1 Introduction and Motivation

Last fall I was returning to the department after a two year "leave of absence," during which I had little contact with psychology apart from a few stints as a TA and coming in to attend this seminar. I had taken this time off because I felt unclear about my own direction after leaving a research job in the artificial intelligence lab at SRI. By the beginning of my leave I had made a decision to devote myself to psychology rather than AI, but in four years as a sometime graduate student while working off campus I had not found a coherent program of research in psychology. My experimental results consisted of a few curiosities obtained in questionnaires, which I now think I understand a little better thanks to support theory. And while I often thought of ideas for experiments, I had a hard time summoning the motivation to do them.

Although it may be rationalizing things a bit, I think one reason for my lethargy stemmed from the fact that I lacked an overall picture of psychology and the problems I wanted to focus on. I had developed such a picture for myself of AI, but I was not really that interested in AI's goal, which is to build intelligent machines. I was intensely interested in psychology and its goals, in human problems, conflicts and so on, but in psychology I lacked the feeling that I could put my arms around the discipline, so to speak — the field seemed too big, too unwieldy, as Brian Knutson once said a "tower of Babel" and I didn't want to contribute to the cacaphony without something truly useful to say. However much this hindsight is true to my feelings at the time, I did begin to become interested in general psychology. I began reading more psychology, I watched Phil Zimbardo's series, I became a TA for psych 1, and I designed and taught an introductory course for high school
students the summer before I returned last fall. I also began to do something that I think many graduate students and others in psychology who share my need for comprehensive understanding have probably done in their private moments: I tried to classify, for myself, the empirical phenomena that I had learned into a small number of categories.

It seemed like an obvious thing to do, and I am sure that many veterans of the field carry these classifications in their heads all the time, noticing that the kind of schemas that lead one to misremember items in an office might also underlie racial stereotypes, that availability is a lot like the salience effects that seem to account for attribution errors, or that the freezing of attitudes has some properties in common with Einstellung. I had also become interested in such classifications for their potential pedagogical value. Having spent a good deal of time being involved in the teaching of introductory psychology, I thought about how I would like to be able, for instance, to write an introductory textbook one day. It seemed to me then (and it still does) that there is a confusion wrought by the present approach to the teaching of introductory psychology which reflects the state of the field as a whole. As the psych 1 coordinator last year I found that students frequently complain that there are too many theories presented, too many results to memorize, and that they leave psychology having been exposed to a hodgepodge of findings and ideas with no overall framework that would facilitate retention. Most often this sentiment manifests itself in the thus perpetuated impression, which I find remarkable for students who have gone through an entire quarter of psychology, that experimental results hold no special place for them in understanding human behavior over and above, say, their own opinions based on experience. Psychology is fun to talk about, some of us are “born psychologists,” there are lots of interesting psychologists on television, and, as one student said who was trying to get into psych 1, it’s a good thing to take if you need a non-science class. Drastically reducing the number of concepts in psychology seemed to me to be one way to get the main messages of the field across more effectively, at least for most of our students who are non-psych majors.

So as I returned last fall from spending two years plugged into Robert

---

1 One student, writing on an evaluation form for psych 1 last year, attacked the notion of an availability bias as being based on experiments that are set up to prove the researcher’s point and not on genuine tests that apply broadly. An interesting criticism, it reveals that the student doesn’t understand that the experiment isn’t meant to imply that the availability heuristic always, or even usually, leads to error, but only that it is revealed in the errors it creates under some conditions.
Nozick's experience machine, I was vaguely excited by my new little friends, the three or four categories of psychological phenomena that I took along to department colloquia and that were helping me to feel better about psychology and myself. There was a category for salience effects, one or two for schemas and overgeneralization-type effects, and one that I thought of as being linked to maintaining control and order (things like cognitive dissonance and ego-enhancement). As I say, many others may think of things in terms of some such taxonomy, but the exact nature of mine and theirs would likely differ quite a lot (as witness the categories mentioned by Eleanor Gibson this summer in her APS Keynote Address calling for broad principles!), and, except for eminent psychologists like Gibson, they would probably have the wisdom to keep their categories to themselves and not try to build a doctoral thesis out of it. But I have made a lengthy project out of my categories that got a boost just as I was returning last year when I went to a talk by Geoffrey Miller, whom some of you will remember as a graduate student here, who was just getting ready to leave for a postdoctoral position at Sussex. Miller has already achieved some recognition as part of a new breed of so-called "evolutionary psychologists," who have become champions of the idea that psychology can be unified under one or a few principles, in this case, coming from evolutionary theory. Miller drew a helpful graph during his farewell talk that has influenced me (FIGURE 1). He said that he thinks of psychology as a field with a theoretical structure very close to its set of results, and that evolution (natural selection) can play the role of the topmost of principles, guiding those at the bottom. What interested me in this view at the time was not so much the idea that effects could be derived from evolution, but rather that the structure could be built up from the bottom. Indeed this reflected my own initial approach in searching for grand principles of psychology which was to look for similarities in the results and theories as they are currently talked about in the literature. But as time wore on last fall I found myself thinking more and more about the origins, the means for deriving, the principles that I had very loosely put together inductively. I switched to a more deductive style, trying to reason from first principles, seeking to understand more of the structure and interrelationships among the principles of psychology. In early January, building on work I had done over several years at SRI, I felt that I had an insight that has led me down the path that I am about to describe, a way of deriving principles by considering the resources available to an evolving agent. I now find my original

---

2in a meeting of Roger Shepard's lab group, October 5, 1992.
clusterings to be too simplistic in terms of the theory so derived, and I think that it can lead to useful predictions, clarifications, and explanations across a wide range of phenomena.

I'll get more contentful in a moment, but first I'd like to make a few remarks to motivate this particular project of deductive theory-building for me and for psychology. The idea that psychology is ready for greater theoretical unification has achieved some currency among a few of our most prominent elders. The last two APS Keynote addresses (by William Estes (Estes, 1993) and Eleanor Gibson (Gibson, 1993)) have focused on the point that psychologists know more than we are given credit for by society and that we need to pay more attention to what our core concepts are. But I am also interested in the possibility that a more unitary theoretical understanding of our field may help psychology to clear up its own confusions. While the body of experimental results in psychology is truly vast, the majority of experiments and the theories that have been derived from them can be exceedingly difficult to apply, even for someone who is a specialist in the field. In many areas of psychology, there are competing theories that each explain part of the data, and for many real world problems there are multiple psychological effects that have been identified and could be operating in various degrees, corresponding to radically different pictures of what is actually happening. One psychologist, considering issues of censorship, might feel that the third person effect (the result that we tend to think that others will be more influenced by speech than we are) is the most relevant finding from psychology, while another might focus on the vulnerability to harmful stereotypes or the effects of images on attitudes. Who is right? Who is to say? The problem is obviously very complex, but in our emphasis on experimentation and on incremental, partial explanations modified in the dialectic of science, academic psychologists have I think managed to create a field that is weak both in theory and in application.

It seems to me that building psychology on experimental results and limited, tentative theories has been a very reasonable way to proceed over the last half century. Certainly the numerous failures of bold theorizing, as well as the limited success of idiographic approaches in applied and professional psychology, lead one to think that experimentation has been the right level on which to focus. And despite the large database of experimental results so far accumulated, it would usually be wrong to claim that even the totality of these results could adjudicate between all the competing theories. We are clearly going to need more experiments. However, I do think that we have reached a point where consolidation and synthesis are called for, in other
words: theory.

At the same time, I think we have reached a point where new experiments are no longer revealing many big new effects. In literature departments they often say, “The grand narratives have already been told, and what remain to be written are subnarratives.” It has certainly happened in other fields of science, and even within psychology, that a relatively short but fruitful period has yielded most of the macroscopic effects that a theory must then contend with. Pure physicists are not doing experiments in classical mechanics these days. In my case, the field I ostensibly know best in psychology is judgment and decision making. Amos is my advisor and the strongest influence on my thinking, but it is very daunting for a student with big ideas to fall under his tutelage. Amos and his colleagues seem to me to have written the map of experimental phenomena for judgment and decision making in a way that can only be done once, like the discovery of visual illusions that occurred in a flurry over less than a century and virtually ended a few generations ago (this is according to my old officemate, Mike McBeath, a perceptual researcher). As I think about effects I observe in human reasoning and consider testing them experimentally, I usually find that my idea is just an instance of a previous finding in judgment and decision making or in some other area of cognitive or social psychology. And when I think it is probably a new idea, it is usually either incorrect (as I have learned after eight questionnaire studies since I have been a psychology student) or it is, at best, a variation, a clarification, or a twist on something already investigated, adding decimal points to the theory. This does not mean that very interesting experimental work is not left to be done, only that it is not coming easily to me. I have no doubt that a large part of these observations reflect my own lack of ingenuity as a hypothesizer, so my claim on that basis amounts to what David Rumelhart calls the “argument by lack of imagination.” But as, perhaps, better evidence for this idea that new effects are going away, I find that although I am still sometimes surprised by research findings from other psychologists, I can usually explain them to myself in terms of past findings and known effects, though sometimes not uniquely. I am still a relative newcomer to psychology, but if I have at least this level of understanding of the main effects in human inference, it suggests that it may be worthwhile to disseminate, popularize, and explore these concepts in the real world, in other words: to apply them.

---

3This is a very loose paraphrase of some ideas in Jean-Francois Lyotard’s The Postmodern Condition: A Report on Knowledge (Lyotard, 1979).
These, then, are the two directions in which I would like to develop as a psychologist and that I think are fruitful directions for the future of the field, at least as things look to me now: toward theory and applications. I think of the relation between theory, experiments and applications as like an hourglass (FIGURE 2). Theoretical ideas at the top are filtered through experiments and can then be tested by application. But the hourglass can be turned upside down, so that ideas from real world applications lead to experiments and then to theories. What works less well, I think I have learned, is trying to go directly from theory to application or vice versa. For now I am mostly concerned with the relation between theory and experiment, because that is were I would like to focus my energy for the dissertation.

Some of my perspective on theory might be understood as an attempt to synthesize ideas I have learned from the two great spirits who flank this hallway. Amos and Roger Shepard (who is not here this year) have each influenced me a great deal, and yet in many ways I think of them as polar opposites. One only has to look inside their offices to realize this. Amos has taught me that errors and biases are central to psychology, really they are what make it interesting; and that attempts to show that these errors are really “optimality in disguise” that can be accounted for by a teleological appeal to the constraints of the world are based on a faith for which there is little evidence and abundant counterevidence. With Roger, I share the dream of a psychology of universal principles akin to those of the physical sciences. Amos, in my experience, is a fairly staunch believer in inductive science and has become famous as a debunker of a priori theories. Roger, on the other hand, believes that a good deal can be learned about what is necessary in mental life by appeal to the facts of the world and derivations about what must be true of intelligent beings that can survive in it, that, as he says, if we are smart enough, we can figure out what psychological laws should be by the process of thought experiments (which should of course be tested by real ones). I may be wrong about this split, but I find myself attracted to Roger’s methodology when thinking about Amos’s phenomena, the biases and errors that reflect limitations on our ability to internalize the facts of the world. Of course, an undertaking to derive these biases at this moment in history has the tremendous advantage that we know what biases must be explained by virtue of four decades of empirical research, so this is by no means a purely deductive exercise.

The proto-theory I will present in a moment will give rise to some new hypotheses and predictions, but for now its main advantages are that (1) it attempts to explain psychological effects in terms of a few core concepts
derived from primitive physical and informational principles, and (2) it contains a reason to think that these concepts are exhaustive for characterizing the range of biases in human nature. A few years ago, I was talking with my friend and fellow graduate student Evan Heit about a remark Amos had casually made that it might be interesting to teach psych 1 someday. I said, “Amos does teach psych 156 (judgment and decisionmaking). He could just change the number.” And Evan said, “Yeah, that’s all the important stuff anyway.” Well, Evan’s comments may be prophetic, for I see in the principles that I will present really just a recasting of the judgmental heuristics of availability (and accessibility), representativeness, and anchoring. But I have what I think are new arguments for the exhaustiveness of these concepts and a way to derive them (with some holes to be filled in still). The exhaustive characterization of the space of biases lends some credence to the claim that we have run out of basic cognitive biases, because the ones I derive have all been found before in one form or another.

The proto-theory suggests sources for the limitations on human rationality that are buried in our evolutionary history, much more deeply, I think, than is commonly assumed. As such it is akin to work in two areas that have received much attention recently: evolutionary psychology and limited rationality. I would like, first, to clarify the relationship between my approach and the usual approaches to these two areas.

1.1 Evolutionary Psychology

Most work in evolutionary psychology (see, for example, the collection in Barkow et al., 1992) applies the logic of an evolutionary theory (for example, natural selection, or the selfish gene theory) to human behavior, as in the following schema:

(1) We observe behavior $x$, either widespread, universal, or consistent over a long evolutionary timespan. Why has $x$ survived/been selected for in evolution?

This question is sometimes criticized as unscientific because the answer proposed can be difficult to falsify.

A related question schema that sometimes seems to underlie research is:

(2) Behavior $y$ would be advantageous for survival, or would have been so in the past. Do humans/animals display $y$, as we would predict?
This question is also vulnerable to an attack on its usefulness, because whatever the answer, we are certainly not going to be disconfirming evolution—we already take that as given. Questions of form (2) are particularly useful when the answer helps us to discriminate between competing accounts of how evolution has operated.

The schemata (1) and (2) are fascinating questions, often worth asking, though posing them in specific ways and answering them can be very difficult. I want to ask a different type of question about evolution and psychology, namely, what are the factors that constrain the evolution of our minds? More schematically, questions of the following form:

(3) Behavior \( z \) would (or would probably) be advantageous for survival and/or reproduction, yet human beings typically do not exhibit \( z \). Why not?

The usual answer to this question is that human evolution either is not an optimizing process (but merely “satisficing” (Simon, 1981)) or else it has not reached its final state. Perhaps, this line of reasoning would go, there just has not been enough time in evolutionary history for \( z \) to emerge, but eventually it will (if it is truly beneficial). There are other possible explanations, such as that \( z \) may interfere with behaviors more important than \( z \) for survival.

I don’t want to quibble with this response, but I do want to remember that in many cases behavior is suboptimal not just because it doesn’t need to be, but also for deep reasons owing to physical and informational limitations that have impinged on every organism, indeed on every physical structure, since the beginning of chemical evolution. The history of evolution is a history of many triumphs (including some astonishingly simple ones) over these limitations, so that for instance in the present, highly adaptive form of the human brain, the limitations may no longer dictate many of these departures from optimality. But (and this is the thesis statement) because evolution proceeds, temporally, by incremental improvements in fitness over existing structures, our brain’s construction still tends to produce behavior that is systematically biased away from that which would optimize fitness. It is the biases so produced that I wish to focus on. There is another possible category of biases that might be induced by a disparity between behavior or inference that is fit for genetic survival, on the one hand, and that which reflects either morality or objective reality; these are biases induced by self-preservation. For the present work I do not wish to consider this category of bias because it would take a great deal of work even to establish that these
biases definitely exist. But, to admit to a personal bias of mine, I think that judgmental biases away from reality *beyond* those that result from resource limitations can never be shown to serve the survival of a life form. Morality is a different matter, and I would simply prefer to leave that question for another time.

1.2 Limited Rationality

Limited or bounded rationality theories have appeared recently in artificial intelligence (Russell and Subramanian, 1992; Horvitz, 1988), the philosophy of mind (Cherniak, 1986), and, of course, economics (Simon, 1955; Salop and Stiglitz, 1977; Philips, 1988). I would include in this category work by behavioral economists portraying people as adaptive to their informational and temporal resources (for example, Payne, Bettman, and Johnson, 1993), as well as work in so-called “evolutionary economics” that models human adaptation (Arthur, 1993). The questions asked usually have the form:

(1) Given constraints on resources (e.g. time and space), how might one design a cognitive architecture that is optimally suited to achieving particular goals? (Normative Question)

The questions I want to ask, on the other hand, are more like the following:

(2) Given systematic and widespread departures from optimality in human beings, how and to what extent can resource constraints, operating now or in evolutionary history, account for the departures? (Descriptive Question, restatement of (3) above)

and,

(3) Given the pervasive constraints on physical/informational resources for evolving forms, what biases would we expect to evolve, and are these the ones we observe? (Predictive Question)

At some level, I will try to argue in section 2 that we can answer the predictive question with reference to a deductive theory and the effects we observe. But to answer the predictive question in a more complete sense, we will need a strong model, and perhaps some computer simulations.

I am interested in how much of psychology is necessary given our level of complexity/functioning and how much is accidental. And by psychology,
I mean implicitly a theory of human limitations and biases. Surely this is central to what psychology is about, in addition to explaining how we came to be such intelligent creatures. Indeed, one of the defining properties of what it is to be human is to be limited, as when we say “I am only human,” or, simply, “I am human.” If my theory is successful, then, it should tell us that certain psychological biases are inevitable given the structure of the universe. At the same time, it may help us to identify effects that could have been otherwise, and would distinguish us from intelligent extraterrestrials that we may meet.

2 The Proto-Theory

I call the present version of the theory a “proto-theory” to reflect its state of incompleteness. Ultimately, I want the theory to which this is a precursor to be able to derive the basic biases of psychology from first principles, and the biases so derived should be sufficient, either individually or in combination, to explain a broad class of (though probably not all) experimental results. Henceforth, my proposal for how to start. The argument will proceed in several steps.

2.1 An Evolutionary Ontology

The fundamental principle of evolution is incontestable because it is a tautology: That which is fitted to survive and multiply will be likely to survive and multiply. The history of the universe can therefore be viewed as a competition, over time, between abstract forms (chemical and macroscopic structures) for prevalence in space (FIGURE 3). This is a picture of the entire universe. In fact, it goes beyond that: it is a picture of the entire history, and if you like, future of the universe as well. It is obviously highly schematic. It is picture of evolution, probably best thought of as chemical evolution, with the lines delimiting regions of space occupied by a particular form at each point in time since the Big Bang.\(^4\) The spatial dimension is

\(^4\)Physicists currently conjecture that at the instant of the Big Bang the universe consisted of a single form of matter, but that due to the spontaneous breaking of symmetry and the production of variations in the surrounding energy field as the universe expanded and cooled, matter was quickly separated into the many different particles that are part of the standard model (the myriad quarks, leptons, and gluons or force carriers). A precursor of the final breaking was, on this view, induced by a particle known as the Higgs boson, which is predicted from the gauge field theory of Weinberg and Salam as an explanation.
normalized for each point in time by the expanding size of the universe.

The regions are not meant to refer to particular forms and the diagram is not in any sense drawn to scale. Still, it illustrates several possibilities for the evolution of a form. Although some forms are present at the beginning, represented by the bottom of the diagram, some come into existence later, branching off of or evolving from other forms. Sometimes a form will tend to crowd out other forms, reducing the amount of space the others take up, and sometimes it will die out or (more likely) evolve into new forms that will replace it. What defines a particular form and the moment of change into a new form is obviously quite arbitrary, a result of how we see the world and divide it into categories, but it seems reasonable to suppose that there are natural kinds that set some limits on interpretation. The concept of a form, because it is so abstract, can be applied to such diverse structures as elementary particles, macroscopic objects, species, social organizations, and even ideas, and we can speak of evolution occurring for each of these.

From this point of view, the only real notion of success that evolution suggests to us is the amount of space that a form occupies over time. I am tempted, in fact, to draw the diagram upside down, with time flowing down rather than up in order to emphasize that evolution does not necessarily imply progress. But of course, although there are some exceptions, evolution does tend to produce greater complexity of forms because complexity has marginal (incremental) advantages for survival of existing forms when it increases the amount of control that a form has over its fate. This does not, however, imply that a complex form is more fit for survival than a less

for how the electro-weak gluons came to be separated. This particle (if it exists) should be observed at energies that would be produced in the Superconducting Supercollider (if it were ever built). Verifying the Higgs particle’s existence would lend credence to the idea that the universe originated from a single form of matter, or very few. According to the Big Bang theory, from the elementary particles created in the initial moment, cosmic evolution has produced the elements, molecules, and higher forms that we observe (Weinberg, 1977; Pagels, 1983; Lederman, 1993). There are controversies in cosmology about whether there was one Big Bang or several. For the present argument their outcome should not matter much.

Researchers have recently reported that the brain of the common house cat appears to have gotten about a third smaller since its ancestors were domesticated by the Ancient Egyptians (citation?). This is because the extra brain structure is no longer necessary for cats who rely on their masters. Generally, in order for complexity to diminish through natural selection, there must be some advantage in fitness owing to the reduced complexity. In the case of the cat, it is hypothesized that cats with smaller brains have been produced more plentifully because their mothers are then able to fit more of them inside themselves during gestation.
complex one of a different line, because the latter may have a more effective means of survival. Thus, rocks are examples of very successful forms even though they are obviously less complex and less able to control their fate in the face of an aggressive nemesis than were the dinosaurs, whom rocks have outlasted. Rocks simply achieve their survival in a different way from life forms, by resilient solidity rather than by reproduction.

Complexity, or structure, has no representation in the diagram, but could be thought of as a function of the spatio-temporal projections that constitute individual instances of each form, adding a third dimension to Figure 3. The lines of the diagram would then be contour lines representing structure, which ideally could be ordered on a scale, for example, by Kolmogorov's measure or a related complexity function. But evolution, at least in a finite history, guarantees nothing about the achievement of control or increasing complexity for a form. The only things we can say with relative certainty are about the limitations on this control, a topic to which we now turn.

2.2 Resources and Limitations

An evolving form, whose survival and multiplication properties might or might not give it a tendency to take over all of space, is nonetheless kept from doing so by limitations in its resources. Resources are the dimensional quantities of the physical universe, all such as are necessary and helpful for the survival and multiplication (or the growth\(^6\)) of physical forms.

In constructing an ontology of resources, I want to find the most useful way to partition them. This suggests four criteria that the list of resource types or categories should satisfy: First, the list should be complete, meaning that anything describable as a resource should fit into one of the categories. Second, each category should be necessary, meaning that each category should be a type of resource. Third, the each category should be distinct so that nothing describable as a resource fits into more than one category. And fourth, the list should partition resources into natural categories suitable for describing limitations that are common within a type and distinct for different types.

I want to begin by assuming that everything that exists in the universe is potentially a resource. We can then choose a division of reality that satisfies the criteria above, and such that at least some part of each category consists of resources. Specifically, I want to divide the universe into space, time,

---

\(^6\)I will use the term growth to mean both survival and multiplication.
mass-energy, and information. I will call this the fourfold division.

In this division, the first three resources (space, time, and energy) are the irreducible quantities of the universe and are instantiated in the fourth (information). Time is a scalar quantity, space is a three-dimensional vector quantity, and energy can be a scalar, a vector, a tensor, or some even more complicated mathematical object depending on the chosen representation, basically incorporating all of the dimensions needed in order to characterize matter at a particular location and instant (see below). Information can be thought of as a function $I$ relating space, time, and energy. For example, we could define $I$ as taking space-time into energy, yielding the history of the universe. Facts about the universe described in ordinary language would qualify as information because in principle we could reduce them to statements about $I$, and differing levels of information for an agent would correspond to better or worse knowledge of $I$. In addition, I take all of the facts about mathematics, and the universal physical constants, to be part of information since, to the extent that they could be viewed as resources, they are implicit in the empirical facts. All of the resources could be viewed as sets, and the portions available to an agent as subsets of these.

2.2.1 Completeness and Mutual Exclusivity

(What follows under this subheading is an attempt to establish the feasibility of the fourfold division for different physical theories. The reader who is already convinced of the exhaustiveness and mutual exclusivity of the resources, and who is not interested in the details, may want to skip to subsection 2.2.2. In this subsection I have especially relied on the remarkable summary of physics in Roger Penrose's *The Emperor's New Mind* (1989) to supplement my faded undergraduate physics knowledge.)

The main benefit of the fourfold division is that its four resource concepts are intuitive in addition to being precisely definable. I think that for our purposes the intuitive concepts will also, by and large, correspond extensionally to their technical versions. This would obviously not be true of more

---

7 Any fact that is uttered or written in a book is represented in the information function, admittedly in a complicated way, because the sound energy or the particles corresponding to ink on a page bear relations of correspondence with their meanings (the facts they describe), and all of the information about the correspondence must be contained in $I$. How exactly this works is the subject of semantics, and I do not presently have a theory of it more grand than this.

8 This section is especially tentative, representing my current best understanding.
complicated ontologies like the standard model of elementary particles, nor would it be true for some simple ontologies like (space, time, particles, fields). In making the claim of exhaustiveness for the fourfold division I am partially relying on quite recent speculations in physics about the underlying simplicity of the universe. But the ontology should ideally be robust and flexible with respect to our knowledge of physics.

Let us consider how we might characterize reality with progressively more complete physical theories:

1. In rectilinear (nonrotational) Newtonian mechanics, space is Euclidean and its coordinates are defined with respect to an inertial reference frame. The only material dimension we need is mass. Information can then be defined as the assignment of a mass and a position (center of mass) to each object (which corresponds to a spatio-temporal region) at each instant. Mechanical energy can be computed from this information because the velocities of each object are implicit in the information we have about their position at each moment in time. To take the simple case of a body falling directly toward the earth without air resistance, kinetic energy (energy of motion) is $\frac{1}{2}mv^2$; potential energy due to the earth’s gravity is $mgh$, where $h$ is height above the earth, and the total of potential plus kinetic energy remains constant as the object falls (Halliday and Resnick, 1978).9

2. The above formulation of pure mechanics is unsatisfying (and untrue to Newton) because it neglects angular momentum. Also, in order to avoid introducing another category of reality corresponding to objects, we must represent them as complicated regions of space and time. If we move to the Hamiltonian formulation of classical mechanics, named for William Rowan Hamilton (1805-1865), both types of momentum can be represented in a six-vector composed of three positional and three momentum coordinates for a given object. Total energy is then represented by the Hamiltonian function $H$ defined on this information, and changes in energy for an object by its partial derivatives (forces and velocities) (Penrose, 1989). To achieve a non-

---

9Interestingly, we find in the early pages of Newton's *Principia* a reference to the concepts that correspond to the fourfold division and a link to mental error: “I do not define time, space, place, and motion, as being well known to all. Only I must observe, that the common people conceive those quantities under no other notions but from the relation they bear to sensible objects. And thence arise certain prejudices...” (quoted in Barrow, 1991).
object-based representation, we could give mass (using a mass density function) and momentum values for each point in space at a particular time, and use the deterministic equations of classical mechanics to compute the changes over time.\textsuperscript{10} This would correspond to a version of the Hamiltonian formulation in which the phase space has infinite dimensionality.\textsuperscript{11} We could have done something similar in the above Newtonian case (number 1) as well,\textsuperscript{12} illustrating that our choice of physical theory does not determine the representation.

3. When we combine classical mechanics with Maxwell’s electromagnetic theory, we must add three dimensions of current and a scale representing charge density to the mass and momentum, so that we now have an eight-dimensional vector quantity characterizing energy,\textsuperscript{13} with information being a function from space at a particular time into the eight-vector. The phase space history is still deterministic. Energy itself now comes from both gravitational and electromagnetic fields, and we can calculate each type from Newton’s and Maxwell’s laws (Halliday and Resnick, 1978; Penrose, 1989).\textsuperscript{14}

4. Einstein (and, apparently, Poincaré before him\textsuperscript{15}) realized that Maxwell’s equations implied that the “Galilean relativity” assumption of classical mechanics, namely that the laws of physics hold equally for moving and stationary reference frames, is untrue for the combined Newton and Maxwell theory. The result was the special theory of relativity. In special relativity, we can define the space-time grid as the rest frame or simultaneous space of the observer at a temporal origin. The experienced time, space, mass, and information are then all defined relative

\textsuperscript{10} As Penrose (1989) points out, we could not do this in practice in a continuous-space system because the values of the initial conditions would have to be known with infinite precision.

\textsuperscript{11} I will have to check on this point.

\textsuperscript{12} Specifically, we could assign mass and velocity nonzero to all points in space-time corresponding to centers of mass, with zeroes everywhere else.

\textsuperscript{13} I need to check on the exhaustiveness of this, to see that we needn’t specify additional quantities to derive the actual electric and magnetic fields from Maxwell’s equations.

\textsuperscript{14} Even in Maxwell’s time it was known that this did not give a complete theory of the effects of electromagnetism, since it does not specify the effects of the fields on a charged particle’s velocity. Also, as Einstein discovered, since electrons move close to the speed of light, their mass depends on their velocity. These problems required further work by Lorentz, Poincaré, Einstein, and Dirac and were not solved to a satisfactory level until 1938 (Penrose, 1989).

\textsuperscript{15} See (Penrose, 1989)
to the motion of the observer in this frame. Mass and energy are now seen as equivalent \( (E = mc^2) \), and the conserved quantity is an energy-momentum four-vector in space-time whose projection on the reference time axis (or the temporal component) describes the mass of an object according to the observer. The vector’s length corresponds to the rest-mass of the object, and the spatial components describe the momentum. Special relativity is deterministic in that initial values on the simultaneous space determine later events, and, in fact, the events on which something depends are bounded in space by the light cone-sphere extending from the event back to initial time, since information is limited by the speed of light (Penrose, 1989; Tipler, 1978).

5. In general relativity, unlike in the special theory, simultaneous space is no longer well defined because, due to the curvature of space-time induced by gravity, “the arena,” as Roger Penrose puts it, “joins in the very action taking place within itself!” (Penrose, 1989, p. 217) “Instead [of simultaneous space],” as Penrose explains, “one may use the more general notion of a spacelike surface. Such a surface... is characterized by the fact that it lies completely outside the light cone at each of its points—so that locally it resembles a simultaneous space.” (ibid., p. 214) Information about matter and the gravitational and electromagnetic fields at a particular point in an observer’s time in this hyperspace is characterized by an energy-momentum tensor, from which we can compute the curvature of space-time. This tensor requires, in its approximate version, ten components. However, the energy carried by gravitational waves, which are emitted according to general relativity by moving bodies, is not measured by this tensor, so that there is a problem in accounting for the conservation of mass-energy. This should not affect our ontology since our concept of energy can certainly be broad enough to include this form even if the physicists are not sure where it resides. The mutual dependencies of space-time and energy in general relativity apparently lead to the potential for indeterminacy (e.g. in black holes), although probably not in a way that would ever matter at the human scale (Penrose, 1989; Tipler, 1978).

\[16\] Mass and energy are equivalent and interconvertible, but converting all of the mass into the enormous amount of energy contained in it requires splitting the nucleus of the atom, since that is where most of the energy is stored.
6. When we try to move to contemporary quantum physics, the picture is anything but clear\textsuperscript{17} (especially to me, since I know very little about it), but, pending some discussions on this with physicists, I believe we can say that, since physicists now apparently think that the \textit{quantum field} is the core concept of reality, in the sense that the distribution and properties (mass, spin, charge) of particles are determined by variations in the field(s) (Feinberg, 1985; Lederman, 1993), we could in principle characterize the history of the universe as the energy value (quanta) for each of a finite number of quantum fields (corresponding to the elementary particle types) at different points in relativistic space-time. It follows from Heisenberg’s Uncertainty Principle that we cannot specify both precise field values and precise particle distributions in a given region of space, so we would have our choice in selecting what would represent information in a quantum universe.\textsuperscript{18} Heisenberg has thus forced us to give up on the idea of completely characterizing the universe in abstractly measurable terms, and, of course, this new universe is quite possibly not deterministic in any meaningful sense (Feinberg, 1985).

All of this is just an attempt to establish that we should be able to characterize reality as well as it can be characterized, given a particular physical theory, with the sets of the fourfold division. If resources are subsets of these sets, then at least we will not leave anything out by restricting ourselves to them. Also, because the sets are distinct (even if they are all to some extent interdependent, as in Einstein’s theory), we should be able to resolve confusions over what type of resource is being used.

\textbf{2.2.2 Category Necessity}

In order to establish that each category in the ontology constitutes a type of resource, let’s consider how each set is used by a form for its survival and multiplication.

\textit{Space.} As was mentioned before, occupying space is evolution’s definition of success. So for this reason alone space is necessary for survival, although it is not sufficient, since future survival is not guaranteed by present existence. Space is required for expansion of a form, which enhances its chances

\textsuperscript{17}The mathematics is apparently getting increasingly impossible even for the specialist to deal with; or, as Einstein might have said, “God does not play \textit{nice}.”

\textsuperscript{18}According to Feinberg (1985), measuring particle content precisely and averaging field value over a region of space is usually more useful than the reverse procedure.
of survival on the principle that, all other things being equal, large forms are harder to destroy than small forms. In higher forms, space in the neighborhood of the organism is essential for motility, if it is to move toward food or toward a mate. And space is necessary for computation in forms that use information about the world in order to enhance their evolutionary chances, substituting for time by making more parallel computation possible.

**Time.** Anything a form does to survive and multiply takes time, and in many cases the chances for success are increased as more time becomes available. More time implies a greater potential for computation that is temporally bounded, more and better information, greater ability to move, store energy, and expand into space, and greater compensation for spatial constraints via serializing computations.

**Energy.** All matter is energy, so a form’s existence as an entity with recognizable properties requires energy in the form of mass, and the more mass a form has the more resistant it will be to thermal breakdown and consumption by other forms. Free energy is required for computation, movement, and the restoration of form after dissipation via metabolism. Expansion and multiplication also require mass and free energy, and a greater availability of usable energy can compensate for a lack of information about and ability to obtain it elsewhere. Energy is also necessary to the formation of structures that are suited to growth, and more energy is required as the form becomes more complex.

**Information.** Information is useful to a form if the form has space, time, and energy in the structure required for exploiting the information. In animals, information about the locations of food, sunshine and mates, and about where to escape from prey, are all valuable for survival and reproduction. An organism’s genetic code also represents valuable information for ensuring the survival of the form. We can also speak of information being used by very simple forms, to the extent that the form’s structure tends to put it in places that help it to survive. The information is a resource in this case because if the facts were different then the form’s chances for survival would diminish or vanish. For example, beds rely on the fact that people like to sleep on them, just as we human beings rely on the fact that the earth is capable of sustaining us.

It should be apparent, then, that each of the four categories is indispensable as a type of resource. This does not mean that everything in each

---

19 i.e., computation that has a minimal time requirement even given maximal use of space
category is useful, and of course particular time, space, information, and energy may be useful to one form and not to another.

2.2.3 Limiting Factors

Before turning to a discussion of the naturalness of the fourfold division, I'd like now to consider a very important part of the proto-theory, namely, What are the limitations in these resources that restrict a form’s ability to survive and multiply? A resource can assist the growth of a form only if it is both useful and available to that form. In the last section, we tried to establish how the four categories can each be useful to a form. Now we turn to the factors that can affect the availability of each type of resource (FIGURE 4).

1. Space is limited by competition. Every form shares the space of the universe with other forms that are more or less difficult to consume, crowd out, or interpenetrate. Thankfully, for fans of diversity, so far no form has found a way to pervade over all space. In the absence of forms that could compete with a given form, this would not be the case, and the form would eventually have all the space it needed, although if the form was dissipative then at some point it might dwindle in material due to a lack of available energy. For each form at a particular time, the space available to it is limited by its enclosure and sometimes its immediate surroundings, and the growth of this space over time is limited by its own growth, which also depends on the other resources. Thus, while space has its own intrinsic limitation, its availability is also constrained by limitations in the other resources. An analogous statement holds for each of the categories in the fourfold division. We can represent the spatial limitation mathematically by means of a competition function with the origin at the form’s present location. Its value at each point in space is the probability that the point is occupied or unavailable to the form at a given instant over some time interval.

2. Time is limited by threats. The cumulative probability of death for a form increases with longer intervals, because threatening forms then have more time to mobilize. In a strict sense, time is limited to the lifetime

---

20 Physicists periodically worry about particle accelerators generating a self-sustaining reaction that would consume the earth. The most serious speculation that this could occur involved a hypothesized form of matter known as Lee-Wick matter that worried theorists in the 1960s before the operation of some actual accelerators confirmed that it did not exist (Ruthen, 1993).
of the form, but a form usually does not "know" how long it has, and of course this depends on what actions it takes to evade the threat. Some forms, of course, like the rock, are relatively incapable of taking actions to evade threats, and so must just accept whatever time is fated to be available to them. The crucial property of time is that its availability is always diminishing with respect to the threat; it is "evanescent." Time for a particular action that might enhance a form's growth is limited by the total time the form has, so that all such actions are limited in their use of time by purposes that compete with them within an agent or organism. But time is limited for virtually all physical realizations of form, not just for life, because it looks as if all of them eventually die, with the possible exception of the universe itself. The threat functions to which a form must respond change from moment to moment. At a given moment, a particular threat \( i \) can be represented as a logistic function describing the cumulative probability of the form's destruction given the form's present location. It is defined along a temporal axis \( t_i \), with the origin representing the present moment, the threshold parameter representing the critical moment when the threat is most likely to strike, and the slope parameter representing the suddenness of the onset.

3. **Energy is limited by thermal entropy.** The Second Law of Thermodynamics states that the total entropy \( S \), which is a measure of disorder or unavailable energy in a physical system and its surroundings, is always nondecreasing with time (Tinoco et al., 1978). If a form has constant entropy, which is true for the impossible ideal of a perfect crystal at absolute zero (by the Third Law), then there is no increase in its disorder and it can survive (but not multiply) without additional work. But for all real forms, the Second Law implies that if internal order is to be preserved (or increased, as when an organism is developing) then the form must either dissipate energy very slowly or, as in the case of living forms, export entropy to its surroundings so that its own entropy can be held constant or diminished, a fact first made clear by Erwin Schrödinger in an extraordinary series of lectures at Cambridge during World War II that seems to have set much of the agenda for modern biology (Schrödinger, 1944). This is necessary.

---

21I am using the term 'death' here differently from its normal use as the opposite of 'life'. Here it means the end of existence for a form, as when we speak of the 'heat death' of the universe or the extinction of a species. The end of an individual's life does not need to mean the end of their physical form, rather just that the individual reaches a state of thermal equilibrium with the environment and will gradually decay until we choose to say that the form no longer exists.
for survival because otherwise the form will degrade into a less ordered, and hence different, form. Of course, for some nonliving forms (like rocks) this will happen very slowly because their chemical structures are such that entropy within them increases slowly, and we will probably choose to say that the form maintains itself in spite of the gradual changes. Other forms, however, cannot rely solely on their internal available energy store for doing work, because this store will decline as entropy then rises (by conservation of energy, the First Law). Thus, a form like life that maintains its internal structure despite the fact that it is constantly dissipating available energy, must seek mass-energy in its surroundings that it can use to do the work required to maintain this order, by changing the mass-energy from a highly ordered form (lower entropy) to a less ordered form (higher entropy). In life this is known as metabolism, and it takes various forms (fermentation, photosynthesis, respiration, digestion). If the Second Law did not hold (as it might not in a contracting universe—see Lloyd, 1989), then it would be possible for an organism to maintain order without seeking outside energy. But since it does hold, it places a limitation on the available energy that a form can use for growth at a given time, namely to that which is in its immediate surroundings. And, of course, moving to obtain more energy is itself an energy-using (or, we might say, order-using) process internally, so the payoff must be worth the effort. Also, every action that does not result in metabolism has a cost, namely that it requires the additional expenditure of more energy to find a way to replace what has already been lost. The limited energy in a form’s “neighborhood” of the type that is required for growth can severely constrain what would otherwise be rapid expansion of a form. For example, a baker’s yeast would cover the earth in just two weeks if it had enough to eat! (Fernald, 1992) We can use the form of the free energy equations of Gibbs and Helmholtz to represent the available energy $A$ in the environment at a given temperature as the difference between the total energy and the temperature times entropy, neglecting details like pressure and volume constancy. Over time the growth in entropy can be plotted and compared to the maximal growth that could occur as a measure of order in the universe.

4. Information is limited by noise. A form can extract information about its surroundings only from the signals it receives at its periphery. We can construct an analogue of the Second Law of Thermodynamics for information by considering how much information is contained in a signal as

---

$^{22}$The form can also extract information about its internal state through apperception.
a function of the distance (in space or time) of its source from the observer. In the terms used by Claude Shannon (1949), we can state this new “law” as follows.

Along a given channel of communication, the uncertainty or likelihood of error for inferring the content of the original signal $x$ from the value received $y$ (the conditional entropy $H_y(x)$) is monotonically nondecreasing as a function of channel distance (in space or time).\(^{23}\)

To define what this means, we need to face a complication caused by changes in interference, for example increasing channel distance while moving out of the way of an obstacle and thereby increasing the accuracy of transmission. There are two ways of dealing with this. One way is to say that the channel has changed when we move out of the way. This is sometimes appropriate but is generally unsatisfying because it seems to say that a new channel is created every time there is a change in interference (or noise, to use Shannon’s term). The other way, which we will adopt, is to say simply that, given the same changes in interference along a channel, the longer transmission will have more associated uncertainty than the shorter. This is so because the longer transmission may pass through additional points of interference that the shorter one does not pass through. The finite speeds of light and sound can introduce additional factors to increase uncertainty when we try to infer what is happening at a given moment from a signal that has originated at a previous moment, but this effect is negligible compared to interference.\(^{24}\) In general, the rate of change in uncertainty along a channel as a function of spatiotemporal distance is a measure of noise at each location and instant. For uncertainty not to increase, the channel must be noiseless beyond some point. This is nearly true for modern digital communication channels over large distances, but for most organisms throughout history, including human beings, uncertainty has increased rapidly with distance. This is because of two factors: the diminished strength of the signal and the increased probability of occluding intermediaries.\(^{25}\) Across time, noise

---

\(^{23}\)Examples of channels across time include books and pictures.

\(^{24}\)Considerations of this kind could be used to construct a version of the law with strictly increasing uncertainty.

\(^{25}\)In the special case of vision in space, we can say that (1) the signal falls off as the cube of the distance, because the number of photons reaching us is proportional to the cube of distance, and (2) the probability of occlusion increases as the square of the distance because (a) occluding objects along a linear path are more probable as the distance and
comes from living in a dynamic environment, especially with respect to the things that threaten and compete with a form or agent. The increase in uncertainty with distance has the consequence of limiting what an agent can know out of the total information in the universe. In fact, the vastness of space and time, combined with the usually rapid increase in our ignorance or uncertainty as we move away from ourselves at the present moment, implies that the information we are limited to is a tiny subset of the total. For particular channels, we can define a noise function in which the signal to noise ratio declines as a function of distance from the present location.

Because of the interdependencies between the four resources, they are reciprocally limiting. Thus, limited spatiotemporal extent limits the amount of energy and information that are available to the form, and limits on its available energy and information in turn limit the form's spatio-temporal extent. Also, imposing additional limitations on one resource increases the burden on the others.

2.2.4 A Natural Division

The final property we seek for our ontology, in addition to completeness, distinctness of the categories, and category necessity, is naturalness (or usefulness). For the most part, I will let the consequences of the fourfold division speak for themselves in this regard (see subsection 2.4), and the reader can decide whether the psychological effects differentiated by this ontology constitute natural kinds. However, I do want to establish a correspondence between the resources I have identified and the complexity measures arising in the theories of computation and of information, because these are directly concerned with measuring required resources.

A basic Turing machine (FIGURE 5) consists of a finite control that, more or less intelligently, moves its head (arrow) both left and right along a tape consisting of a finite number of inputs \( (a_1, a_2, \ldots, a_n) \) followed by an infinite number of blank spaces (denoted \( B \)), printing symbols based on what it has read. The finite control represents the program that it follows in doing this. The Turing machine is a favorite device in computational theory because, depending on how you construct the finite control, it can compute anything that has been shown to be computable by other machines. Computational complexity theory focuses on two measures: (1) the minimum number of steps that a Turing machine must go through in order to solve a

(b) the required size for two-dimensional (retinal) occlusion falls off as the square of the distance.
problem, as a function of the problem's size (e.g., the number of input symbols), and (2) the minimum length of tape that the machine must have in order to solve a problem, also as a function of the number of inputs. These two measures are $T(n)$ (the temporal complexity) and $S(n)$ (the spatial complexity) of a given problem.

But we can also specify, for any given problem, two other types of complexity. One is the minimum number of inputs required in order to solve the problem unambiguously, and the other is the minimum length of the program that would be required for a universal Turing machine (a Turing machine that is completely programmable by its inputs) to simulate the finite control. Complexity measures of these two types have been studied by information theorists. The first type is usually characterized by the Shannon entropy $H$, and the second is commonly measured by the Kolmogorov complexity $K$.

There are additional complexity measures that have been studied in information theory, such as "logical depth" (or the number of steps required given the shortest program for solving a problem) (Bennett, 1987) and other "resource-bounded" complexity measures which tell how much of a resource is required to solve a problem given bounds on the availability of other resources (Li and Vitanyi, 1990), but the four measures that correspond to components of the Turing machine are the basic types in terms of which other measures are defined. Moreover, these four complexity measures exhaustively characterize the resources required for any problem that is computable by a Turing machine, because each part of the Turing machine is covered by one of the measures.

Now I want to argue that there is a correspondence between the measures and the resources of my fourfold division. Specifically, $S(n)$ and $T(n)$ measure space and time, Shannon entropy measures the information that must be given to a structure, program, person, etc., to solve the problem, and Kolmogorov complexity measures the required complexity of the controlling structure itself. If we make a bit of a leap from "measure of complexity of the controlling structure" to "measure of thermodynamic order for a physical form," then this last complexity measure will correspond, in its more physical version, to the quantity of (ordered) energy that must be available to the form in order to solve the problem. In fact, characterizing physical order in just this way has been the goal of a new complexity measure introduced by Heinz Pagels and Seth Lloyd (1988). They call the measure "thermodynamic depth," and it attempts to define how hard it is to assemble a structure as "the amount of information required to specify the trajectory
that the system has followed to its present state." We could think of this as the Kolmogorov complexity of specifying this trajectory. The authors define the thermodynamic version of depth as the difference between the coarse- and fine-grained thermal entropy of a system, and they claim it is an objective quantity, but I do not understand it well enough to explain what they mean. I will study the paper on thermodynamic depth some more, but others interested in information complexity, who have spent much more time analyzing this notion than I have, have also had difficulty intuiting it, and thermodynamic depth has apparently met with mixed reviews as a measure of physical structure.\(^\text{26}\) Nonetheless, the idea of trying to characterize the structure in nature in complexity terms seems sound, and others are working on how to do it better. In fact, we might class this work with that of a fair number of physicists, biologists, chemists, and philosophers who have been working on integrated theories of evolution, thermodynamics, and information (e.g. Weber, Depew, and Smith (eds.), 1988; Brooks and Wiley, 1988; Wicken, 1987; Kuppers, 1990; Bonner, 1988; Locker (ed.), 1973; Denbigh and Denbigh, 1985; Pagels, 1983; Prigogine and Stengers, 1984; and Layzer, 1990).

The important point for us, I think, is that, definitely in the case of time, space, and information, and with some handwaving about reducing program complexity to thermodynamic order in the case of energy, we have essentially the same categories of resources being measured in complexity and information theory as we have proposed in the fourfold division. Because the complexity measures are the prevailing way of talking about resource requirements in the information sciences, it seems to me that we have a \textit{prima facie} argument that the fourfold division is a natural one for talking about problems at the physical level as well.

2.3 The Responsive Principles

With the physical basis thus established, we come now to the core of the proto-theory: the principles that govern how a living form responds to the factors that limit its resources. At this point I will mostly drop the general talk of "forms" and focus on living organisms I will call "creatures."

We begin with a topmost principle, laying out the dimensions along which evolution selects.

\textbf{The Fitness Principle.} Creatures tend to survive and multiply

\(^{26}\)This is according to conversations with Jose Mosterin and Andrew Nobel.
to the extent that their use of resources promotes reproduction in a way that is responsive to (1) the prevalence of competition, (2) the urgency of threats, (3) the scarcity of available energy (level of entropy), and (4) the noise in the environment (i.e., how much uncertainty increases with spatiotemporal distance).

This is a consequence of the argument we have given so far, which tried to establish that the only things limiting a form’s evolution are constraints on its available resources, and that these constraints are completely characterized by the factors (1) through (4). The term ‘responsive’ is highlighted because it implies that a creature can use resources for its survival within the limitations imposed on the resources. From the viewpoint of optimization, the use of resources should maximize reproduction subject to the limiting factors as constraints.

From what we have established so far, we can spell out four principles by which a creature can respond to the limited availability of resources. I will call these four the responsive principles. They can be viewed as defining the optimal use of a resource, but in each case there is an implied dual principle, namely that the loosening of a constraint (for instance, the decline of threats) frees the creature from having to restrict its use of the resource in the manner prescribed by the principle. This point will be important when we consider the reasons for observed suboptimality. In terms of Geoffrey Miller’s graph (FIGURE 1), these four principles would lie just below the top, fitness principle.

The responsive principles all have a general form, which I claim is the correct one given that we are concerned with creatures that have survived over some period. This schematic form is

\[
\text{Responsive Principle. To the extent that the limiting factor reduces the availability of the resource in a creature’s environment, the creature will tend to reduce its use of the resource toward the minimal envelope.}
\]

I defined the limiting factors for each resource earlier. This schema introduces the concept of the minimal envelope. The minimal envelope is that portion of a resource whose availability is implied by a creature’s existence and survival.

Let’s consider each resource in turn. The first responsive principle is

\[
\text{The Principle of Internality. To the extent that competition limits the availability of space in a creature’s environment, the}
\]
creature will tend to reduce its use of space toward that which the creature itself occupies.

This principle simply says that if a creature cannot use space in its surroundings, then it must use its internal space, in particular that within its own brain. For example, if an animal cannot rely on the markings it makes in its surroundings, it will have to remember where things are or else lose whatever capability this facilitates. Since there are other limitations on the creature’s use of space owing to the other resource constraints, the use of internal space is also a way of overcoming those other constraints. This will be true of each of the principles, even though we will not write it explicitly into each principle. For example, the use of internal space is a way of responding to the limiting factor on information, which is noise produced by distance, because it is too hard to use distant space to store information due to the uncertainty of being able to retrieve it. If I make a mark on a piece of paper in Calcutta, it will be very difficult to see it without great effort and a lot of modern technology. But if I store this information in my head, then I can carry it around with me and use it whenever I want. Internality, then, helps to overcome limitations in the other resources even as it imposes a limitation on the use of space. This is another instance of the interdependence of resources that was mentioned previously. The principle of internality is becoming less important for human beings as we rely more on the external environment to store information for us, through language (which allows other people to be sources of information), books, computers, and other media, because in effect we have been successful in competing with other forms for the control of this external space. It is conceivable that this may eventually lead to a shrinking of the human brain in our future evolution, but of course new functions could emerge to use up the space that was once used for memory.

The second responsive principle is

The Principle of Automaticity. To the extent that threats reduce the availability of time in the creature’s environment, the creature will tend to reduce its use of time toward that which is required for performing evasive actions.

This principle says that the creature will benefit from minimizing its use of time, or at least acting within the time it has prior to a threat. This seems obvious, but it gives rise to some interesting strategies, as we will see later on. Intuitively, time is often more scarce than space, so a creature
is for that reason driven to internalize more information so that it can be accessed quickly, rather than using time to do computation or to seek more information. Like internality, automaticity is less important than it used to be for human beings, since we have managed to reduce the threats in our environment to a manageable number (except in some parts of our cities).

The third responsive principle is

**The Principle of Stability.** To the extent that thermal entropy reduces the availability of energy in a creature's environment, the creature will tend to reduce its use of energy toward that which is required for maintaining its physical structure.

I think that principle requires a lengthier explanation than the others. Survival places on a physical form the burden of finding and maintaining a stable structure in the face of the Second Law of Thermodynamics, which is the limiting factor in chemical evolution. In saying this, we are merely recognizing that the very concept of a "form that endures" is, by definition, a stable structure. Indeed, the kinds of things that we could have a theory about are likely to be biased toward stability of some sort—otherwise there would be no well-defined object to analyze. If we consider the space of possible objects to be the set of all spatiotemporal regions (after all, each of these "objects" could be said to have "survived" in evolution), then the set of stable systems within the group is an extremely small subset and is roughly coincident with the set of recognizable objects. So stability of some sort seems to be required if we are to analyze an object at all. Of course, we have theories about gases, nonequilibrium systems, and chaotic systems, but even these must have some properties (like the principles of the kinetic theory, or the presence of strange attractors) that are stable across the class we are analyzing, or we would not really be learning anything.

In biological evolution, the forms that we study have a great deal of physically stable structure, and the complexity of any given form owes itself largely to the ability of its predecessors to maintain increasingly complex stable structures. The structures we observe have been selected for their stability, which, when we realize that this may be the only reason we wish to preserve ourselves, becomes a point almost too deep to fathom. A consequence is that complex forms that have tended to survive tend to be biased toward initial conditions. The reason is that, as a form becomes more complex, the energy landscape required for stability tends toward more local minima, an insight from connectionism among other fields. Thus, equi-

---

27This is actually a very tentative statement, but I think it is true. I will be attempting
librium states tend to represent states closer to the initial state than is the global minimum, and the presence of local equilibria can make it difficult to escape from the stable well.

Biological systems represent what we might call a triumph of stability over equilibrium, because an organism is never in equilibrium (until it is dead) and yet it manages to maintain stability. In fact, we can distinguish four broad classes of stability that have evolved since the dawn of time. The first kind to evolve was what we might broadly call physiochemical stability, incorporating a large number of types that include stable particles, nonradioactive isotopes, crystals, and gases and liquids in macroscopic equilibrium. Rocks are our paradigm for this type of stability, and it is very effective as an evolutionary strategy. Many organic compounds are not particularly stable in this sense, but they have an incredible flexibility as a class that has allowed them to develop the other forms of stability we associate with life. Probably the next form of stability to evolve was regulative stability, or the ability to metabolize.28 This corresponds to “autonomy in the face of external perturbations” (Wicken, 1987) and is maintained by homeostasis, or the tendency to return to a reference state when, for example, temperature or chemical balance departs from this state. This is why we crave salt when we haven’t had any, why we yearn for a heater on a cold night, but air conditioning on a hot one, and probably why we sleep. Human beings require, in addition to the air, water, and heat, about 1500 nutritionist’s calories a day of energy in order to maintain ourselves in a state of rest (the minimal envelope for human digestive energy), compared to 2500 calories for normal activity (Tinoco et al., 1978), so we can see the benefits for energy conservation in the more stable state of rest. Homeostasis is necessary for preservation of the species because it serves to maintain the right balance of molecules for reproduction, which is the third form: replicative stability. Replication serves to preserve a life form after its metabolic structure has died. The fourth, and most recently evolved, form of stability I want to call anticipatory stability. Whereas metabolism is a purely responsive mechanism, anticipatory life forms (see Rosen, 1985, for a related concept) act to preserve stability even before it is perturbed, for example by avoiding threats to stability, or stocking up on food.

to verify how rigorously this and a related statement about (see section 2.4 on modularity) has or can be established.

28Some biologists believe in what Freeman Dyson calls the “dual origin hypothesis,” meaning that metabolism and replication arose at the same time, but he provides a good argument that they evolved separately and that metabolism evolved first (Dyson, 1985).
I think that the propensity for anticipatory stability can explain a great deal about human behavior, realizing that it is a bias rather than an absolute tendency. In particular, it gives rise to a phenomenon we might call *change aversion*,\(^{29}\) which is closely related to loss aversion, as well as security-seeking. This is evident in the cost we attach to change itself, even when the change is beneficial. We anticipate instability and therefore have some desire to avoid it. This can lead to things like a hardening of one’s attitudes, not probably because of the relatively small amount of energy required to percolate these effects through the brain, but because a change in attitude may require us to change the way we live, and that can require tremendous energy in effecting the transition. A good deal, perhaps all, of our emotional life can be explained as an effect of seeking stability of one form or another. Why do we desire to have a great deal of money? Perhaps because we anticipate the greater stability associated with it, the lesser requirement that we expend energy (living a life of leisure, etc.) This last point is related to a more basic one about stability, that we are biased toward idleness when available energy is scarce. The sloth is a good example of this, as are humans who are tired. All of these are introspective observations, of course. I will try to tie the principle to experimental prediction and phenomena in the last part of the talk.

The fourth, and final, responsive principle is

**The Principle of Locality.** To the extent that noise reduces the availability of information in a creature’s environment, the creature will tend to reduce its use of information toward that which is contained in signals appearing at its present location.

This principle says simply that the creature is informationally biased toward what is immediately present. There are three aspects or consequences of the locality principle that I would like to discuss in order to set up the next section of the talk on psychological effects.

First, it follows as a consequence of the increase in noise with distance that we (speaking now of “us” creatures) make less use of information originating at some spatial or temporal distance.

Second, locality, to the extent that it holds, constrains the measuring of quantities in received signals to occur via comparison with quantities that are present at the moment and location of reception. For example, we can

\(^{29}\)This has been referred to in the literature as a “status quo bias” (Samuelson and Zeckhauser, 1988; Hartman, Doane, and Woo, 1991).
compare the intensities of neighboring regions in the receptive field, we can
correlate new information with whatever is present in our minds, or we can
observe changes in quantities received from moment to moment, but, to the
extent that locality holds, we cannot make comparisons to quantities that
lie outside the local zone. This has the effect of confining us to bounded
log-interval scales in which, like a pan balance, we can measure ratios but
not differences. This leads to the familiar diminishing sensitivity inherent
in Weber's Law: the difference between quantities that is required in order
for us just to notice it is proportional to the absolute magnitudes involved
rather than being a constant that is insensitive to those quantities. Signals
whose ratio falls below Weber's proportionality constant are like those that
tip the pan balance to its limit. Of course, our minds contain abundant
information that might enable us to measure quantities on more refined
scales that do not imply diminishing sensitivity. For example, we can, at
least linguistically, distinguish the difference between ten pounds and eleven
pounds just as easily as between zero and one pound, but this represents a
departure from the conditions required for locality to hold in toto because
there is information that is carried across time analogous to the numbers on
a bathroom scale. To measure magnitudes in this way, as opposed to being
merely affected proportionally by incoming quantities, we must have a way
to equate quantities at different points in time to which the scale has been
applied, so that for example someone in the past had to place the numbers
on the bathroom scale to indicate how much it was displaced by different
weights. This is a subtle set of issues that I have not really mastered, and
I am planning to read a paper by Duncan Luce that should improve my
understanding. But it would seem that to the extent that we move toward
linear or even superlinear sensitivity to quantities, we are making use of
information that is either in memory or that has been built up in our brains
by evolution. Locality, like all the principles, is a bias that may hold to
greater or lesser extent.

The third and most profound implication of locality is that we are much
more apt to use or be affected by information that lies in, to paraphrase
Vygotsky, our zone of proximal experience, or as Lewin would say, the field.
We are more influenced (on average) by information we are receiving here
and now or have experienced recently, less influenced by information we
receive from afar or experienced long ago, and less still by facts that we
would have to infer or imagine about people and places that lie outside our
direct experience. For my money, this is the most profound fact about our
cognitive lives.
Locality partly compensates for stability and internality, because while those principles emphasize the storage and preservation of memory across time and space, locality emphasizes a responsiveness to present experience.

These, then, are the core principles of the proto-theory: internality, automaticity, stability, locality, and the topmost principle, fitness. From them, I have said that we should be able derive psychological effects. Some have already been mentioned. But there is a basic set of consequences to which I would now like to turn.

2.4 Consequences for Information Processing

At this point the proto-theory will begin to look even more “proto-ish,” the justifications become more oblique, and there is clearly a good deal of work to be done. But by now I am also beginning to feel, like George Miller, that I am persecuted by an integer. This time the magic number is seven minus three, as we establish additional correspondences between the fourfold division and its responsive principles, on the one hand, and the properties of logical systems, information processing, and psychology, on the other.

When we are considering the representation of information and procedures for reasoning with it, the pared down theory of this that I like to consider first is formal logic. In logic, perhaps not so coincidentally, there are four properties that we typically seek for sets of axioms, and by extension, I want to argue for other types of representational and reasoning systems. In order of importance for logical systems, they are: consistency, soundness, independence, and completeness. Each of these has both a semantic and a syntactic (or proof theoretic) interpretation, and since I want to consider how resources constrain representation and reasoning within the mind itself, I want just to consider the syntactic versions of these concepts. Consistency is an absolute requirement in logic, because inconsistency implies that you can prove anything and hence your logic is useless. Soundness, the property that statements in the logic are true, is obviously important, but the logic will not collapse if you have a false statement. However the syntactic requirement for soundness is validity (that everything derivable from the axioms should indeed follow from them), which begins to seem very important indeed for a logical system. Logical independence is kind of a luxury, conforming to the aesthetic of minimality or simplicity. And completeness, the property that all truths are represented, is impossible for a finite axiom system incorporating the domain of arithmetic, as Gödel showed, unless of course the system is inconsistent. The syntactic portion of completeness,
to which I will restrict consideration, is decidability: the property that any proposition in the logic can be either proven or disproven from the axioms.

When we soften the hard edges of logic into a more flexible representational system allowing for uncertainty, we can speak of properties analogous to the logical ones. The new properties (and their logical analogues) are modularity (minimality/independence), vividness (completeness/decidability), harmony (consistency/satisfiability), and likelihood (soundness/validity). Each of the logical properties now becomes something that can hold to a greater or lesser degree without the entire system collapsing, as it would in the case of an inconsistent set of axioms. As I said before, the waters are getting a bit murkier in the sense that this list of terms hardly sounds like a standard one. But I have chosen the terms to draw attention to the fact that each of them has some theoretical basis on which we can draw for making these notions more precise and for exploring their consequences. Modularity and vividness are aspects of representation that have been studied in artificial intelligence. Harmony is a concept from connectionism, for which a theory has been developed, and likelihood is the object of numerous statistical theories, including theories about how to extract information from signals.

Now, if you hadn’t already guessed, I am going to claim that there is a correspondence between these four properties and the resources of the fourfold division with their responsive principles. Specifically, I think that spatial limitations and internality push the system toward modularity, that time pressure and automaticity drive it toward vividness, that entropy and stability put a premium on harmony, and that noise and locality put a premium on likelihood. Let’s consider how all this follows and try to infer some consequences for psychology. We will now speak not just of animals and creatures but of brains and human cognition.

*Modularity* is representation of information in a form that restricts dependencies. In networks, modularity is at its height when there are no connections between any of the elements: each one operates completely independently. This is obviously not workable for an inferential system that needs to combine information. Complete absence of modularity occurs when all of the elements are connected directly with all the others, as in the case of a fully connected network with nth order connections, where n is the number of elements. This, and even full binary connectivity, are not feasible for networks of any size. So the brain, for example, gets by with an intermediate degree of modularity, with neurons being remote from the influence of some and closely connected to others. For linear increases in the number of neurons, and hence in the volume of the brain, the volume of fibers required
for full interconnection would grow as the square, so that some degree of modularity seems to be called for by the limits of space within the brain. When we wish to represent all the dependencies that are present between concepts, minimal modularity corresponds, on this semantic criterion, to data compression. For example, when a dependency between A and C can be *computed* from the dependencies between (A and B) and (B and C), the third connection would be redundant. If it is present, it may take less time to update C given A than if we had to go the two steps through B. This illustrates a tradeoff, the fundamental one between space and time in computation, namely that more redundant representations save time but require more space, whereas compressed representations save space but require more time. Now the key insight in moving from the presence of modularity in the brain to its associated psychological effect is provided by a result from connectionism: greater modularity means an increased sensitivity to initial conditions in determining what a neural network can access, because of the increased number of local minima in the dynamic landscape.\(^{30}\) The effect that we see is context sensitivity, and its examples in cognitive psychology are abundant: encoding specificity, *Einstellung*, and problem-specific script behavior. The judgmental heuristic that, I claim, corresponds to this limitation is accessibility (Tulving and Pearlstone, 1966).

**Vividness** is a concept that was introduced into the technical lexicon in the 1980s by the artificial intelligence researcher Hector Levesque (Levesque, 1986). It means that information is maximally represented in a way that is explicit, so that the computational time required to derive consequences implicit in information can be avoided. In contrast to minimality, it tends to lead to redundant representations from an informational standpoint. Representations of this kind have been a topic of study by designers of knowledge-based systems as a way of avoiding the so-called “frame problem,” or the problem of newly deriving all the consequent changes and non-changes when a fact becomes known or an event happens in the world. The solution has been to choose representations that have what is called the finite model property, in which a complete (i.e. vivid) model of the situation is represented and changes percolate throughout the model. The property incorporates a closed-world assumption, or the assumption that the representation will choose an action or inference in every situation and hence will not “hang” indefinitely. This is quite consistent with automaticity, in which the goal is to react with maximal speed. If the brain keeps a model of the world

\(^{30}\)I need to verify this.
and is updating it according to the changes it notices between data and the model, then it can respond quickly to such changes, but it must also engage in the procedure of "completion" or filling-in, i.e. going beyond the information available in constructing the model so that it will know when new information is surprising and needs to be added. This, along with the context sensitivity given by spatial limitations, is one of the properties of schemas that is observed in experiments. The judgmental heuristic that roughly results from these effects is representativeness.

Harmony is the term introduced, also in the 80s, by Paul Smolensky to describe the tendency of connectionist systems that corresponds to energy minimization (Smolensky, 1986). Scarcity of energy implies that changes in the representation will have some cost because new computation must be performed in order to propagate the effects of the change. This fact leads the brain to seek an understanding of the world that is in "informational equilibrium" (Jerry Fodor's term—Fodor, 1993) in the sense that it is relatively impervious to new information, and that is in reflective equilibrium in the sense that it will not change with further thinking, and the model should have harmony or consonance. When a change is incompatible with the model that a mind has of the world, energy and the stability principle lead in the direction of discounting the importance of the change because of its computational costs. These observations imply the general phenomenon of, to use Lee Ross's term, biased assimilation of new information, meaning that the information can have less impact than it ought to when the cognitive dissonance that would be created by its full assimilation is high. I have already indicated how important I think stability is for emotion, and it seems that harmony-seeking in the service of stability is responsible for many experimental effects, including especially primacy effects. As with Einstellung it appears that both internality and stability may be partly responsible for primacy, so that we can begin to see how effects can be dependent on more than one of the resources. The judgmental heuristic that corresponds to these effects, I think, is the bias of anchoring.

Likelihood is simply the property of a representation that makes it a reliable carrier of information about the real world. Just as temporal considerations run counter to spatial ones, informational constraints lead in the opposite direction of energy limitations. Uncertainty about what is out there in the distance leads the brain, at least at the periphery to try to make the most use it can of the information it has in front of it. This leads to a what in signal processing is often called “enhancement” or “restoration,” in which an image or signal is filtered so as to recover information and discard
noise. In our everyday awareness, violations of expectation and changes relative to the current model are focused upon and highlighted for the possible information they carry. The psychological effects that bear all this out are those associated with contrast enhancement. The diminishing sensitivity and ratio comparison associated with locality lead us to be biased in favor of high-contrast or salient information. I think that the closest judgmental heuristic to this is availability, understood in a somewhat narrower sense that distinguishes it from accessibility.

Thus, I have argued for a correspondence between the fourfold division and a set of what seem to be primary psychological effects: context sensitivity, filling-in, biased assimilation, and contrast enhancement. The overall pattern of cognition that this suggests is what we might call “model-based assimilation,” in which the mind carries around a model that it deems appropriate to given classes of situations and responds to new information according to how much strain it puts on the model.

We can now construct a table showing all of the psychological correspondences we have established (though not yet rigorously) for the four resources (FIGURE 6).

From the standpoint of information processing, we can also draw an interesting graph (FIGURE 7) that shows the tradeoffs between space and time along one axis, and between energy and information on the other axis. If the bare signal is the origin, the vertical axis (coding) expresses the apparent fact that for encoding the signal in the brain, temporal limitations lead toward greater redundancy and spatial limitations lead toward compression. The horizontal axis (filtering, or processing) expresses that informational limitations tend to lead in the direction of enhancement or recovery of information, whereas energy limitations lead in the opposite direction, toward destroying or discarding information that would interfere with stability.

2.5 The Arrow of History

Having moved from the beginning of the universe through physics, chemistry, biology, and psychology, we now set our sights on a view of history, or at least the evolutionary history of human beings. A possible objection to the four responsive principles as explanations for human biases is that the errors we observe in psychology are seldom explainable solely in terms of limitations on available resources. The brain is more than large enough to apply simple rules of probability, for instance, or logic, and in some cases (for example mental accounting) it seems to do more computation than
is necessary. How can we reconcile the violations of rationality that we observe with the abundances of resources that now seems available. Do human beings have any legitimate excuse?

The answer I want to give is one I have already mentioned as the thesis statement of this paper. It is that the brain has evolved from a succession of earlier forms that were smaller, less complex, and therefore more responsive to the limiting factors on resources. The structures they built are the ones on which our cognitive architecture is overlaid, and because the limitations we have noted have applied consistently and from the beginning, they have had a tremendous constraining influence on evolution. This is a vestigial theory of the biases, arguing that they are the shadows of physical limitations that applied at an earlier stage of evolution, indeed traceable to constraints on available quantities that apply even to the bedrock of the earth. But my sense is that we are gradually overcoming our past, growing into the larger brains and greater control that we have achieved, so that there is reason to hope that we may loosen the chains imposed on us by the four responsive biases. If this view is right, we should be moving from stability toward greater flexibility/adaptiveness, from locality toward more knowledgeable judgments, from internality toward more communication, and from automaticity toward more deliberation and consciousness, but how fast this will happen is obviously very difficult to say.

2.6 Some Hypotheses

The evolutionary thesis, which may be amenable to demonstration via computer-simulations of natural selection for problem solving architectures, gives rise to a hypothesis that is one of several I would like to put forth and find ways to test. The basic empirical hypotheses that seem to me to grow out of the proto-theory are as follows. (Each of them is almost surely not true in general, but may be more or less true for particular behaviors and may tend to be true overall).

*Evolution.* The biases will diminish for a form over the course of its evolutionary history.

*Development.* The biases will diminish for an individual over the life span.

*Adaptation.* The biases will diminish for an individual as the constraints on resources are loosened.
Submergence. Human beings tend to assume smaller effects of the bias than result from earlier phases in evolutionary history and development, and from tighter constraints on resources.

Each of these should be viewed as the schema for many more specialized hypotheses. The experimental work that I have paid the most attention to relates to the adaptation hypothesis, and as I think about it, it seems to me that one of the best ways to test this, as well as perhaps the others, is with rats. I like the thought, because after a year of theorizing I feel ready to become a good old-fashioned experimentalist.

3 Bibliography


Fernald, R. Evolution and Behavior. Lecture delivered at Stanford University, April 2, 1992.

Fodor, J. Lecture delivered at Stanford University, February 9, 1993.


40


Figure 1. Geoffrey Miller’s Graph
Fig. 2. The Hourglass
Figure 3. A Picture of the History of the Universe
Space is limited by competition.

Time is limited by threats.

Energy is limited by entropy.

Information is limited by noise.
Figure 5: A Basic Turing Machine
(form from Hopcroft and Ullman, 1979)
<table>
<thead>
<tr>
<th></th>
<th>Likelihood</th>
<th>Locality</th>
<th>Understanding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conflict</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Argument</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>bias vs. assertion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>representation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>filling-in</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>process-skill</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Context-sensitivity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Psychological effects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Consequence</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Intra-</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Inter-</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Resource</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Space**

**Time**

**Stability**

**Vividness**

**Harmony**

**Competence**

**Responsibility**

**Principal**

**Resource**