3: Monopoly exploitation: **(A)** Supply restriction. **(B)** Price scheduling

Although it is mutually advantageous, we might characterize this transaction as 'exploitative' if one of the players captures much more of the surplus than the other. Ysolde can do this if she has a *monopoly* on the supply of cake. As a monopolist, Ysolde can capture part of Xander's consumer surplus using in two ways[7]:

Figure 8.3(A) By *restricting supply* (or equivalently, inflating the price), Ysolde can force Xander to purchase Q^* units of cake (where Q^* is less than the equilibrium quantity Q) at a price p^*, which is higher than the equilibrium price p, and certainly *much* higher than p^c, which is Ysolde's real marginal cost for producing the Q^*th slice of cake. In this way, Ysolde makes her producer surplus much larger, by making Xander's smaller, and by imposing an 'efficiency loss' on the market (less units of cake and cash are exchanged in total).

Figure 8.3(B) By *price scheduling*, Ysolde forces Xander to pay price p_1 for his first q_1 slices of cake (which he needs the most), p_2 for his next q_2 slices of cake (which he needs slightly less), and so on. In this way, she can 'chisel' away almost all of his surplus.

We can interpret the 'cliffhanger' example **(C)** as a monopoly: I have a monopoly on the ability to pull you to safety, so I can dictate a price. If several would-be Samaritans were competing to save you, then we would likely bid one another down to an almost zero price (since the 'marginal cost of production' of pulling you to safety is virtually zero).

[7] A third strategy, called *market segmentation*, involves isolating buyers from each other, and customizing a price schedule for each buyer's individual demand curve.

But this is a rather uncomfortable fit. In the 'cliffhanger' example, the 'commodity' (ie. safety) cannot be traded in small quantities —it is an 'all or nothing'. Thus, we cannot really model the cliffhanger using the standard economic formalism of supply and demand curves. Also, in the case of a monopoly, the lack of freedom seems to come *first* (Xander's inability to deal with other suppliers), and *causes* the exploitation. This doesn't represent situations where it seems that exploitation *causes* the lack of freedom.

Finally, there are free market situations where your freedom may still be compromised. Consider the 'terminal illness' example **(F)**. It may be that the marginal cost of production of the 'lifesaving treatment' really *is* \$300 000; perhaps I am one of five competing hospitals, and I am offering you a competitive rate. So there is no monopolistic exploitation *per se*. Nevertheless, if you simply can't afford \$300 000, we might agree that your freedom (to live) has been compromised.

One thing is clear. Political rhetoric which focuses on 'freedom' as an end in itself is naïve and spurious. We are all 'free'. The question is not, 'Are you free?' The question is: 'What are your options?'

Notes

§(i) discussed the semiotics of verbal communication, distinguishing *explicit* communication from *coded communication*, *metacommunication*, etc. This analysis is similar to the *speech act* theory of Austin [2] and Grice [15]. See Bechtel [4, pp.28-29] for a discussion; I owe the 'reference letter' example to Bechtel.

§(v) presents a model of society as an attractor within a game dynamical system is similar to other 'strategic equlibria', such as as the *Nash equilibrium* of classical game theory [31], or the *evolutionarily stable strategies* of evolutionary game theory [19]. The difference is in the nature of the player interactions. In classical game theory, players have *full information* about the game state and each other's motives and abilities, so the game requires no probabilistic analysis. The Nash equlibrium yields the unique mixed strategy for each player which maximizes his *worst-case* (not *expected*) utility. In evolutionary game theory, 'mindless' organisms compete and evolve, and are driven toward an optimal distribution of genotypes (ie. mixed strategy) by selection pressure, *not* by stochastically optimizing their utilities based on imperfect information.

These appendices provide a brief and nontechnical introduction to some mathematical concepts used in the text. My advice is to read these on a 'need to know' basis.

A Functions

Let $\mathcal{P}$ be the set of all people in the world, and let $\mathcal{M} \subset \mathcal{P}$ be the subset of all men. For any person $p \in \mathcal{P}$, let $f(p)$ be the father of p. Thus, $f(\text{Penelope}) = \text{Michael}$ means that Penelope is Michael's daughter.

Now let $\mathcal{W} \subset \mathcal{P}$ be the set of women, and let $m(p)$ be the *mother* of p. Thus, $m\Big(f(\text{Penelope})\Big)$ is the mother of the father of Penelope —in other words, the Penelope's *paternal grandmother*. On the other hand, $f\Big(m(\text{Penelope})\Big)$ is the father of the mother of Penelope —in other words, the Penelope's *maternal grandfather*.

These are examples of *functions*. If $\mathcal{P}$ and $\mathcal{M}$ are two sets, then a **function** from $\mathcal{P}$ to $\mathcal{M}$ is some mechanism which *assigns* a unique element of $\mathcal{M}$ to every element of $\mathcal{P}$. We normally indicate this by writing "$f : \mathcal{P} \longrightarrow \mathcal{M}$". In the previous examples:

- $\mathcal{P}$ is the set of people, $\mathcal{M}$ the set of men, and $f : \mathcal{P} \longrightarrow \mathcal{M}$ was the function assigning, to each $p \in \mathcal{P}$, the *father* of p.
- $\mathcal{P}$ is the set of people, $\mathcal{W}$ the set of men, and $m : \mathcal{P} \longrightarrow \mathcal{W}$ was the function assigning, to each $p \in \mathcal{P}$, the *mother* of p.

Depending upon their intended application, functions are often given other names, such as *transformations*, *labellings*, *representations*, or *mappings*.

- A *transformation* is a function $f : \mathcal{A} \longrightarrow \mathcal{B}$ which 'transforms' objects in $\mathcal{A}$ into objects in $\mathcal{B}$.

  For example, let $\mathcal{A} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \ldots \mathtt{z}\}$ be the set of lower-case Roman letters, and let $\mathcal{B} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \ldots \mathtt{Z}\}$ be the set of upper-case letters, and define $f(\mathtt{a}) = \mathtt{A}$, $f(\mathtt{b}) = \mathtt{B}$, etc.

- A *labelling* is a function $f : \mathcal{A} \longrightarrow \mathcal{B}$ which attaches a 'label' in $\mathcal{B}$ to each object in $\mathcal{A}$.

  For example, let $\mathcal{A}$ be a set of people, and let $\mathcal{B}$ be the set of their names. For each person $a \in \mathcal{A}$, let $f(a)$ be that person's names.

- A *representation* is a function $f : \mathcal{A} \longrightarrow \mathcal{B}$ which 'represents' objects in $\mathcal{A}$ with objects in $\mathcal{B}$.

  For example, let $\mathcal{A}$ be the set of countries of the world, and let $\mathcal{B}$ be the set of delegates of the United Nations General Assembly. For each country $a \in \mathcal{A}$, let $f(a)$ be the delegate of that country at the General Assembly.

- A *mapping* is a function $f : \mathcal{A} \longrightarrow \mathcal{B}$ which draws a 'map' of $\mathcal{A}$ on $\mathcal{B}$.

  For example, let $\mathcal{A}$ be the set of all points on the Earth's surface, and let $\mathcal{B}$ be the set of all points on a globe map. For any point $a \in \mathcal{A}$, let $f(a)$ be the corresponding point on the map $\mathcal{B}$.

On important function is the **identity function Id**, which simply transforms any object into itself. That is, for any a,

$$\mathbf{Id}(a) \quad = \quad a.$$

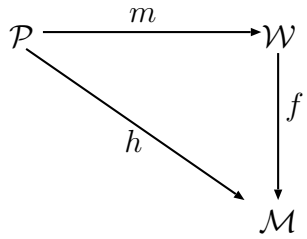
This may seem kind of stupid, but it is often useful to introduce this function into mathematical expressions, for the same reason that it is often useful to 'multiply by one' or 'add zero' when manipulating algebraic expressions.

Function Composition: We can create a new function by applying two functions in succession. This is called *function composition*. For example, suppose $f : \mathcal{P} \longrightarrow \mathcal{M}$ is the 'father' function, and $m : \mathcal{P} \longrightarrow \mathcal{W}$ is the 'mother' function. We define a new function $h : \mathcal{P} \longrightarrow \mathcal{M}$ by $h(p) = f\Big(m(p)\Big)$. Thus, $h(p)$ is the *father of the mother of* p —that is, the *maternal grandfather* of p. We indicate this by writing:

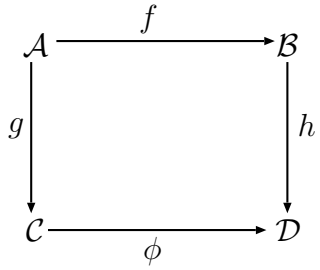
$$h \quad = \quad f \circ m.$$

Notice that, in general, $f \circ m$ is not the same function as $m \circ f$. In the above example, $f \circ m$ is the *maternal grandfather* function, whereas $m \circ f$ is the *paternal grandmother* function.

Commuting Diagrams: Mathematical arguments often involve composing together networks of functions, in various orders. It is often important to know when two different sequences of compositions produce the same function. We represent this diagrammatically via a *commuting diagram*. For example, the diagram:



says that $h = f \circ m$. The diagram:



says:

- $\mathcal{A}, \mathcal{B}, \mathcal{C}$ and $\mathcal{D}$ are sets;

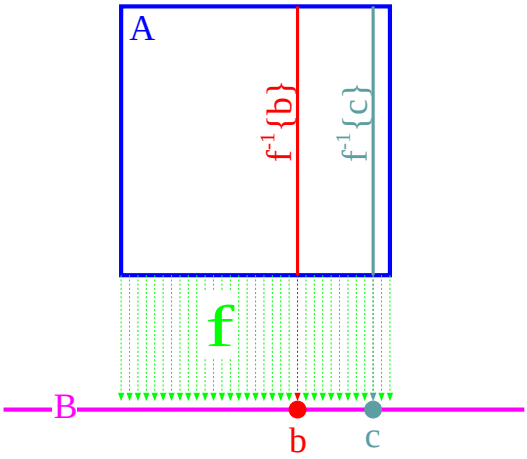


Figure A.1: $f^{-1}\{b\}$ is the **fibre** over b

- $f : \mathcal{A} \longrightarrow \mathcal{B}$, $g : \mathcal{A} \longrightarrow \mathcal{C}$, $h : \mathcal{B} \longrightarrow \mathcal{D}$, and $\phi : \mathcal{C} \longrightarrow \mathcal{D}$ are functions; and
- $h \circ f \;=\; \phi \circ g$.

Preimages Consider the set of all *children* of Mary —that is, the set of all people having Mary as their mother. Symbolically, we write this set:

$$\{p \in \mathcal{P} \;;\; m(p) \;=\; \text{Mary}\}.$$

This is called the **preimage** of Mary under the function m, and is written $m^{-1}\{\text{Mary}\}$. Likewise,

$$f^{-1}\{\text{Michael}\} \quad = \quad \{p \in \mathcal{P} \;;\; f(p) \;=\; \text{Michael}\}.$$

is the set of Michael's children, and

$$f^{-1}\left(f^{-1}\{\text{Michael}\}\right) \quad = \quad \Big\{p \in \mathcal{P} \;;\; f\left(f(p)\right) \;=\; \text{Michael}\Big\}.$$

is the set of Michael's grandchildren through his sons.

If $f : \mathcal{A} \longrightarrow \mathcal{B}$ is a function, and $b \in \mathcal{B}$, then we sometimes refer to the preimage $f^{-1}\{b\}$ as the **fibre** over b. This is illustrated by Figure A.1.

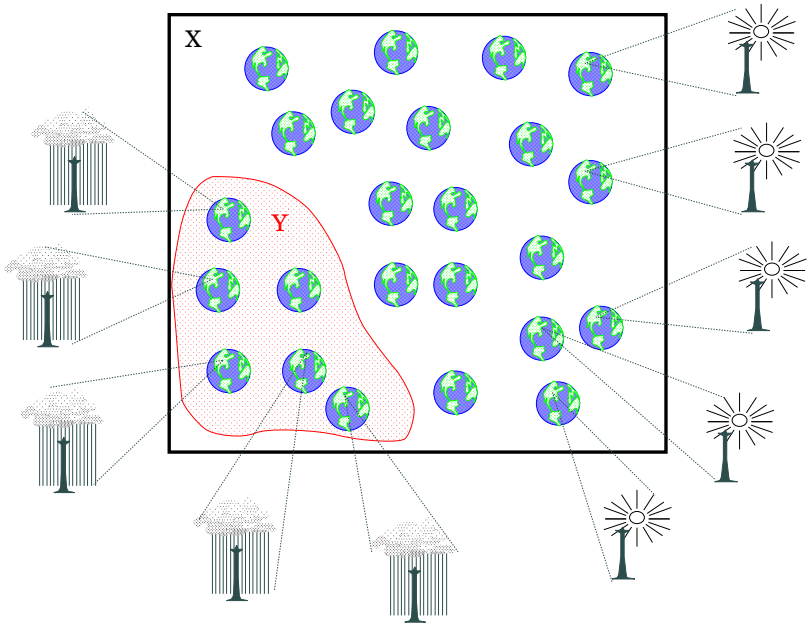


Figure B.1: **Probability Measures:** $\mathbf{X}$ is the 'set of all possible worlds'. $\mathbf{Y} \subset \mathbf{X}$ is the set of all worlds where it is raining in Toronto.

B Probability

Imagine that $\mathbf{X}$ is the space of possible **states** of some "universe" $\mathcal{U}$. For example, in Figure B.1, $\mathcal{U}$ is the weather over Toronto; thus, $\mathbf{X}$ is the set of all possible weather conditions. Subsets of $\mathbf{X}$ are called **events**; each event corresponds to some assertion about $\mathcal{U}$. For example, in Figure B.1, the assertion "It is raining in Toronto" corresponds to the event $\mathbf{Y} \subset \mathbf{X}$; the set of of all states in $\mathbf{X}$ where it is, in fact, raining in Toronto.

A *probability measure* is an object ρ which assigns a 'size' to each subset[1] of $\mathbf{X}$. The *probability* of the assertion, 'It is raining in Toronto' is the *size* of $\mathbf{Y}$, which is denoted $\rho[\mathbf{Y}]$.

As a measure of size, ρ must satisfy certain natural properties:

(I: Additivity[2]) *If* $\mathbf{R}$ *and* $\mathbf{S}$ *are* disjoint *subsets of* $\mathbf{X}$ *(ie.* $\mathbf{R} \cap \mathbf{S} = \emptyset$*), then* $\rho[\mathbf{R} \cap \mathbf{S}] \quad = \quad \rho[\mathbf{R}] + \rho[\mathbf{S}]$.

For example, if $\mathbf{R}$ is the event, 'It is raining in Toronto', and $\mathbf{S}$ is the event 'It is snowing in Toronto', (and we assume that these events are mutually exclusive), then **(I)** says:

[1] Technically, ρ must be defined on a sigma-algebra of 'measurable' sets, but we will ignore this issue.

[2] Actually, this is formulated in terms of *countable* disjoint unions, but I'm suppressing this to keep things simple.

The probability that it is either raining *or* snowing is the probability that it is raining, *plus* the probability that it is snowing.

The second property is:

(II: Monotonicity) *If* $\mathbf{R} \subset \mathbf{P}$*, then* $\rho[\mathbf{R}] \leq \rho[\mathbf{P}]$.

For example, if **R** is the event, 'It is raining in Toronto', and **P** is the event 'There is some kind of precipitation in Toronto', then **(II)** says,

The probability that it is raining is less than or equal to the probability that there is some kind of precipitation.

Another way to express **(II)** is as follows:

(II') *If* **R** *and* **W** *are two events, then* $\rho[\mathbf{R} \cap \mathbf{W}] \leq \rho[\mathbf{R}]$ *and* $\rho[\mathbf{R} \cap \mathbf{W}] \leq \rho[\mathbf{W}]$.

For example, if **R** is the event, 'It is raining in Toronto', and **W** is the event, 'It is windy in Toronto', then **(II')** just says

The probability that it is raining *and* windy is lesser or equal to the probability that it is raining, and lesser or equal to the probability that it is windy.

Notice that, in a probability space, *smaller* events correspond to assertions with *more* information. The assertion, 'it is raining *and* windy' carries more information than either the assertion 'it is raining' or the assertion 'it is windy'; hence, the corresponding event $\mathbf{R} \cap \mathbf{W}$ is *smaller* than the events **R** and **W**.

Intuitively, then, since **X** is the *largest* set in the space, it represents the *least* information. Indeed, since **X** is just the entire space of weather-states of Toronto, the event **X** is the vacuous assertion 'Something is true'.

Conversely, since $\emptyset$ is the smallest set, it represents the *most* information. Indeed, $\emptyset$ essentially represents *impossibility:* an assertion so restrictive that it *cannot* be true. Hence we have the third property:

(III) $\rho[\emptyset] = 0$, and $\rho[\mathbf{X}] = 1$.

If ρ satisfies axioms **(I)**, **(II)** and **(III)**, it is called a **probability measure**. The pair $(\mathbf{X}, \rho)$ is then called a **probability space**. Let's look at some other examples:

Dice: Imagine a six-sided die. In this case, the $\mathbf{X} = \{1, 2, 3, 4, 5, 6\}$. The probability measure ρ is completely determined by the values of $\rho\{1\}, \rho\{2\}, \ldots, \rho\{6\}$. For example, suppose:

$$\begin{array}{lcllcl} \rho\{1\} & = & 1/12 & \rho\{4\} & = & 1/6 \\ \rho\{2\} & = & 1/12 & \rho\{5\} & = & 1/6 \\ \rho\{3\} & = & 1/3 & \rho\{6\} & = & 1/6 \end{array}$$

Then the probability of rolling a 2 *or* a 3 is $\rho\{2, 3\} = 1/12 + 1/3 = 5/12$.

Urns: Imagine a giant clay 'urn' full of 10 000 balls of various colours. You reach in and grab a ball randomly. What is the probability of getting a ball of a particular colour?

In this case, $\mathbf{X}$ is the set of balls. If we label the balls with numbers, then we can write $\mathbf{X} = \{1, 2, 3, \ldots, 10\ 000\}$. Suppose that the balls come in colours RED, GREEN, BLUE, and VIOLET. For simplicity, let

$$\begin{array}{rcll} \mathbf{R} & = & \{1, 2, 3, \ldots, 500\} & \text{be the set of all RED balls;} \\ \mathbf{G} & = & \{501,\ 502,\ 503, \ldots, 1500\} & \text{be the set of all GREEN balls;} \\ \mathbf{B} & = & \{1501,\ 1502,\ 1503, \ldots, 3000\} & \text{be the set of all BLUE balls;} \\ \mathbf{V} & = & \{3001,\ 3002,\ 3003, \ldots, 10\,00\} & \text{be the set of all VIOLET balls;} \end{array}$$

Thus (assuming the urn is 'well-mixed' and all balls are equally probable), the probability of getting a RED ball is

$$\rho\left[\mathbf{R}\right] \quad = \quad \frac{500}{10000} = 0.05$$

while the probability of getting a GREEN ball *or* a VIOLET ball is

$$\rho\left[\mathbf{G} \sqcup \mathbf{V}\right] \quad = \quad \rho\left[\mathbf{G}\right] + \rho\left[\mathbf{V}\right] \quad = \quad \frac{1000}{10000} + \frac{7000}{10000} \quad = \quad 0.1 + 0.7 = 0.8$$

The unit interval: Let $\mathbb{X}$ be the set of all real numbers between 0 and 1. Suppose $\mathbf{V}$ is a *line-segment* inside $\mathbb{X}$; say, the set of all points between 1/8 and 1/2. Then the probability of $\mathbf{V}$ is just it's *length*; in this case, 3/8. This is just the odds of picking a 'random number' inside $\mathbf{V}$.

Conditional Probability

Suppose that, over a historical period of 10000 days:

- It rained in Toronto on 4500 days.
- It rained in Montréal on 3500 days.
- It rained in Toronto *and* Montréal on 3000 days.

Assuming this sample accurately reflects the underlying statistics, we can conclude, for example:

The probability that it will rain in Montréal on any given day is $\frac{3500}{10000} = 0.35$.

This probability estimate is made assuming 'total ignorance'. Suppose, however, that you already *knew* it was raining in Toronto. This might modify your wager about Montréal. Out of the 4500 days during which it rained in Toronto, it rained in Montréal on 3500 of those days. Thus,

Given that it is raining in Toronto, the probability that it will *also* rain in Montréal, is $\frac{3000}{4500} = \frac{2}{3} = 0.6666\ldots$.

If **T** is the event "It is raining in Toronto", and **M** is the event "It is raining in Montréal", then **M**∩**T** is the event "It is raining in Toronto *and* Montréal". What we have just concluded is:

$$\rho\left[\mathbf{M} \text{ given } \mathbf{T}\right] \quad = \quad \frac{\rho\left[\mathbf{M} \cap \mathbf{T}\right]}{\rho\left[\mathbf{T}\right]}$$

we call this the **conditional probability** of **M**, **given T**.

In the previous example, meteorological information about Toronto modified our wager about Montréal. Suppose instead that the statistics were as follows: over 10000 days,

- It rained in Toronto on 5000 days.
- It rained in Montréal on 3000 days.
- It rained in Toronto *and* Montréal on 1500 days.

Then we conclude:

- The probability that it will rain in Montréal on any given day is $\frac{3000}{10000} = 0.3$.
- *Given* that it is raining in Toronto, the probability that it will *also* rain in Montréal, is $\frac{1500}{5000} = 0.3$.

In other words, the rain in Toronto has *no influence* on the rain in Montréal. Meteorological information from Toronto is *useless* to predicting Montréal precipitation. If **M** and **T** are as before, we have:

$$\frac{\rho\left[\mathbf{M} \cap \mathbf{T}\right]}{\rho\left[\mathbf{T}\right]} \quad = \quad \rho\left[\mathbf{M}\right].$$

Another way to write this:

$$\rho\left[\mathbf{M} \cap \mathbf{T}\right] \quad = \quad \rho\left[\mathbf{M}\right] \cdot \rho\left[\mathbf{T}\right].$$

We say **M** and **T** are **independent**.

C Stochastic Processes

A stochastic process is a particular kind of probability measure, which represents a system *randomly evolving in time*.

Imagine $\mathcal{S}$ is some complex system, evolving randomly. For example, $\mathcal{S}$ might be a die being repeatedly rolled, or a publically traded stock, a weather system. Let $\mathbf{X}$ be the set of all possible *states* of the system $\mathcal{S}$, and let $\mathbb{T}$ be a set representing *time*. For example:

- If $\mathcal{S}$ is a rolling die, then $\mathbf{X} = \{1, 2, 3, 4, 5, 6\}$, and $\mathbb{T} = \mathbb{N} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, \ldots\}$ indexes an infinite sequence of successive dice rolls.
- If $\mathcal{S}$ is the weather over Toronto, then $\mathbf{X}$ is the 'space of all weather states.' Since the weather evolves continuously, $\mathbb{T}$ is the set of all real numbers, denoted $\mathbb{R}$.

We represent the (random) evolution of $\mathcal{S}$ by assigning a probability to every possible *history* of $\mathcal{S}$. A history is an assignment of a state (in $\mathbf{X}$) to every moment in time (i.e. $\mathbb{T}$); in other words, it is a function $h : \mathbb{T} \longrightarrow \mathbf{X}$. The *set of all possible histories* is thus the space $\mathbf{H} = \mathbf{X}^{\mathbb{T}}$.

An *event* —an subset of $\mathbf{H}$ —thus corresponds to a some collection of possible histories. Usually we specify such an 'event' by stipulating that specific worldstates occurred at specific points in time. A *probability measure* on $(\mathbf{H}, \mathcal{H})$ is then a way of assigning probabilities to such assertions. Examples:

Toronto's weather: Suppose y and t are two points in time —say, yesterday and tomorrow. Let $\mathbf{R}$ and $\mathbf{S}$ are two subsets of $\mathbf{X}$ —say, the set of 'raining' weather states and the set of 'snowing' weather states. Then the event:

$$\mathbf{E} = \{h \in \mathbf{H} \; ; \; h(y) \in \mathbf{R} \text{ and } h(t) \in \mathbf{S}\}$$

is the set of all histories described by the assertion, 'It rained yesterday and will snow tomorrow.' Thus, if $\rho[\mathbf{E}] = 0.2$, this means that there is a 0.2 probability that it rained yesterday and will snow tomorrow.

A (fair) six-sided die: Now $\mathbf{X} = \{1, 2, \ldots, 6\}$ and $\mathbb{T} = \mathbb{N}$. Thus, $\mathbf{H} = \{1, 2, \ldots, 6\}^{\mathbb{N}}$ is the set of all possible infinite sequences $\mathbf{x} = [x_1, x_2, x_3, \ldots]$ of elements $x_n \in \{1, 2, \ldots, 6\}$. Such a sequence represents a record of an infinite succession of dice throws. The sigma algebra $\mathcal{H}$ is generated by all cylinder sets of the form:

$$\langle y_1, y_2, \ldots, y_N \rangle \quad = \quad \left\{\mathbf{h} \in \{1, 2, \ldots, 6\}^{\mathbb{N}} \; ; \; h_1 = y_1, \ldots, h_N = y_N\right\}$$

where $N \in \mathbb{N}$ and $y_1, y_2, \ldots, y_N \in \{1, 2, \ldots, 6\}$ are constants. For example,

$$\langle 3, 6, 2, 1 \rangle \quad = \quad \left\{\mathbf{h} \in \{1, 2, \ldots, 6\}^{\mathbb{N}} \; ; \; h_1 = 3, \; h_2 = 6, \; h_3 = 2, \; h_4 = 1\right\}$$

This event corresponds to the assertion, "The first time, you roll a *three*; the next time; a *six*, the third time, a *two*; and the fourth time, a *one*".

Assuming the die is fair, this event should have probability $\frac{1}{6^4} = \frac{1}{1296}$. More generally, for any $y_1, y_2, \ldots, y_N \in \{1, 2, \ldots, 6\}$, we should have:

$$\rho\langle y_1, y_2, \ldots, y_N\rangle = \frac{1}{6^N}$$

If the probabilities deviate from these values, we conclude the dice are loaded.

D Boolean Algebras and Information

Let $\mathbf{X}$ be a space. A **Boolean algebra** is a collection $\mathcal{B}$ of *subsets* of $\mathbf{X}$ so that, for any subsets $\mathbf{A} \subset \mathbf{X}$ and $\mathbf{B} \subset \mathbf{X}$,

(BA1) If $\mathbf{A}$ and $\mathbf{B}$ are elements of $\mathcal{B}$, then $\mathbf{A} \cap \mathbf{B}$ is also an element of $\mathcal{B}$.

(BA2) If $\mathbf{A}$ and $\mathbf{B}$ are elements of $\mathcal{B}$, then $\mathbf{A} \cup \mathbf{B}$ is also an element of $\mathcal{B}$.

(BA3) If $\mathbf{A}$ is elements of $\mathcal{B}$, then $\mathbf{X} \setminus \mathbf{A}$ is also an element of $\mathcal{B}$.

For example, the **power set** of $\mathbf{X}$ is the *set of all subsets* of $\mathbf{X}$:

$$\mathcal{P}(\mathbf{X}) = \{\mathbf{S} \,;\, \mathbf{S} \subset \mathbf{X}\}$$

the **null algebra** contains only two elements: the empty set and all of $\mathbf{X}$:

$$\mathcal{N} = \{\emptyset, \mathbf{X}\}$$

It is easy to check that both $\mathcal{P}(\mathbf{X})$ and $\mathcal{N}$ satisfy the properties **(BA1)**, **(BA2)**, and **(BA3)**.

Boolean algebras are a good way to mathematically represent a state of partial knowledge. Suppose we want to know the location of an unknown point $x \in \mathbf{X}$. The *knowledge represented by* $\mathcal{B}$ is the information about x that one obtains from knowing, for every $\mathbf{B} \in \mathcal{B}$, whether or not x is an element of $\mathbf{B}$. Thus, if $\mathcal{C} \supset \mathcal{B}$ is a larger Boolean algebra, then $\mathcal{C}$ contains 'more' information than $\mathcal{B}$. The power set $\mathcal{P}(\mathbf{X})$ is the 'largest' Boolean algebra, and thus, contains the 'most' information. At the opposite extreme, the null algebra represents a state of total ignorance.

To understand this, we'll examine some examples.

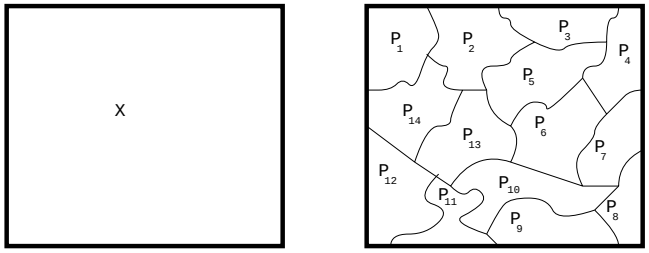


Figure D.1: $\mathcal{P}$ is a partition of $\mathbf{X}$.

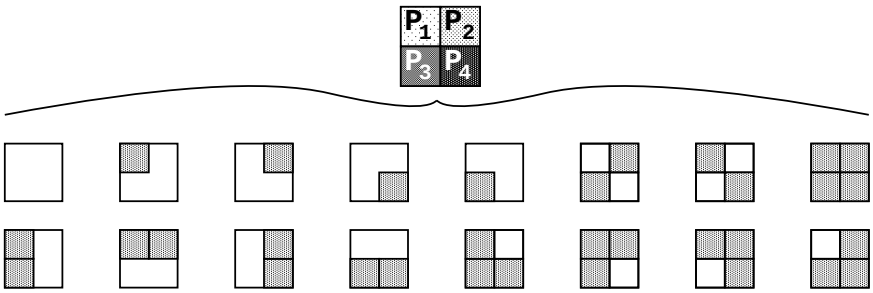


Figure D.2: **The Boolean algebra generated by a partition:** Partition the square into four smaller squares, so $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4\}$. The corresponding Boolean algebra contains 16 elements.

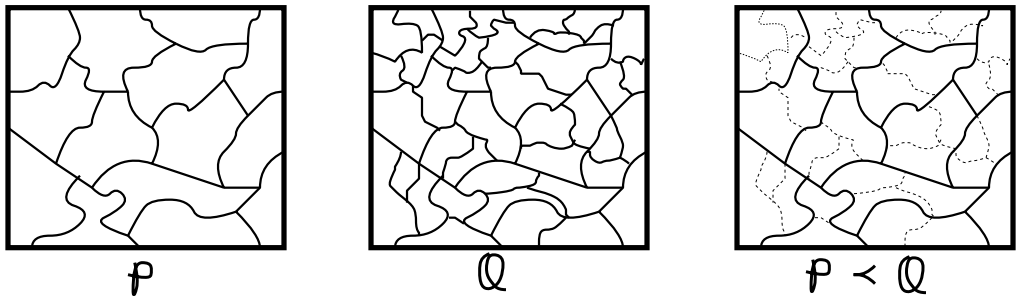


Figure D.3: Partition $\mathcal{Q}$ refines $\mathcal{P}$ if every element of $\mathcal{P}$ is a union of elements in $\mathcal{Q}$.

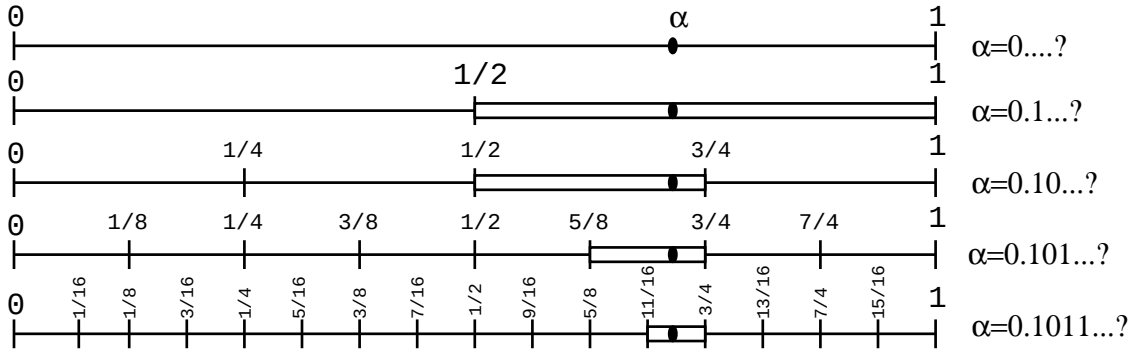


Figure D.4: Finer dyadic partitions reveal more binary digits of α

Partitions: The simplest Boolean algebras are those generated by *partitions*. Figure D.1 shows a **partition** of $\mathbf{X}$: a collection $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N\}$ of disjoint subsets, such that $\mathbf{X} = \bigsqcup_{n=1}^{N} \mathbf{P}_n$. The sets $\mathbf{P}_1, \ldots \mathbf{P}_N$ are called the **atoms** of the partition. Figure D.2 shows the Boolean algebra *generated* by $\mathcal{P}$: the collection of all possible unions of $\mathcal{P}$-atoms:

$$\sigma(\mathcal{P}) \;=\; \{\mathbf{P}_{n_1} \sqcup \mathbf{P}_{n_2} \sqcup \ldots \sqcup \mathbf{P}_{n_k} \;;\; n_1, n_2, \ldots, n_k \in [1..N]\}$$

Thus, if $\mathsf{card}\,[\mathbf{P}] = N$, then $\mathsf{card}\,[\sigma(\mathbf{P})] = 2^N$.

If $\mathcal{Q}$ is another partition, we say that $\mathcal{Q}$ **refines** $\mathcal{P}$ if, for every $\mathbf{P} \in \mathcal{P}$, there are $\mathbf{Q}_1, \ldots, \mathbf{Q}_N \in \mathcal{Q}$ so that $\mathbf{P} \;=\; \bigsqcup_{n=1}^{N} \mathbf{Q}_P$; see Figure D.3. We then write "$\mathcal{P} \prec \mathcal{Q}$" It follows that

$$\Big(\; \mathcal{P} \prec \mathcal{Q} \;\Big) \iff \Big(\; \sigma(\mathcal{P}) \subset \sigma(\mathcal{Q}) \;\Big).$$

Dyadic Partitions of the Interval Let $\mathbb{R}$ be the *real line* —the set of all real numbers (imagined as an infinite, straight line). If a and b are two real numbers, and $a < b$,

then the **interval** from a to b is just the line segment connecting these points. We use the notation $[a, b)$ to indicate this interval[1]. Formally,

$$[a, b) \quad = \quad \{r \in \mathbb{R} \; ; \; a \leq r < b\}.$$

For example, $[0, 1) \; = \; \{r \in \mathbb{R} \; ; \; 0 \leq r < 1\}$ is the line segment from 0 to 1, which is called the **unit interval**.

Let $\mathbb{I} = [0, 1)$, and consider the following sequence of **dyadic partitions**, illustrated in Figure D.4:

$$\begin{aligned}
\mathcal{P}_0 &= \{\mathbb{I}\} \\
\mathcal{P}_1 &= \left\{ \left[0, \frac{1}{2}\right), \; \left[\frac{1}{2}, 1\right) \right\} \\
\mathcal{P}_2 &= \left\{ \left[0, \frac{1}{4}\right), \; \left[\frac{1}{4}, \frac{1}{2}\right), \; \left[\frac{1}{2}, \frac{3}{4}\right), \; \left[\frac{3}{4}, 1\right) \right\} \\
\mathcal{P}_3 &= \left\{ \left[0, \frac{1}{8}\right), \; \left[\frac{1}{8}, \frac{1}{4}\right), \; \left[\frac{1}{4}, \frac{3}{8}\right), \; \left[\frac{3}{8}, \frac{1}{2}\right), \; \left[\frac{1}{2}, \frac{5}{8}\right), \; \left[\frac{5}{8}, \frac{3}{4}\right), \; \left[\frac{3}{4}, \frac{7}{8}\right), \; \left[\frac{7}{8}, 1\right) \right\} \\
&\vdots \quad \vdots \quad \vdots
\end{aligned}$$

Suppose α is an unknown element of $\mathbb{I}$. Then:

$$\begin{aligned}
&\left(\text{ knowing which element of } \mathcal{P}_0 \text{ contains } \alpha \right) \iff \left(\text{ knowing } \alpha \text{ with precision } \frac{1}{2} \right). \\
&\left(\text{ knowing which element of } \mathcal{P}_1 \text{ contains } \alpha \right) \iff \left(\text{ knowing } \alpha \text{ with precision } \frac{1}{4} \right). \\
&\left(\text{ knowing which element of } \mathcal{P}_2 \text{ contains } \alpha \right) \iff \left(\text{ knowing } \alpha \text{ with precision } \frac{1}{8} \right).
\end{aligned}$$

and, in general,

$$\left(\text{ knowing which element of } \mathcal{P}_n \text{ contains } \alpha \right) \iff \left(\text{ knowing } \alpha \text{ with precision } \frac{1}{2^{n+1}} \right).$$

Thus, the 'finer' the partition $\mathcal{P}_n$, the more 'information' about α it provides.

PARTITIONS OF A SQUARE Let $\mathbb{I}^2$ be a unit square, and let $\mathcal{P}$ be a partition of $\mathbb{I}^2$, with associated Boolean algebra $\sigma(\mathcal{P})$. The information contained in $\sigma(\mathcal{P})$ is the information about $x \in \mathbb{I}^2$ you obtain from knowing which atom of $\mathcal{P}$ contains x.

[1]Technically, this is called a **half-open interval**.

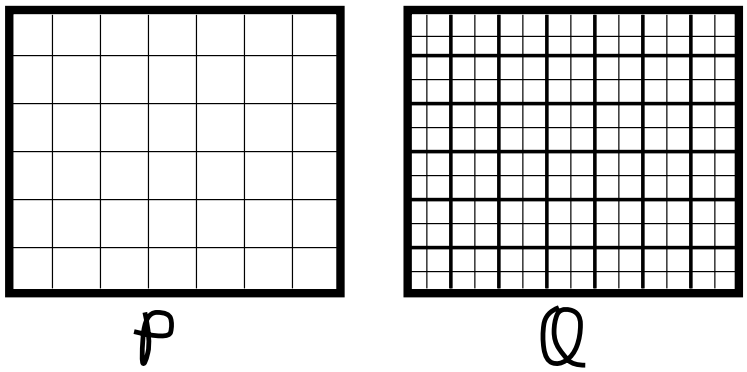


Figure D.5: A higher resolution grid corresponds to a finer partition.

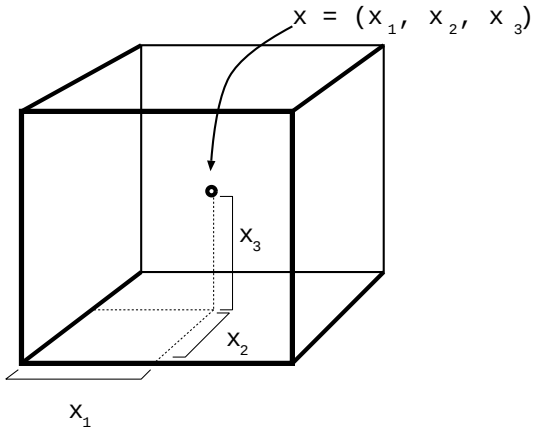


Figure D.6: Our position in the cube is completely specified by three coordinates.

Recall that partition $\mathcal{Q}$ **refines** $\mathcal{P}$ if every element of $\mathcal{P}$ can be written as a union of atoms in $\mathcal{P}$, as shown in Figure D.3 on page 109. Thus, $\sigma(\mathcal{Q})$ contains *more* information than $\sigma(\mathcal{P})$, because it specifies the location of x with greater 'precision'. For example, suppose $\mathcal{P}$ and $\mathcal{Q}$ were *grids* on $\mathbb{I}^2$, as in Figure D.5. If $\mathcal{P} \prec \mathcal{Q}$, then $\mathcal{Q}$ is a *higher resolution* grid, providing proportionately better information about spatial position.

Projections of the Cube Consider the **cube** $\mathbb{I}^3$ of sidelength 1 (see Figure D.6). We will imagine this cube to be the set of all points (x_1, x_2, x_3) whose coordinates are all between 0 and 1. In other words

$$\mathbb{I}^3 \quad = \quad \{(x_1, x_2, x_3)\ ;\ 0 \leq x_1 < 1;\ 0 \leq x_2 < 1 \text{ and } 0 \leq x_3 < 1\}$$

Recall that $\mathbb{I}$ is the unit interval (see page 110). Then we could also write:

$$\mathbb{I}^3 \quad = \quad \{(x_1, x_2, x_3)\ ;\ x_1 \in \mathbb{I};\ x_2 \in \mathbb{I};\ \text{and } x_3 \in \mathbb{I}\}.$$

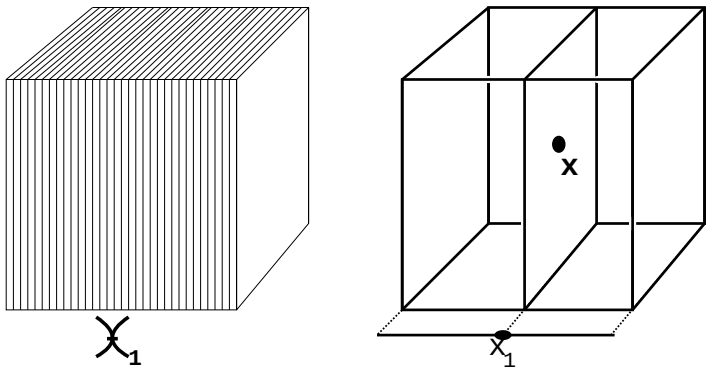


Figure D.7: The Boolean algebra $\mathcal{X}_1$ consists of "vertical sheets", and specifies the x_1 coordinate.

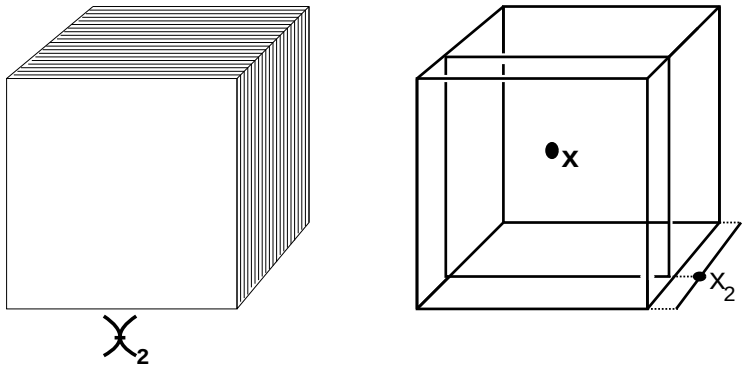


Figure D.8: The Boolean algebra $\mathcal{X}_2$ consists of "vertical sheets", and specifies the x_2 coordinate.

As shown in Figure D.6, the position of $x \in \mathbb{I}^3$ is completely specified by three coordinates (x_1, x_2, x_3). The information embodied by each coordinate corresponds to a certain Boolean algebra.

Consider the *projection* onto the *first* coordinate, $\mathbf{pr}_1 : \mathbb{I}^3 \longrightarrow \mathbb{I}$. In other words, if $\mathbf{x} := (x_1, x_2, x_3) \in \mathbb{I}^3$, then $\mathbf{pr}_1(\mathbf{x}) = x_1 \in \mathbb{I}$.

Consider the *pulled back* Boolean algebra $\mathcal{X}_1 := {\mathbf{pr}_1}^{-1}(\mathcal{I})$, (where $\mathcal{I}$ is the power set of $\mathbb{I}$). Roughly speaking, $\mathcal{X}_1$ consists of all "vertical sheets" in the cube (Figure D.7). Thus knowing the coordinate x_1 is equivalent to knowing which of these vertical sheets contains $\mathbf{x}$.

Next, consider the *projection* onto the *second* coordinate, $\mathbf{pr}_2 : \mathbb{I} \longrightarrow \mathbb{I}$. That is, $\mathbf{x} := (x_1, x_2, x_3) \in \mathbb{I}^3$, then $\mathbf{pr}_2(\mathbf{x}) = x_2 \in \mathbb{I}$.

The pulled back Boolean algebra $\mathcal{X}_2 := {\mathbf{pr}_2}^{-1}(\mathcal{I})$ consists of the 'vertical sheets' shown in Figure D.8. Knowing the coordinate x_2 is equivalent to knowing which of these sheets

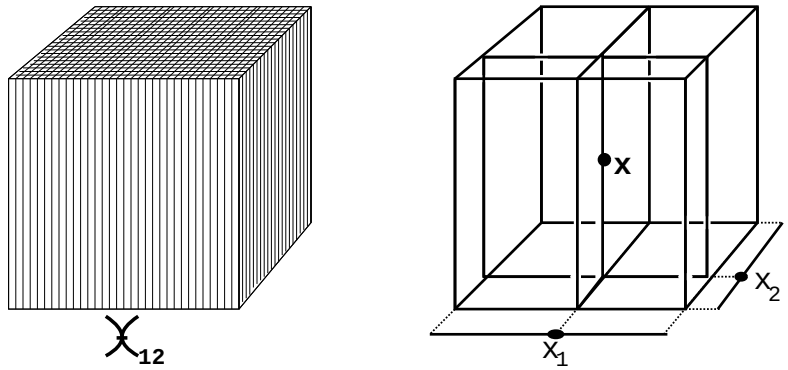


Figure D.9: The Boolean algebra $\mathcal{X}_{12}$ completely specifies (x_1, x_2) coordinates of $\mathbf{x}$.

contains $\mathbf{x}$.

Finally, let $\mathbb{I}^2$ be the unit square, and consider the *projection* onto the *first two* coordinates, $\mathbf{pr}_{1,2} : \mathbb{I}^3 \longrightarrow \mathbb{I}^2$. That is, if $\mathbf{x} := (x_1, x_2, x_3) \in \mathbb{I}^3$, then $\mathbf{pr}_{1,2}(\mathbf{x}) = (x_1, x_2) \in \mathbb{I}^2$.

Let $\mathcal{X}_{12} := {\mathbf{pr}_{1,2}}^{-1}(\mathcal{I}^2)$ (where $\mathcal{I}^2$ is the power set of $\mathbb{I}^2$). Elements of $\mathcal{X}_{12}$ look like "vertical fibres" in the cube (Figure D.9). Thus, $\mathcal{X}_{12}$-related information specifies exactly which of these vertical fibres contain $\mathbf{x}$, and exactly which vertical fibres *don't* contain $\mathbf{x}$. From this, we can reconstruct arbitrarily accurate information about the coordinates x_1 and x_2. In other words, the information contained in $\mathcal{X}_{12}$ is exactly the same information contained in the (x_1, x_2) coordinates of $\mathbf{x}$.

Remarks: A *sigma algebra* is like a Boolean algebra, but it is closed under *countable* unions and intersections, not just finite ones. For most applications, mathematicians work with sigma-algebras, not Boolean algebras. I've confined this discussion to Boolean algebras to minimize the technical complexity.

Bibliography

[1] Keith J. Arrow. *Individual Values and Social Choice*. John Wiley & Sons, New York, 2nd edition, 1963.

[2] J.L. Austin. How to do things with words. In J.O. Urmson and G.J. Warnock, editors, *Philosophical Papers of J.L. Austin*. Oxford UP, Oxford, 1962/1970.

[3] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[4] William Bechtel. *Philosophy of Mind: An overview for cognitive science*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[5] Gregory J. Chaitin. *Algorithmic Information Theory*. Cambridge UP, Cambridge, MA, 1987.

[6] Gregory J. Chaitin. *The Limits of Mathematics*. Springer-Verlag, New York, 1997.

[7] Tim Crane. *The Mechanical Mind*. Penguin, London, 1995.

[8] Robert Cummins. *Meaning and Mental Representation*. MIT Press, Cambridge, MA, 1989.

[9] Peter Danielson. *Artificial Morality: Virtuous Robots for Virtual Games*. Routledge, New York, 1992.

[10] Daniel C. Dennet. *Consciousness Explained*. Little, Brown & Co., Boston, 1991.

[11] Frances Egan. Individualism, computation, and perceptual content. *Mind*, 101, 1992.

[12] G. Frege. Über sinn and bedeutung. *Zeitshrift fur Philosophie und philosphische Kritik*, 100:25–50, 1892.

[13] David P. Gauthier. *Morals by Agreement*. Oxford UP, Oxford, 1986.

[14] Kurt Gödel. *On Formally Undecidabe Propositions of Principia Mathematica and Related Systems*. Dover, New York, 1962.

[15] H.P. Grice. Logic and conversation. In P. Cole and J.L. Morgan, editors, *Speech acts*, pages 45–58. Academic Press, New York, 1975.

[16] J. Flum H. D. Ebbinghaus and W. Thomas. *Mathematical Logic*. Springer-Verlag, New York, 1984.

[17] Douglas Hofstadter. *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books, New York, 1979.

[18] John E. Hopcroft and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Co., 1979.

[19] Jörgen W. Weibull. *Evolutionary Game Theory*. M.I.T. Press, Cambridge, MA, 1995.

[20] Stuart Kauffman. *Investigations*. Oxford UP, 2002.

[21] Ki Hang Kim and Fred W. Roush. *Introduction to Mathematical Consensus Theory*. Marcel Dekker, New York, 1980.

[22] Naomi Klein. *No Logo: taking aim at the brand bullies*. Vintage, Toronto, 2000.

[23] David C. Korten. *When Corporations Rule the World*. Kumarian Press, Bloomfield, CT, 2nd edition, 2001.

[24] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1st phoenix edition edition, 1964.

[25] Kenneth Kunen. *Set Theory: An introduction to independence proofs*. North-Holland, London, 1980.

[26] J.R. Lucas. Minds, machines, and Gödel. *Philosophy*, 36:112, 1961.

[27] John Mighton. *Possible Worlds*. Consortium Book Sales, March 2001.

[28] Noam Chomsky and Edward S. Herman. *Manufacturing Consent: The Political Economy of Mass Media*. Pantheon, 2002.

[29] Roger Penrose. *The Emperor's New Mind*. Oxford University Press, 1989.

[30] Karl Petersen. *Ergodic Theory*. Cambridge University Press, New York, 1989.

[31] R. Duncan Luce and Howard Raiffa. *Games and Decisions*. Dover, New York, 1957.

[32] Richard J. Barnet and John Cavanagh. *Global Dreams, Imperial Corporations, and the New World Order*. Simon and Schuster, New York, 1994.

[33] J.M. Roberts. *The Penguin History of the Twentieth Century*. Penguin Books, London, 1999.

[34] Donald G. Saari. *Basic Geometry of Voting*. Springer-Verlag, New York, 1995.

[35] Edward Said. *Orientalism*. Vintage, New York, October 1979.

[36] Edward Sapir. *Language*. Harcourt Brace, New York, 1921.

[37] Edward Sapir. *Selected Writings in Language, Culture, and Personality*. Univ. of California Press, Berkeley, 1949.

[38] A. Sen. *Collective choice and social welfare*. Holden-Day, San Francisco, 1970.

[39] A. Sen. The impossibility of a Paretian liberal. *Jornal of Political Economy*, 78:152–157, 1970.

[40] Patrick Suppes. *A Probabilistic Theory of Causality*. North-Holland, Amsterdam, 1970.

[41] Patrick Suppes. *Axiomatic Set Theory*. Dover, New York, 1972.

[42] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[43] Peter Walters. *An Introduction to Ergodic Theory*. Springer-Verlag, New York, 1982.

[44] Benjamin Lee Whorf. *Language, Thought and Reality*. M.I.T. Press, Cambridge, MA, 1956.