CHAPTER 10 Epistemology, Access, and Computational Models

Introduction: The Epistemological Question

The study of epistemology considers how the human agent knows itself and its world, and in particular whether this agent/world interaction can be considered as an object of scientific study. The empiricist and rationalist traditions have offered their specific answers to this question. I propose a constructivist, model-refinement approach to epistemological issues and offer a Bayesian characterization of agent-world interactions. I present several Bayesian-based models for diagnostic reasoning and point out epistemological aspects of this approach. I conclude this chapter with some discussion of possible cognitive correlates of this class of computational model.

After finishing a PhD at the University of Pennsylvania in 1973, I accepted a Postdoctoral Research Fellowship at the University of Edinburgh. My research at Penn had focused on models for human problem solving in the spirit of Allen Newell and Herbert Simon (1972, Goldin and Luger 1975). There was at that time in the Psychology, Linguistics, Epistemics, and Artificial Intelligence Departments at the University of Edinburgh a wealth of cognitive tasks under analysis, as well as an exciting research community directing these efforts. These projects included the study of seriation skills in primates and young children (Young 1976, McGonigle and Chalmers 1977), object permanence studies with infants (Bower 1977, Luger et al. 1983, 1984), algebra problem solving by adults (Luger 1981, Bundy 1983), the effects of problem structure on problem solving behavior (Luger1976, Luger and Bauer 1978), and even the understanding of perception in moving environments. Besides the normal university teaching and research activities, the fairly regular interdisciplinary seminars in Stewart's Conference Room on Drummond Street gave us the opportunity to understand each others' projects and research, as well as to map out much of the common intellectual territory that supported and mediated our research goals.

In a foundational sense, many of our common problems were epistemological. Whether human or primate, how does an agent work within and manipulate elements of a world that is external to, or more simply, is *not*, that agent? And consequently, how can the human agent address the epistemological integration of the agent and its environment? And how is the human agent able to characterize this integration? The roots of these epistemological issues predate Greek philosophy, and across the centuries have had many articulate proponents, including Aristotle, Plato, and in more modern times, Descartes, Locke, and the Scots philosopher David Hume, near whose memorial building at the University of Edinburgh many of our experiments and discussions took place.

Epistemological access, or the question of whether/how an agent can understand its own understanding, became a key question for research. In our science at Edinburgh we attempted to assess both how our subjects understood and dealt with their world as well as how we ourselves could understand and characterize that subject-environment interaction.

In fact, epistemological access also addresses the foundations for bias and ignorance in ourselves and society, as well as provides clear directives for understanding and creating science itself. Bias and ignorance often follow a lack of knowledge and the subsequent misunderstanding of situations. How can an agent appreciate that its judgments about a situation are simply incorrect? Aberrations often occur because agents do not have the intellectual courage and/or confidence to deconstruct a situation or the accumulated knowledge and commitment to make well-founded judgments. A common example of this is the proposition that creationism, intelligent design, and selection-based evolution have equal explanatory power in understanding the state of the natural world.

This chapter offers the author's rapprochement with the issue of epistemology and addresses the deeper problem of epistemological access. The following section takes a philosophical stance, finding in constructivism a plausible integration of the empiricist and rationalist views of the world. The third section, utilizing the insights of Bayes' theorem, offers a computational model able to integrate prior expectations with posterior perceptions within the phenomenal world. The fourth section demonstrates this integration with several examples of diagnostic reasoning. I conclude the chapter with more speculative comments on how schema-based diagnosis and model-refinement might be instantiated in the human cortex.

An Integration of Rationalism and Empiricism

Over the past sixty years work in artificial intelligence can be understood as an ongoing dialectic between the empiricist and rationalist traditions in philosophy and epistemology. It is only natural that a discipline that as its focus engages in the design and building of artifacts that are intended to capture intelligent activity should intersect with philosophy, and in particular, with epistemology. I describe this intersection of disciplines in due course, but first we consider philosophy itself.

Perhaps the most influential rationalist philosopher was Rene Descartes (1680), a central figure in the development of modern concepts of the origins of thought and theories of mind. Descartes attempted to find a basis for understanding himself and the world purely through introspection and reflection. Descartes (1680) systematically rejected the validity of the input of his senses and even questioned whether his perception of the physical world was "trustworthy". Descartes was left with only the reality of thought: even the reality of his own physical existence had to be reestablished. His physical self was established only after making his fundamental assumption: "Cogito ergo sum". Establishing his own existence as a thinking entity, Descartes inferred the existence of a God as an essential creator and sustainer. Finally, the reality of the physical universe was the necessary creation and reflected the veridical trust in this benign God.

Descartes' mind/body dualism was an excellent support for his later creation of mathematical systems including analytic geometry where mathematical relationships could provide the constraints for characterizing the physical world. It was a natural next step for Newton to describe the orbits of planets around the sun in the language of elliptical relationships of distances and masses. Descartes clear and distinct ideas themselves became a sine qua non for understanding and describing "the real". His physical (res extensa) non-physical (res cogitans) dualism supports the body/soul or mind/matter biases of much of modern life, literature, and religion.

The origins of many of Descartes' ideas can be traced back at least to Plato. The epistemology of Plato supposed that as humans experience life through space and time we gradually came to understand the pure forms of real life separated from material constraints. In his philosophy of reincarnation, the human soul is made to forget its knowledge of truth and perfect forms as it is reborn into a new existence. As life progresses the human, through experience, gradually comes to remember the pure forms of the disembodied life: learning is remembering. In his cave experience, in the final book of The Republic, Plato introduces his reader to these pure forms, the perfect sphere, beauty, and truth. Mind/body dualism is a very attractive exercise in abstraction. Especially for agents confined to a physical embodiment and limited by senses that can mislead, confuse, and even fail.

The empiricist tradition, espoused by Locke, Berkeley, and Hume, distrusting the abstractions of the rational agent, reminds us that nothing comes into the mind or to understanding except by passing through the sense organs of the agent. On this approach the perfect sphere or *absolute truth* simply do not exist. What the human agent "perceives" are the things of a physical existence; what it "knows" are loose associations of these physical stimuli. The extremes of this tradition, expressed through the Scots philosopher David Hume, include a denial of causality and the very existence of an all-powerful God. There is an important distinction here, the foundation of an agnostic position: it is not that a God can't exist, it is that the human agent can't know or prove that he/she *does* exist.

Aristotle was one of the first proponents of the empiricist tradition, although his philosophy also contained the ideas of "form" and abstraction from a purely material existence. For Aristotle the most fascinating aspect of nature was change. In his Physics, he defines his "philosophy of nature" as the "study of things that change". He distinguishes the *matter* from the *form* of things: a sculpture might be "understood" as the material bronze taking on the form of a specific human. Change occurs when the bronze takes on another form. This matter/form distinction supports the modern computer scientists' notions of symbolic computing and data abstraction, where sets of symbols can represent entities in a world and abstract relationships can describe these entities sharing common characteristics. Abstracting form from a particular material existence supports computation, the manipulation of abstractions, as well as theories for data structures and languages as symbol-based representations.

The modern empiricist has a deep dilemma, however: how can a human come to know/understand new relationships in the physical world? This issue was described more than two thousand years ago by the slave Meno in the Platonic dialog bearing his name (Plato 1961):

> And how can you enquire, Socrates, into that which you do not already know? What will you put forth as the subject of the enquiry? And if you find out what you want, how will you ever know that this is what you did not know?

Plato handled this dilemma, as noted above, with his theory of "learning as remembering" and that experience in the physical world gradually brings the human back to its (pre-birth) understanding of pure forms. But for the modern empiricist the deeper question remains: What is abstraction, learning, the understanding of new relationships? What is causality, induction, and generalization?

Modern artificial intelligence practitioners have adopted both the empiricist and rationalist views of the world. To offer several simple examples: From the rationalist perspective came the expert system technology, knowledge was seen as a set of clear and distinct relationships (rules) that could be encoded within a production system architecture and that could then be used to compute decisions in particular situations. Traditional robotic planning was seen as presenting the world as a set of explicit constraints that were to be used to accomplish a particular task. Case based reasoning was a data base of collected and clearly specified cases that could then be used to address new and related problems (Luger 2009, Chapter 7).

From the empiricist perspective of AI there is the creation of semantic networks and conceptual dependencies. These structures, deliberately formed to capture the concept and property associations of the human agent, were used to "understand" human language and interpret meaning in specific contexts. Neural networks were designed to capture associations in collected data and to interpret and understand patterns in the world. (There were even the obvious – and scientifically useless - claims that neural connectivity was the way intelligent humans performed these tasks). Later approaches to robotics created a "subsumption" architecture that was supposedly a knowledge-free model for operation in world situations (Luger 2009, Chapter 7). Artificial life and genetic algorithms were proposed as structures that captured survival of the fittest and thus deserved to be seen as plausible models of the products of evolution.

It is not surprising that these approaches only met with limited successes. To give them their due, they have been useful in many of the application domains in which they were designed and deployed. But as models of human cognition, able to generalize to new related situations, or even to generalize or interpret their various results, they were failures. The thought that they could adapt to new categories and problem domains was a delusion. The success of the AI practitioner as the designer of new and important software tools is beyond question; the notion of the cognitive creditability of the products of these tools is simply naive.

We view a constructivist epistemology as a rapprochement between the empiricist and rationalist viewpoints. The constructivist hypothesizes that all understanding is the result of an interaction between energy patterns in the world and mental categories imposed on the world by the intelligent agent (Piaget 1954, 1970; von Glasersfeld 1978). Using Piaget's descriptions we *assimilate* external phenomena according to our current understanding and *accommodate* our understanding to phenomena that does not meet our prior expectations.

Constructivists use the term *schemata* to describe the *a priori* structure used to mediate the experience of the external world. The term schemata is taken from the British psychologist Bartlett (1932) and its philosophical roots go back to Kant (1781/1964). On this viewpoint observation is not passive and neutral but active and interpretative.

Perceived information, Kant's *a posteriori* knowledge, never fits precisely into our preconceived and *a priori* schemata. From this tension the schema-based biases a subject uses to organize experience are either modified or replaced. The use of *accommodation* in the context of unsuccessful interactions with the environment drives a process of cognitive *equilibration*. The constructivist epistemology is one of cognitive evolution and continuous refinement. An important consequence of constructivism is that the interpretation of any perception-based situation involves the imposition of the observers (biased) concepts and categories on what is perceived. This constitutes an *inductive bias*.

When Piaget proposed a constructivist approach to understanding the external world, he called it a *genetic epistemology*. When encountering new phenomena, the lack of a comfortable fit of current schemata to the world "as it is" creates a cognitive tension. This tension drives a process of schema revision. Schema revision, Piaget's *accommodation*, is the continued evolution of the agent's understanding towards *equilibration*.

Schema revision and continued movement toward equilibration is a genetic predisposition of an agent for an accommodation to the structures of society and the world. It combines both these forces and represents an embodied predisposition for survival. Schema modification is both an *a priori* reflection of our genetics as well as an *a posteriori* function of society and the world. It reflects the embodiment of a survival-driven agent, of a being in space and time.

There is a blending here of the empiricist and rationalist traditions, mediated by the requirement of agent survival. As embodied, agents can comprehend nothing except that which first passes through their senses. As accommodating, agents survive through learning the general patterns of an external world. What is perceived is mediated by what is expected; what is expected is influenced by what is perceived: these two functions can only be understood in terms of each other. In the following sections I propose several Bayesian models where prior experience conditions current interpretations and current data supports selection of interpretative models.

We, as intelligent agents, are seldom consciously aware of the schemata that support our interactions with the world. As the sources of bias and prejudice both in science and society, we are more often then not unaware of our *a priori* schemata. These are constitutive of our equilibration with the world and are not usually a perceptible component of our conscious mental life.

Finally, we can ask why a constructivist epistemology might be useful in addressing the problem of understanding intelligence itself? How can an agent within an environment understand its own understanding of that situation? I believe that constructivism also addresses this problem of *epistemological access*. For more than a century there has been a struggle in both philosophy and psychology between two factions: the positivist, who proposes to infer mental phenomena from observable physical behavior, and a more phenomenological approach which allows the use of first person reporting to enable the access of cognitive phenomena. This factionalism exists because both modes of access to cognitive phenomena require some form of model construction and inference.

In comparison to physical objects like chairs and doors, which often, naively, seem to be directly accessible, the mental states and dispositions of an agent seem to be particularly difficult to characterize. We contend that this dichotomy between the direct access to physical phenomena and the indirect access to mental phenomena is illusory. The constructivist analysis suggests that no experience of the external (or internal) world is possible without the use of some model or schema for organizing that experience. In scientific enquiry, as well as in our normal human cognitive experiences, this implies that *all* access to phenomena is through exploration, approximation, and continued model refinement.

In the following section we consider mathematical (computational) approaches to this exploratory model refinement process. We begin our analysis with Bayes' methods for probabilistic interpretations and schema building and refine this approach to a form of naïve Bayes, which we call the *greatest likelihood* measure, which uses continuous data acquisition to do real-time diagnosis through model refinement.

A Bayesian-Based and Constructivist Computational Model

We can ask how the computational epistemologist might build a falsifiable model of the constructivist worldview. Historically, an important response to David Hume's skepticism, described briefly in the previous section, was that of the English cleric, Thomas Bayes (1763). When challenged to defend the gospel's and other believers' accounts of Christ's miracles in the light of Hume's demonstrations that such "accounts" could not attain the credibility of a "proof", Bayes' genius responded with a mathematical demonstration of how an agent's prior expectations could be related to its current perceptions. Bayes' approach, although it didn't do much for the creditability of miracles, has had an important effect on the design of probabilistic models. In the final section we conjecture that Bayes can support an exciting computational model of epistemological phenomena.

We make a simple start; suppose we have a single symptom or piece of evidence, **e**, and a single hypothesized disease, **h**: we want to determine how a bad headache, for example, can be an indicator of a meningitis infection. We can visualize this situation with Figure 1, where we see one set, **e**, containing all the people having bad headaches and a second set, **h**, containing all the people that have the disease, meningitis. We want to get a measure of what the probability is of a person having a bad headache also having meningitis.

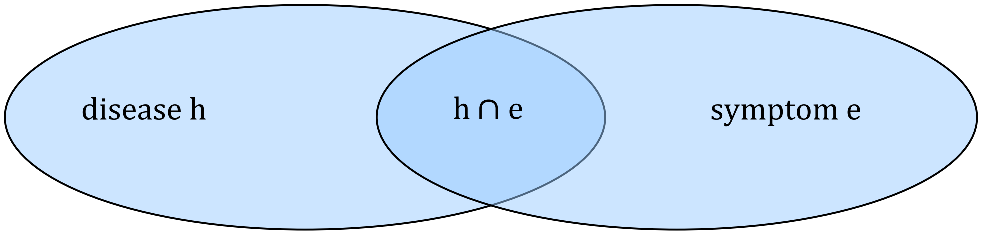

Figure 1. A representation of the numbers of people having a symptom, **e**, and a disease, **h**. Note that what we want to measure is the probability of a person having the disease, given that they suffer the symptom, **p(h|e)**.

We now determine the probability that a person having the symptom, **e**, also has the hypothesized disease, **h**. This probability can be determined by finding the number of people having both the symptom and the disease divided by the number of people having the disease. (We will concern ourselves with the processes for obtaining these actual numbers later.) Since both these sets of people are normalized by the total number of people, we can represent each number as a probability. We represent the probability of the symptom **e** given the disease **h** as **p(e|h)**:

$$\mathbf{p(e|h) = p(e \cap h) / p(h).}$$

The value of **$p(e \cap h)$** can now be determined:

$$\mathbf{p(e \cap h) = p(e|h)\ p(h)}$$

We wish to determine the **$p(e \cap h)$** value and to do so we have other information from Figure 1, including the number of people that have both the symptom and the disease **$e \cap h$,** as well as the total number of people that have the symptom, **e**. So we can determine the value for **$p(e \cap h)$** with this information: The probability of the disease **h**, given the evidence **e**, **p(h|e)**:

$$\mathbf{p(h|e) = p(e \cap h) / p(e)}$$

Finally, we have a measure of the probability of the hypothesized disease, **h**, given the evidence, **e**, in terms of the probability of the evidence given the hypothesized disease:

$$\mathbf{p(h|e) = p(e|h)\ p(h) / p(e)}$$

This last formula is Bayes' law for one piece of evidence and one hypothesized disease. But what have we just accomplished? We have created a relationship between the posterior probability of the disease given the symptom **p(h|e)** and the prior knowledge of the symptom given the disease **p(e|h)**. Our (or in this case the medical doctor's) experience over time supplies the prior knowledge of what should be expected when a new situation – a patient with symptoms – is encountered. The probability of the new person with symptom **e** having the hypothesized disease **h**, is represented in terms of the collected knowledge obtained from previous situations where the diagnosing doctor has seen that a diseased person had a particular symptom **p(e|h)** and how often the disease itself occurred, **p(h)** .

We can make the more general case, along with the same set-theoretic argument, of the probability of a person having a possible disease given two symptoms, say of having meningitis while suffering from both a bad headache and high fever. Again the probability of meningitis given these two symptoms will be a function of the prior knowledge of having the two symptoms when the disease is present along with the probability of the disease.

Next we present the general form of Bayes' law for a particular hypothesis, **h_i**, from a set of hypotheses, given a set of symptoms (evidence, **E**). The denominator of Bayes' theorem represents the probability of the set of evidence occurring. With the assumption of the hypotheses being independent, given the evidence, the intersection of each **h_i** with its piece of the evidence set forms a partition of the full set of evidence, **E**, as seen in Figure 2.

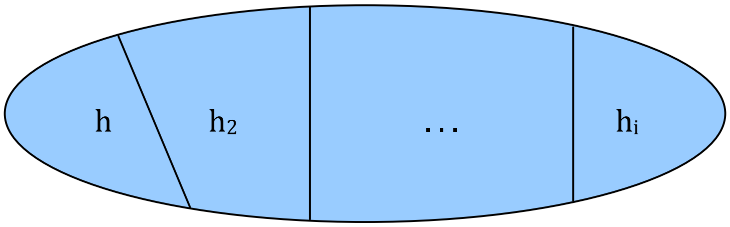

Figure 2. The set of evidence, **E**, is partitioned by the set of possible hypotheses, **h_i**.

With the assumption of this partitioning, the earlier equation we presented:

$p(e \cap h) = p(e|h)\ p(h)$

can be summed across all the **h_i** to produce the probability of the set of evidence, **p(E),** and the denominator of Bayes' relationship for the probability of a particular hypothesis, **h_i**, given evidence **E**:

$$p(h_i \mid E) = \frac{p(E \mid h_i)p(h_i)}{\sum_{k=1}^{n} p(E \mid h_k)p(h_k)}$$

$p(h_i|E)$ is the probability that a particular hypothesis, **h_i**, is true given evidence **E.**

$p(h_i)$ is the probability that **h_i** is true overall.

$p(E|h_i)$ is the probability of observing evidence **E** when **h_i** is true.

n is the number of possible hypotheses.

The general form of Bayes' theorem offers a functional (computational) description (model) of the probability of a particular situation happening given a set of perceptual clues. Epistemologically, the right hand side of the equation offers a "schema" describing how prior accumulated knowledge of occurrences of phenomena can relate to the interpretation of a new situation, the left hand side of the equation. This relationship can be seen as an example of Piaget's *assimilation* where encountered information fits the accepted pattern created from prior experiences.

To describe further the pieces of Bayes formula: The probability of an hypothesis being true, given a set of evidence, is equal the probability that the evidence is true given the hypothesis times the probability that the hypothesis occurs. This number is divided by (normalized by) the probability of the evidence itself. The probability of the evidence occurring is seen as the sum over all hypotheses presenting the evidence times the probability of that hypothesis itself.

There are several limitations to using Bayes' theorem as just presented as an epistemological characterization of the phenomenon of interpreting new (a posteriori) data in the context of (prior) collected knowledge and experience. First, of course, is the fact that the epistemological subject is not a calculating machine. We simply don't have all the prior values for all the hypotheses and evidence that can fit some problem situation. In a complex situation such as medicine where there can be well over a hundred hypothesized diseases and thousands of symptoms, this calculation is intractable (Luger 2009, Chapter 5).

A second objection is that in most realistic diagnostic situation the sets of evidence are NOT independent, given the set of hypotheses. This makes the mathematical version of full Bayes just presented unjustified. But the rationalization of the probability of the occurrence of evidence across all hypotheses can also be seen as simply a normalizing factor, supporting *the calculation of* a realistic probability measure for the probability of the hypothesis given the evidence (the left side of Bayes' equation). The same normalizing factor is utilized in determining the actual probability of any of the $\mathbf{h_i}$, given the evidence.

Finally, diagnostic reasoning is not about the calculation of probabilities, it is about the determination of the most likely *explanation*, given the accumulation of pieces of evidence. Humans are not doing real-time complex mathematics, rather we are looking for the most coherent explanation or possible hypothesis, given the amassed data.

A much more intuitive form of Bayes rule – often called *naïve Bayes* – ignores this $\mathbf{p(E\ )}$ denominator entirely. Naïve Bayes simply determines the likelihood of any hypothesis given the evidence, as the product of the probability of the evidence given the hypothesis times the probability of the hypothesis itself $\mathbf{p(E|h_i)\ p(h_i)}$.

In most diagnostic situations we are often required to determine which of a set of hypotheses $\mathbf{h_i}$ is most likely to be supported. We refer to this as determining the *argmax* across all the set of hypotheses. Thus, if we wish to determine which of all the $\mathbf{h_i}$ has the most support we look for the largest $\mathbf{p(E|h_i)\ p(h_i)}$:

$$\mathbf{argmax(h_i)\ \ p(E|h_i)\ p(h_i)}$$

In a dynamic interpretation, as sets of evidence themselves change across time, we will call this argmax of hypotheses given a set of evidence at a particular time the *greatest likelihood* *of that hypothesis at that time*. We show this relationship, an extension of the Bayesian *maximum a posteriori* (or *MAP*) estimate, as a dynamic measure over time **t**:

$$\mathbf{gl(h_i|E_t) = argmax(h_i)\ p(E_t|h_i)\ p(h_i)}$$

This model is both intuitive and simple: the most likely interpretation of new data, given evidence **E** at time **t**, is a function of which interpretation is most likely to produce that evidence at time **t** and the probability of that interpretation itself occurring.

We now ask how the argmax specification can produce a computational model of epistemological phenomena. First, we see that the argmax relationship offers a falsifiable approach to explanation. If more data turns up at a particular time an alternative hypothesis can attain a higher argmax value. Furthermore, when some data suggests an hypothesis, $\mathbf{h_i}$, it is usually only a subset of the full set of data that can support that hypothesis. Going back to our medical hypothesis, a bad headache can be suggestive of meningitis, but there is much more evidence that is also suggestive of this hypothesis including fever, nausea, and the results of certain blood tests.

We view the evolving greatest likelihood relationship as a continuing tension between possible hypotheses and the accumulating data collected across time. The presence of changing data supports the revision of the greatest likelihood hypothesis, AND, because data sets are not always complete, the possibility of a particular hypothesis motivates the search for data that either supports or falsifies it. Thus, greatest likelihood represents a dynamic equilibrium evolving across time of hypotheses suggesting supporting data and the presence of data combinations supporting particular hypotheses.

When, because of changing data, no new hypothesis is forthcoming, a greedy local search on the data points can suggest (create) new hypotheses. This technique supports *model induction*, the creation of a most likely model to explain the data, an important research component of machine learning (Luger 2009, Chapter 13).

The following section presents several computational examples from my own research group that utilizes this greatest likelihood dynamic equilibration process.

Computational Examples of Model Refinement and Equilibration

In recent research into the diagnosis of failures in discrete component semiconductors (Stern and Luger 1997, Chakrabarti et al. 2005) we have an example of creating the greatest likelihood for hypotheses across expanding data sets. Consider the situation of Figure 3, where we have two discrete component semiconductor failures.

Figure 3 shows examples of two different failures of discrete component semiconductors. This failure type is called an “open”, or the break in a wire connecting a component to others in the system. For the diagnostic expert the presence of a break supports a number of alternative hypotheses. The search for the most likely explanation for the failure broadens the evidence search: How large is the break? Is there any discoloration related to the break? Were there any sounds or smells on its happening? What are the resulting conditions of the components of the system?

Driven by the data search supporting multiple possible hypotheses that can explain the “open”, the expert notes the *bambooing* effect in the disconnected wire, Figure 3a. This suggests a revised greatest likelihood hypothesis that explains the open as a break created by metal crystallization that was most probably caused by a sequence of low frequency high current pulses. The greatest likely hypothesis for the open of the example of Figure 3b, where the break is seen as *balled* is melting due to excessive current. Both of these diagnostic scenarios have been implemented by an expert system-like search through an hypothesis space (Stern and Luger 1997) as well as with a Bayesian belief net (Chakrabarti et al. 2005). Figure 4 presents a Bayesian belief net (BBN) capturing this and other related diagnostic situations.

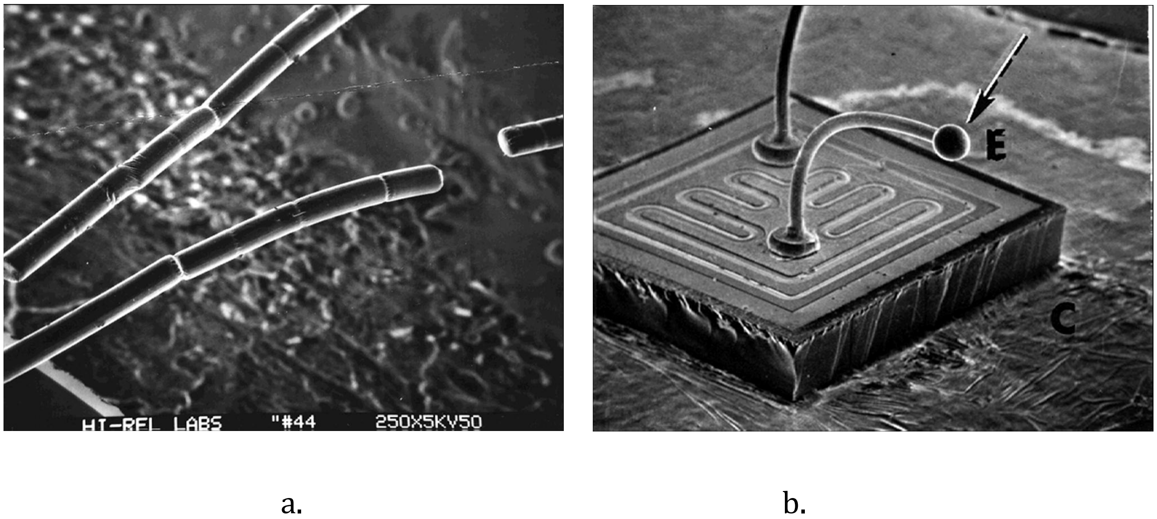

Figure 3. Two examples of discrete component semiconductors, each exhibiting the "open" failure.

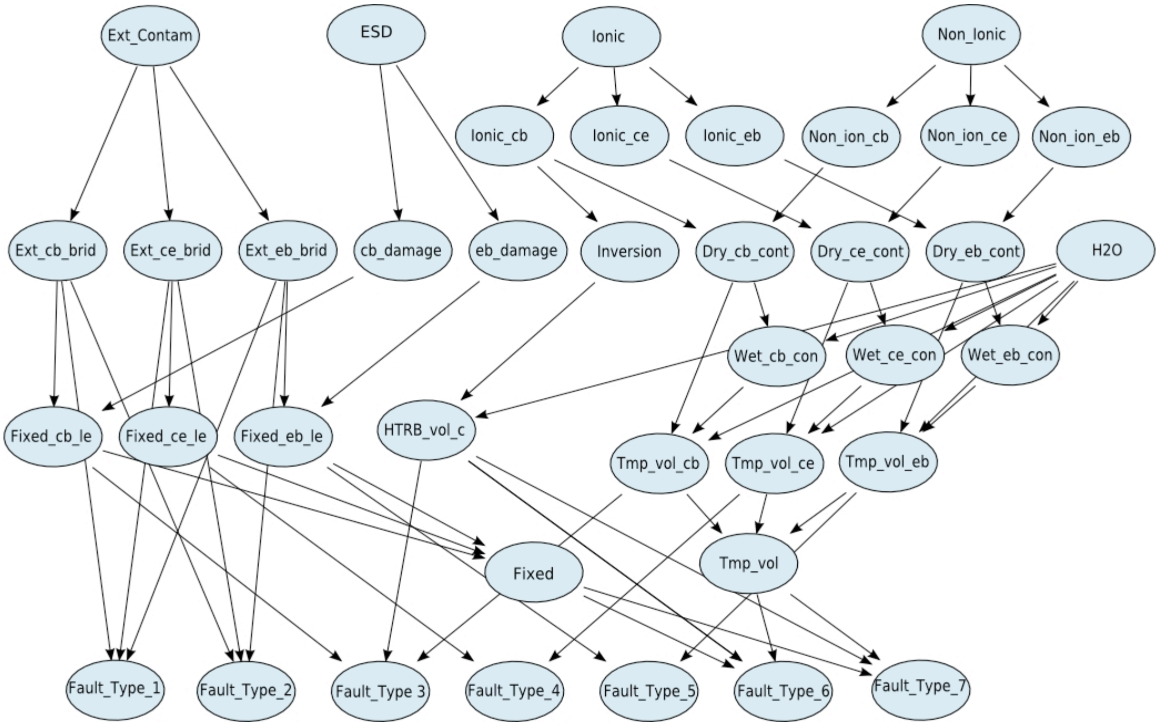

Figure 4. A Bayesian belief network representing the causal relationships and data points implicit in the discrete component semiconductor domain. As data is "discovered" the (a priori) probabilistic values change.

A Bayesian belief net (Pearl 1988) is a graph of probabilistic causal relationships that reflects the understanding of an application domain. Because a BBN is a causal model, its graph has two properties it is directed and without cycles (causes are directed to their effects and no effect can cause itself). These properties also support a further (drastic) reduction of the computational costs of full Bayesian inference: each node in the graph is independent of its non-descendents, given knowledge of its parents.

The BBN, without new data, represents the a priori state of an expert's knowledge of an application domain. In fact, these networks of causal relationships are usually carefully crafted through many hours working with human experts in that application domain. Thus, they can loosely be said to capture expert knowledge implicit in a domain of interest. When new (a posteriori) data are given to the BBN, e.g., the wire is "bambooed", the color of the copper wire is normal, etc, the belief network "infers" the most likely probabilities within its (a priori) model, given this new information. There are many inference rules for doing this (Luger 2009, Chapter 9), we will describe one of these, loopy belief propagation (Pearl 1988) later. An important result of using the BBN technology is that as one hypothesis achieves its greatest likelihood other related hypotheses are "explained away", i.e., their likelihood measures decrease.

In the failure of discrete component semiconductors, the discovery of new evidence occurred in a discrete fashion, that is, one piece of information at a time, the first evidence often suggesting consideration of related evidence. In our second example, (Chakrabarti et al. 2005) we have a continuous data stream from a set of distributed sensors. In monitoring the "health" of the transmission of Navy helicopter rotor systems, we receive a steady stream of sensor readings, mainly of temperatures, vibrations, and pressure information distributed across the various components of the running system. An example of this data can be seen in the top portion of Figure 5, where we have broken this continuous data stream into discrete and partial time slices.

We then use a fast Fourier transform to translate these signals into the frequency domain, as shown on the left side of the second row of Figure 5. These frequency readings were compared across time cycles to diagnose the running health of the rotor system. The method for diagnosing rotor health is by using the auto-regressive hidden Markov model (A-RHMM) of Figure 6. The observable states of the system are made up of the sequences of the segmented signals in the frequency domain while the hidden states are the imputed health values of the helicopter rotor system, as seen in the lower right of Figure 5.

The A-RHMM technology is used rather than a simple HMM because the human analysts suggested that the greatest likelihood value of the hidden nodes at any time would have some correlation with the values of these nodes at the previous time period (see the A-RHMM description in Luger 2009, Chapter 13). Training this system on streams of normal data allows the system to make the correct greatest correct greatest likelihood measure of the transmission when breakdowns occurred. The Navy supplied data both for the normal running system as well as for transmissions that contained seeded faults (Chakrabarti et al. 2007). Thus the hidden state $\mathbf{S_t}$ of the A-RHMM reflects the greatest likelihood hypothesis of the state of the rotor system, given the observed evidence $\mathbf{O_t}$ at time $\mathbf{t}$.

A final computational example of determining the greatest likelihood measure for hypotheses considers the model-calibration problem itself. What can be done if the data stream cannot be interpreted by the present state's (a priori) model? The problems we have considered to this point simply ask, what is the greatest likelihood hypothesis, given a model and a set of data across time. Now we ask what we can do when there is no interpretation of the model that fits the current data.

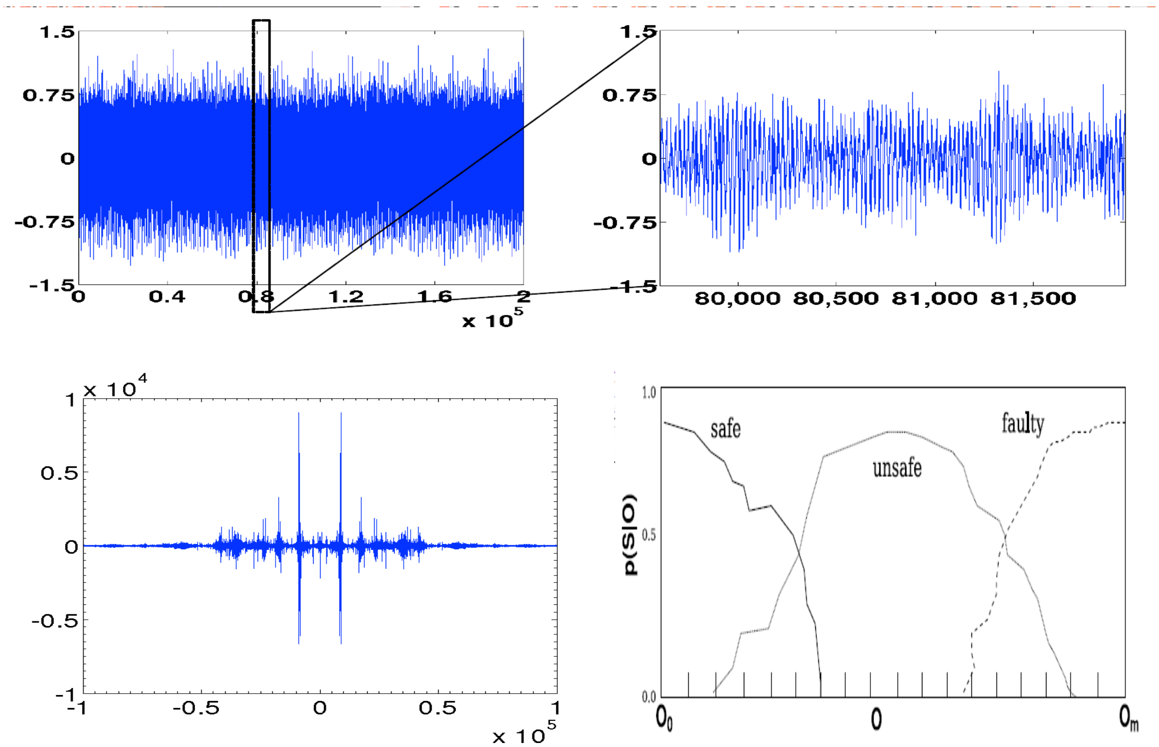

Figure 5. Real-time data from the transmission system of a helicopter's rotor. The top component of the figure presents the data stream and an enlarged time slice. The lower left figure is the result of the fast Fourier transform of the time slice data (transformed) to the frequency domain. The lower right figure represents the hidden states of the rotor system.

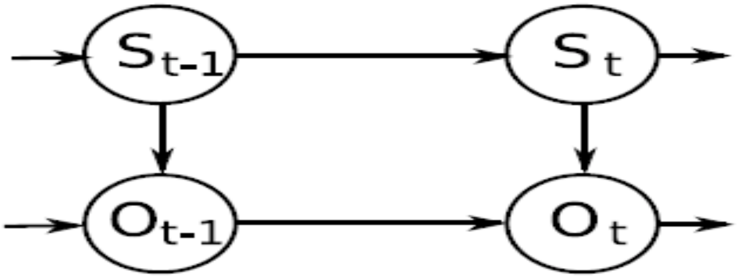

Figure 6. The data of Figure 5 is processed using an auto-regressive hidden Markov model as in Figure 6. States **O$_t$** represent the observable values at time **t**.

The $\mathbf{S_t}$ states represent the hidden "health" states of the rotor system, **{safe, unsafe, faulty}** at time **t**.

Figure 7 presents an overview of this situation, where, on the top row, a cognitive model either offers an interpretation of data or it does not. Piaget has described these situations as instances of *assimilation* and *accommodation*. Either the data fits, possibly requiring the model to slightly adjust its (probabilistic) expectations (assimilation), or the model must reconfigure itself, possibly adding new variable relationships (accommodation). The lower part of Figure 7 presents our (COSMOS) architecture (Sakanenko and Luger 2009) that addresses both these tasks.

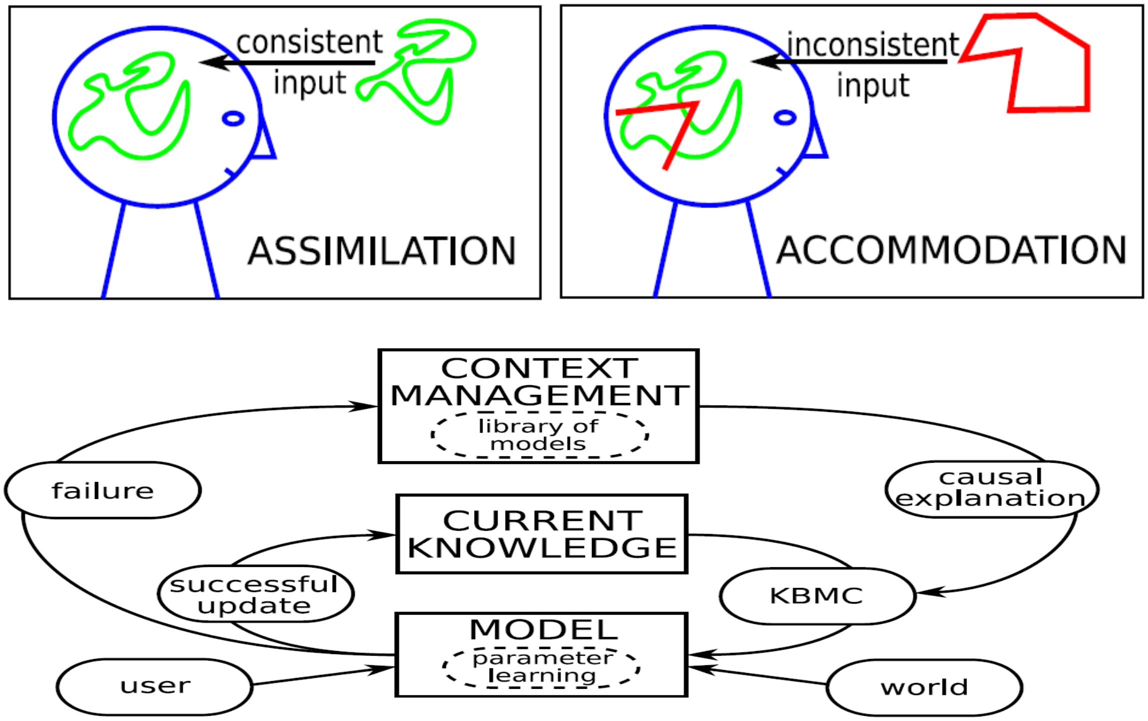

Figure 7. Cognitive model use and failure, above; a model-calibration algorithm, below, for assimilation and accommodation of new data.

Although this model calibration algorithm has been tested in complex tasks such as that of pumps, pipes, filters, and liquids, complete with real-time measures of pressure, pipe flow, filter clogging, vibrations, and alignments, I describe this "model calibration" idea in a simpler situation of home burglar alarms.

Suppose we have developed a probabilistic home burglar alarm and monitoring system. We then deploy many of these alarm systems in a certain city and test their outputs across multiple situations, in particular monitoring these systems for false positive predictions. Suppose this system is deployed successfully over four winter months where we learn the probabilistic values for the outputs of alarm monitoring system. The day-to-day deployment produces data that are used to condition the system. After a time of training the new daily data is easily assimilated into the model and the resulting trained system successfully reports both false alarms and actual robberies.

We then find ourselves in the spring months of the year where we encounter multiple fierce desiccating winds that shake the components – those mounted on doors and windows - of the alarm systems and dry out their connections. When our monitoring system sees many more false positive results that no longer fit comfortably into the previously trained system it is necessary to readjust the probabilities of the model and add new parameters reflecting the spring wind conditions. The result will be a new model for the spring monitoring situation.

Furthermore, when our alarm systems are sold in a new city it will need to determine which of its library of models will best fit that new situation. If there are other important variables, such as lots of small earthquake tremors, this variable will probably also need to be modeled. Although the problem of model induction in general is intractable, we feel in these knowledge intensive situations we can create useful new models, using ideas such as causality and the greedy local search of constraints located near the points of model failure (Sakhanenko et al. 2006, Sakhanenko and Luger 2009).

There other examples in the literature of reasoning to the greatest likelihood in diagnostic and prognostic systems, for example, in computer based speech processing where, with a series of n-grams, sounds and/or words, are tested against the contents of corpora of language data. In these situations techniques such as the probabilistic form of dynamic programming, the Viterbi algorithm (Luger 2009, Chapter 13), process continuous streams of data to produce greatest likelihood results.

More recent utilization of the greatest likelihood hypothesis is the study of an agent's active intervention in the world, for example, that agent's direct manipulation of the parameters that change the data stream itself, in an attempt to better understand the most likely model, including possible causality relationships, that "support" an interpretation of the data. This interventionist approach to interpretation was first proposed by Judea Pearl (2000) and his students (Tian and Pearl 2001), and has been proposed by psychologists as a methodology for an agent's causal understanding of its world (Gopnik et al., 2004). This active interventionist approach for model discovery and refinement is also being explored in my own research group (Rammohan and Luger 2010).

The final section offers some more speculative conjectures about possible cognitive architectures that could support the computational calculation of the greatest likelihood schemas.

A Possible Cognitive Architecture for Greatest Likelihood Calculation

Neuroscience has seen radical change, based mostly on better imaging technology, over the past fifty years. The neurophysiology and the brain science lectures we all had in the 1950s and 1960s now seem like exercises in the primitive practice of phrenology. Some insights of earlier times still remain quite relevant, however, including Hebbian (1949) learning.

Thanks to new sophistication in fMRI and other neuro-imaging techniques along with the development of sophisticated stochastic tools for model induction, calibration, and computational inference we now know much more about cortical architecture and processing. In particular, pre-frontal cortex or Broadmann's areas are often seen as the primary "location" for model creation and hypothesis generation.

If we consider the Bayesian-based schemas and in particular our formula for calculating a greatest likelihood measure, we note a balance or tension between the left hand side, the set of hypotheses, and the right hand side, the scoring of perceptual phenomena, that support these hypotheses. Consider Figure 8, a diagram of related components of this situation, where each of a set of hypotheses, $\mathbf{h_i}$, is linked to a number of perceptual indicators, $\mathbf{e_i}$ elements of evidence set **E** that support that hypothesis. It should be noted that even though each hypothesis is assumed to be unique, the pieces of evidence may each support more than one hypothesis.

First, we consider each of these hypotheses and their supporting data sets to be attractor networks (Luger 2009, Section 11.5). An attractor network, sometimes referred to as a *basin of attraction*, serves as a generalized pattern that can be either partially of totally matched. The insight behind the attractor network model is that stimuli rarely match perfectly to their archetypical patterns, but are messy, partial, and usually compromised. Thus the basin of attraction is essential for partial and incomplete matches. But this is a sign of natural intelligence, where perfect information is most likely an unattainable ideal – and yet incomplete partial perceptions are often sufficient.

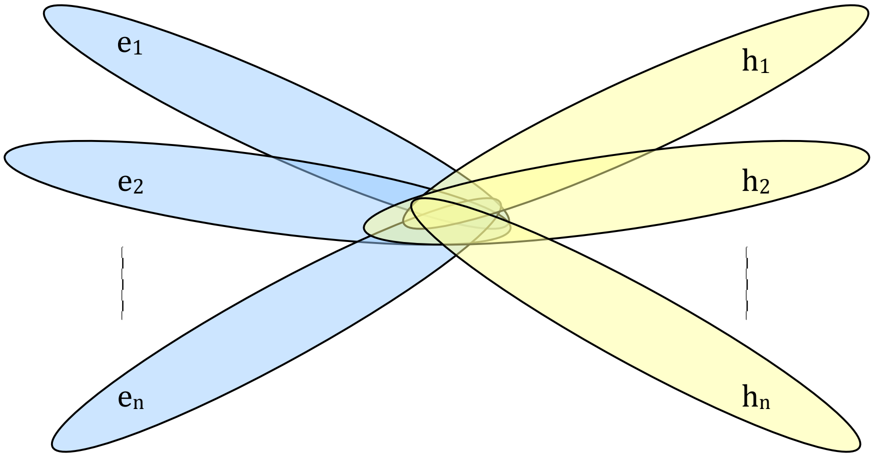

Figure 8. A set of hypothesis $\mathbf{h_i}$ are linked to multiple (supporting) perceptions, evidence $\mathbf{e_j}$**.** It is hypothesized that the hypothesis space is located in the pre-frontal cortex (Broadmann's area), while the intersection of the hypothesis space with the various sense modalities is located in the hippocampus.

The second conjecture supporting the diagnostic model of Figure 8 is that any perceptual information, given our basins of attraction model, is able to trigger potential hypotheses. That is, as with natural intelligence, data patterns, even partial matches, are able to trigger entire attractor networks. In this sense what we see triggers its explanation, just like appropriate expectations condition our interpretation of perceptual information: the data triggers appropriate models and suggested hypotheses motivate the search for particular pieces of perceptual information.

Third, Hebbs' reinforcement rule supports the calculation of the $\mathbf{p(h_i)}$ parameter in the greatest likelihood calculation of $\mathbf{p(E_t|h_i)\ p(h_i)}$. The likelihood of a particular hypothesis is a measure that is conditioned across time, as the more a particular hypothesis is seen the more it is expected. Similarly, even though a piece of evidence might be indicative of an hypothesis, this fact alone is not sufficient to trigger the greatest likelihood condition for that hypothesis. Thus the importance of the perception of various pieces of data is conditioned on the likelihood of that hypothesis itself, and when the hypothetical situation is extremely rare, the search can continue for other more plausible hypotheses – given the same pieces of evidence.

The discussion to this point has shown how a cognitive a priori equilibrium can integrate and interpret a posteriori evidence. Our model also demonstrates that a greatest likelihood hypothesis is supported both by the a posteriori evidence suggestive of particular hypotheses and the likelihood of that hypothesis. This approach to model support and calibration is suggestive of Piaget's *assimilation*.

I would suggest that our architecture is also maximally flexible for the integration of new hypothesis/evidence relationships. New relationships may be learned though an educational process or simply conditioned through observations of a talented diagnostician. Regardless of the method of learning, the new hypothesis/evidence relationships are integrated as further related (Figure 8) basins of attraction: new hypotheses are conjoined with their related evidence sets and integrated into the larger diagnostic model. Piaget might refer to this phenomenon as *accommodation*. Similar arguments support the determination of the greatest likelihood for combinations of hypotheses, given sets of evidence. Thus, model calibration and revision is an active process of the intelligent agent interacting with its environment: as it learns, it rewards; when it fails it revises.

A cortical column based cognitive architecture (Hawkins 2004) would support both the representation for and algorithms to compute a greatest likelihood measure. The hypothesis space of Figure 8 would be located in the columns of pre-frontal cortex or Broadmann's areas. The evidence sets would be linked to the cortical receptors for the human sensory and memory modalities. The critical overlap of the hypothesis space with the sensory modalities is the hippocampus area of cortex. Thus, sense and memory stimulation trigger and support related hypotheses and the possibility of a particular hypothesis would suggest related sensory and memory data.

There exist a number of algorithms created for real-time integration of posterior information and its propagation into (a priori) stochastic models (Luger 2009, Chapters 5, 9, 13). There are also many computational inference rules for calculating the greatest likelihood in probabilistic modeling situations.

For "cognitive creditability" we propose a form of the *loopy belief propagation* (Pearl 1989) algorithm as it reflects a system constantly iterating towards equilibrium (or *equilibration*, as Piaget might describe it). A cognitive system is in what can be called *a priori equilibrium* with its continuing states of learned diagnostic knowledge. When presented with the novel information characterizing a new diagnostic situation, this a posteriori data perturbs the equilibrium. The (cognitive) system then iterates by sending "messages" between near-neighbors' prior and posterior components of the model until it finds convergence or equilibrium, usually with support for a greatest likelihood hypothesis.

Iteration of this system can be (intuitively) seen as integrating small perturbations of the values of neighbors in the system, aimed at achieving compatible equilibrating measures, and a stable state of the system is reached. The iteration process itself can be visualized as continuous message passing between near neighbors ("I've got these values; what are yours? Let's each make slight adjustments moving toward a probabilistic compatibility") attempting to determine the most appropriate set of values for the entire system once the a posteriori information is added to the previous (a priori) equilibrium.

This iteration process can also be seen as a method to account for incomplete or missing information in a situation given a priori equilibrium. The iterative message passing suggests most likely values for missing or lost data, given the state of a priori equilibrium. In our research (Sakhanenko et al. 2008) this is shown to be a form of expectation-maximization (EM) learning (Dempster et al. 1977). Furthermore, research has demonstrated that the generalized loopy belief propagation algorithm iterates to optimal states of equilibrium when the original a priori belief state is reflected with a non-cyclic directed graph (Pear 1988).

In concluding these discussions we note that we are not claiming the human diagnostic system is *doing* loopy belief propagation on an explicit graphical model when it moves to finding equilibration through interpreting a posteriori information. This type reduction of cognitive phenomena to computational representations and algorithms has long been questioned by researchers including Anderson (1978, see *representational indeterminacy*) and Luger (1994, Chapter 4).

Rather, the claim of this chapter is that stochastic models coupled with the loopy belief propagation algorithm offer a *sufficient* account of the cortical computation of the *greatest likelihood* measure given a priori cognitive equilibrium and the presentation of novel information. Further, I suggest that this greatest likelihood calculation IS cognitively plausible, and thus supports an epistemological stance on understanding the phenomena of human diagnostic and prognostic reasoning, as well as addresses the larger question of how agents can come to understand their own acts of interpreting a complex and often ambiguous world.

References

Anderson, J. R., 1978. Arguments Concerning Representation for Mental Imagery. *Psychological Review*, 85: 249-277.

Bartlett, F., 1932. *Remembering*. London: Cambridge University Press.

Bayes, T., 1763. Essay Towards Solving a Problem in the Doctrine of Chances. Philosophic Transactions of the Royal Society of London, London: The Royal Society, pp 370-418.

Bower, T. G. R., 1977. *A Primer of Infant Development*. San Francisco: WH Freeman.

Bundy, A. 1983. *Computer Modeling of Mathematical Reasoning*. New York: Academic Press.

Chakrabarti, C., Rammohan, R., and Luger., G. F., 2005. A First-Order Stochastic Modeling Language for Diagnosis, in *Proceedings of the 18th International Florida Artificial Intelligence Research Society Conference, (FLAIRS-1*8). Palo Alto: AAAI Press.

Chakrabarti, C., Pless, D. J., Rammohan, R., and Luger, G. F., 2007. Diagnosis Using a First-Order Stochastic Language That Learns, *Expert Systems with Applications*. Amsterdam: Elsevier Press. 32 (3).

Dempster, A.P., 1968. A Generalization of Bayesian Inference, *Journal of the Royal Statistical Society*, 30 (Series B): 1-38.

Descartes, R., 1680. *Six Metaphysical Meditations, Wherein it is Proved That there is a God and that Man's Mind is really Distinct from his Body*. W. Moltneux, translator, London: Printed for B. Tooke.

Goldin G. A. and Luger G. F. 1975. Problem Structure and Problem Solving Behavior. In *Proceedings of IJCAI, 1975, Tiblisi, USSR*. Cambridge Mass: MIT Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D., 2004. A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological Review*, 111(1), p 3-32.

Hawkins, J., 2004. *On Intelligence*. New York: Times Books.

Hebb, D.O., 1949. *The Organization of Behavior*, New York:Wiley.

Kant, I. 1781/1964. *Immanuel Kant's Critique of Pure Reason*. Smith, N.K., translator, New York: St. Martin's Press.

Luger, G. F., 2009. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, 6th edition, Boston: Addison-Wesley Pearson Education.

Luger, G. F., 1994. *Cognitive Science: The Science of Intelligent Systems*. New York: Academic Press.

Luger, G. F. 1981. Mathematical Model Building in the Solution of Mechanics Problems: Human Protocols and the MECHO Trace. *Cognitive Science* (5), p 55-77.

Luger, G. F. 1978. *Cognitive Psychology: Learning and Problem Solving. Unit 28: Formal Analysis of Problem Solving Behaviour*. Milton Keynes: The Open University Press.

Luger, G. F., 1976. The Use of the State Space to Record the Behavioral Effects of Subproblems and Symmetries in the Tower of Hanoi Problem. *Int. Journal of Man-Machine Studies*, 8.

Luger, G. F., Lewis, J. A., and Stern, C., 2002. Problem Solving as Model-Refinement: Towards a Constructivist Epistemology. *Brain, Behavior, and Evolution*, Basil: Krager, 59: 87-100.

Luger, G. F., Bower, T. G. R., and Wishart, J. G., 1983. A Model of the Development of the Early Infant Object Concept. *Perception*, 12(1) p 21-34.

Luger, G. F., Wishart, J. G., and Bower, T. G. R., 1984. Modeling the Stages of the Identity Theory of Object-Concept Development in Infancy. *Perception* (13) p 97-113.

Luger, G. F. and Bauer, M. A. 1978.Transfer Effects in Isomorphic Problem Situations. *Acta-Psychologica* (42) p 121-31.

McGonigle, B. O.and Chalmers, M., 1977. Are monkeys logical?. *Nature* **267**: 694–696

Newell, A, and Simon, H. A. 1972. *Human Problem Solving*, Englewood Cliffs, NJ: Prentice-Hall.

Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Los Altos CA: Morgan Kaufmann.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge UK: The University Press.

Peirce, C.S., 1958. *Collected Papers 1931 – 1958*. Cambridge MA: Harvard University Press.

Piaget, J., 1954. *The Construction of Reality in the Child*. New York: Basic Books.

Piaget, J. 1970. *Structuralism*, New York: Basic Books.

Plato, 1961. *The Collected Dialogues of Plato*. Hamilton, E. and Cairns, H. eds. Princeton: Princeton University Press.

Rammohan, R. and Luger, G. F., 2010. *Causal Learning from Context Transitions,* (in press, copies available from authors).

Sakhanenko, N. A., Rammohan, R. R., Luger, G. F. Stern, C. R., 2008. A New Approach to Model-Based Diagnosis Using Probabilistic Logic, in *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference (FLAIRS-21)*, Palo Alto: AAAI Press, 2008.

Sakhanenko, N. A. and Luger, G. F., 2009. *Model Failure and Context Switching Using Logic-based Stochastic Models* – in press, copies available from authors.

Stern, C.R. and Luger, G. F., 1997. Abduction and Abstraction in Diagnosis: A Schema-Based Account. In *Expertise in Context*. Feltovitch, P. J. et al. eds, Cambridge MA: AAAI/MIT Press.

Tian, J. and Pearl, J., 2001. Causal Discovery from Changes. In *Proceedings of UAI '2001*, p 512-521 Morgan Kaufman.

von Glaserfeld, E., 1978. An Introduction to Radical Constructivism. In *The Invented Reality*, Watzlawick, ed., pp17-40, New York: Norton.

Young, R. M., 1976. *Seriation by Children: A Production System Approach*. Basel: Birkhauser-Verlag.